Imperial College London

BioMedIA
Imperial College London

Probabilistic and Causal Reasoning in Deep Learning for Imaging

Author Nick Pawlowski
Submitted 30th September 2021

Supervisors Dr. Ben Glocker[1]
Dr. Aditya Nori[2]
Prof. Daniel Rueckert[1]

[1]BioMedIA, Department of Computing, Imperial College London, UK.
[2] Microsoft Research, Cambridge, UK.

A thesis submitted in fulfilment of the requirements for the degree of *Philosophiae Doctor*

Biomedial Image Analysis Group, Department of Computing,
Imperial College London

BioMedIA,
Department of Computing,
Huxley Building,
Imperial College London,
180 Queens Gate,
South Kensington,
SW7 2AZ

I may not have gone where I intended to go, but I think I have ended up where I needed to be.

Douglas Adams - *The Long Dark Tea-Time of the Soul*

I often stumbled and changed directions throughout this PhD;
but I believe I arrived where I should be.

To my family – the ones who see me finishing this as well as the ones who don't.

Declaration

I, Nick Pawlowski, hereby declare that this thesis is my own work unless otherwise specified. All published papers that are reproduced in this thesis are appropriately referenced and the reproduction adheres to the relevant copyright licenses. The thesis template has been adapted from Rob Robinson (Robinson 2020), who made his template available on his Github https://github.com/mlnotebook.

Copyright

Acknowledgements

First, I would like to thank my family, especially my parents, for their everlasting love and support – regardless of whether I was physically close or on another continent – and for always having my back and providing the strength to dealing with the unexpected things that life throws at you; as well as generally giving me the environment to pursue things like this PhD.

I would like to express my gratitude to my supervisor, Dr. Ben Glocker, for the constant guidance and unwavering support throughout the last five years. Thank you for creating an inspiring and encouraging research group – as well as always having time to discuss potentially absurd research ideas, and never getting tired of me jumping from one project to the next or joining some industry group for an internship. Also a big thank-you to Marc Deisenroth for the ongoing mentoring and advice, and the wider BioMedIA group, especially Benhard Kainz and Daniel Rückert, for the collaborative research environment.

I am hugely thankful to David Dao, who has been a good friend since our first meeting at CERN almost a decade ago and is the one who sparked the idea of me moving into the field of machine learning. Furthermore, I want to thank all mentors, advisers and managers, that I got to work with on my path to this thesis, and that continue to provide invaluable insights and advice: Amos Storkey, my MSc supervisor in Edinburgh, and the imaging platform at the Broad institute, especially Shantanu Singh, for allowing me to experience real-world research so early in this journey and for still providing guidance whenever it is needed; the Cortex team at X, especially Albin Jones, Georgios Evangelopoulos and Bin Ni, for the opportunity to see how moonshots are pursued; Michal Drozdzal and the FAIR team in Montreal for the chance of working in an industry lab (with access to unreasonable numbers of GPUs) and allowing me to approach research in my own pace, in a time when things seemed difficult; the Frontiers team at Google Health, especially Jim Winkens and Alan Karthikesalingam, and its friends for the opportunity to work on problems related to real-life deployment of ML systems for medical applications.

Thanks also go to my close collaborators and friends within BioMedia – Daniel Castro, Miguel Monteiro, Sebastian Popescu, Stefan Winzeck and Matthew Lee – and outside – Miguel Jaques, Steindor Sæmundsson, Andrew Brock and Maximilian Ilse – with whom I was able to discuss and explore exciting research ideas or the simple (or sometimes not so simple) problems of life; I am especially thankful to those ones who took the time to provide comments on this thesis. Moreover, I want to thank all those in the BioMedIA group, Party {Bus, Boat, Rocket, . . . }, din dins, fellowship of the crimp, 5pm pints, and the 2016 CDT space cohort for the fun times and for making this whole PhD experience a lot of fun. Lastly, I want to thank my cheezy flatmates, Jeevan and Margherita,

as well as my partner, Friederike, for all the fun, food and wine, and for the support to stay sane and push this thesis forward throughout this ongoing pandemic.

My PhD was supported by a Microsoft Research PhD Scholarship as well as the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, grant ref EP/L016796/1). All UK Biobank data used in this work is accessed under Application Number 12579. The template for this thesis has been adapted from Robinson (2020), who kindly made the template publicly available at https://github.com/mlnotebook.

Abstract

Typical machine learning research in the imaging domain occurs in clearly defined environments on clean datasets without considering realistic deployment scenarios. However, applied machine learning systems are exposed to unexpected distribution shifts and still need to produce reliable predictions without relying on spurious correlations. Similarly, such systems encounter ambiguous or unseen inputs and need to communicate their uncertainty. Often, AI systems support a human operator and should provide interpretable explanations of their decisions. This thesis argues for a probabilistic and causal approach to machine learning that is robust to spurious correlations, improves interpretability, and communicates uncertainty. First, we investigate the learning abilities of neural networks that are constrained to extracting information from image patches. We show that careful network design can prevent shortcut learning and that restricting the receptive field can improve the interpretability of predictions. We tackle uncertainty estimation by introducing a Bayesian deep learning method to approximate the posterior distribution of the weights of a neural network using an implicit distribution. We verify that our method is capable of solving predictive tasks while providing reliable uncertainty estimates. Moving on, we frame various medical prediction tasks within the framework of outlier detection. We apply deep generative modelling to brain MR and CT images as well as histopathology images and show that it is possible to detect pathologies as outliers under a normative model of healthy samples. Next, we propose deep structural causal models as a framework capable of capturing causal relationships between imaging and non-imaging data. Our experiments provide evidence that this framework is capable of all rungs of the causal hierarchy. Finally, with further thoughts on applications of uncertainty estimation, robust causal estimation, and fairness we conclude that the safe and reliable deployment of AI systems to real-world scenarios requires the integration of probabilistic and causal reasoning.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Humankind has long been fascinated by the idea of developing intelligent automata – both to decrease the burden of work on people as well as to better understand intelligence, following the notion of "what you cannot create, you do not understand". Some of the first ideas of artificially intelligent objects can be found in Greek mythology, e.g. with the giant Talos. Talos was a bronze giant built by the Greek god Hephaestus to protect the island of Crete from invaders. Additionally, Hephaestus made other self-moving objects – some of which he used as servants (Mayor 2020). In the early middle ages, Ramon Llull created the foundation which inspired Gottfried Leibniz to work on an "alphabet of human thought" (Fidora and Sierra 2011; Press 2020; Schmidhuber 2021). The last century saw the formalisation and establishment of the field of artificial intelligence as its own field of research (Haenlein and Kaplan 2019).

In recent years the main driver of progress in the field of artificial intelligence was the ever-growing research on deep learning (LeCun et al. 2015). The current interest in deep learning was initially spurred by the significant win of AlexNet (Krizhevsky et al. 2012) in the ImageNet challenge (Deng et al. 2009). From there neural networks have been trained to achieve superhuman performance on a range of Atari games (Mnih et al. 2015), beat humans at the game of Go (Silver et al. 2016) or Dota (OpenAI et al. 2019), or predict the structure of proteins (Tunyasuvunakool et al. 2021). While some of those advances seem rather playful, they tackle important problems in the field of artificial intelligence and combine capabilities such as perception, complex reasoning and decision making. Other advances, such as the work on protein folding, help to advance human knowledge outside the field of artificial intelligence.

Image analysis is one of the application areas of artificial intelligence and machine learning that already sees a lot of commercial use cases based on newly developed methods, such as self-driving cars or checkoutless shops like Amazon Go. A special application area is the field of medical image analysis with the aim of detecting pathologies or otherwise supporting human clinicians (Esteva et al. 2021). Research articles have proposed machine learning models with performances equivalent to or even surpassing human experts in fields such as dermatology (Liu et al. 2020), breast cancer screening (McKinney et al. 2020), or diabetic retinopathy (Gulshan et al. 2016). However, the deployment of this research into clinical routine is lengthy and few systems are in active use.

Some of the issues of deploying deep learning models into the real world stem from the disparity of the approaches to reasoning between humans and machine learning systems. Many of the potential medical imaging applications are set out to provide decision support to human expert operators. However, while humans are capable of expressing their reasoning in logical statements[1] about the relation of the state of various objects or patterns as well as their corresponding certainty of a found conclusion, many artificial systems are not. Thus, the often rudimentary output of an AI system that e.g. simply predicts a scan to be 'cancerous' or 'healthy' without further specifying why that might be the case hinders human-AI interactions, as it is hard for the operator to interpret and trust. Human trust in the reliable working of such systems is also eroded by the presence of adversarial examples as well as the habit of neural networks to learn shortcut predictors that might be independent of the actual predictive task but confounded in the used training dataset.

Human-AI interaction would benefit from human-understandable explanations of why a machine learning system makes a certain prediction. A wide range of approaches aim to improve the interpretability of decisions made from visual inputs (Zhang and Zhu 2018). However, many of the proposed solutions simply present parts of the image that influence the decision. They therefore require significant understanding of the method from the human operator, as it can be hard for a human to understand how a certain area in an image is related to the prediction made.

Furthermore, machine learning models are often seen as universal function approximators and flexible on purpose. Instead, human decision making often relies on Occam's razor principle which prefers simpler solutions. In machine learning, this is often encouraged through various forms of regularisation techniques that penalise functional complexity. However, this notion of complexity is often hard to interpret in a humanly understandable way. An easier way to explain the complexity of a machine learning model would be through hard limitations such as "This network has a receptive field of $5 \times 5\mathrm{cm}$ and can only extract relations on that scale." or "This network uses a hard attention mechanism which only allows it to analyse data in a specific region of the input.". This form of explanation offers specific and easy to understand limits that allow for an operator to judge whether this mechanism is adequate for the task at hand. Additionally, it is possible to

[1] One could argue that humans sometimes only rationalise their conclusions or decisions post-factum. As such, it is not clear whether humans can actually logically express their reasoning or rather come up with a plausible explanation. However, this question probably could fill a PhD thesis on its own and is left for the reader to ponder over.

use domain knowledge to design networks with task-specific limitations that guide the learning process, making it harder for the model to pick up on unwanted spurious correlations.

Lastly, trust and safety largely rely on the notion of uncertainty. Perception in the real-world is ambiguous and noisy. Objects might be occluded or not fully visible and therefore leave their full nature to our imagination. Measurement devices always produce noisy measurements due to physical limitations and imperfections of sensors. Knowledge in itself can be in the process of being learned and therefore deductions might be uncertain. As such it is important for a machine learning system to express its certainty of a prediction to allow for human interpretation and potential deferral to a more experienced fallback system – human or not.

1.2 Research Aims and Thesis Outline

Following the above motivations, this thesis aims to explore research that allows for better human-AI interactions through three complementary goals:

1. **To train neural networks that ignore spurious correlations:** Neural networks often learn shortcuts to solve predictive tasks. Those shortcuts are features that are only correlated with the quantity of interest in the specific training set and therefore do not generalise to unseen data. Can we build neural networks that ignore spurious correlations? Can hard constraints on the flexibility of the learned functions eliminate shortcuts?

2. **To teach machines to know when they do not know:** Focusing on the specific scenario in which a model is uncertain due to limited amounts of training data, the estimation of the parameters of the model is constrained to only approximate the true parameters and benefits from probabilistic treatment. Whenever a model can reliably tell whether it is uncertain, it is possible to use this as a cue to defer to a human expert or flag abnormalities.

3. **To enable neural networks to leverage causal relations:** Causal reasoning offers tools for explaining decisions through causal factors as well as the imagination of counterfactual examples. Imagining counterfactuals requires the learning or prior specification of the assumed causal relationships. Similarly, neural networks that use prior knowledge to inspire its architecture can learn the true function more easily.

The thesis explores different angles at addressing these goals and is structured as follows. Each chapter is prefaced with a box stating which publication(s) the chapter is based on.

Chapter 2 describes background knowledge necessary to frame the contributions of this thesis. It introduces basic concepts in probabilistic modelling and Bayesian inference. Later it illustrates the evolution of deep learning for imaging applications as well as some core differences between medical and natural image analysis. Finally, it presents deep learning techniques applicable to modelling the generative process of high-dimensional data.

Chapter 3 aims to build neural networks that ignore spurious correlations (**Goal 1**) by studying the effect of constraining convolutional neural networks to small receptive fields – only allowing the network to model the relations within patches of the original image. The first part of the chapter is based upon (Pawlowski et al. 2019) and explores how different design choices influence the capabilities of a neural network to extract information from larger images in which all the relevant information is contained within small patches. This is an unusual scenario for natural images but regularly occurs in medical imaging, e.g. histopathology. The second part of that chapter, based on (Pawlowski and Glocker 2019), aims to answer the question of whether patches from brain magnetic resonance imaging contain enough information to reliably predict global information, such as a subject's age or biological sex.

- **Pawlowski, N.**, Bhooshan, S., Ballas, N., Ciompi, F., Glocker, B., and Drozdzal, M. (2019). "Needles in Haystacks: On Classifying Tiny Objects in Large Images". In: *arXiv preprint arXiv:1908.06037* – (Pawlowski et al. 2019)
- **Pawlowski, N.** and Glocker, B. (2019). "Is Texture Predictive for Age and Sex in Brain MRI?". In: *Medical Imaging with Deep Learning Abstract track* – (Pawlowski and Glocker 2019)

Chapter 4 introduces a novel technique, *Bayes by Hypernet*, for capturing the model uncertainty of neural networks – and teaching machines to know when they do not know (**Goal 2**). *Bayes by Hypernet*, published in (Pawlowski et al. 2017a), uses hypernetworks (Ha et al. 2017) to build an implicit variational approximation to the posterior of the weight distribution of a neural network. The implicit variational distribution enables highly complex distributions to be modelled. *Bayes by Hypernet* achieves competitive predictive performances while allowing for reliable uncertainty estimates.

- **Pawlowski, N.**, Brock, A., Lee, M. C., Rajchl, M., and Glocker, B. (2017a). "Implicit Weight Uncertainty in Neural Networks". In: *NeurIPS Workshop on Bayesian Deep Learning* – (Pawlowski et al. 2017a)

Chapter 5 expands on the notion of "knowing when one does not know" (**Goal 2**) and explores the use of deep generative modelling to detect whether a sample belongs to the seen training distribution or not. This task of outlier detection can be medically relevant as it can flag unusual regions or samples as items of interest and therefore guide a clinicians attention. The chapter starts with considering the detection of lesions caused by traumatic brain injuries from brain computed tomography images using variational autoencoders (Kingma and Welling 2014), as published in (Pawlowski et al. 2018). We show with experiments that in certain conditions it is possible to flag tumours as abnormal regions within the scan. Next, the chapter studies the application of the same technique to the detection of lesions on brain magnet resonance images, published in (Chen et al. 2018b). This application suffers from domain shift caused by differences in MR scanners and scanning protocols between healthy and unhealthy subjects. Lastly, the chapter applies normalising flows to the detection of cancerous tissue on histopathology images. The experiments, published in

(Pawlowski and Glocker 2021), suggest that cancerous tissue can be detected using normalising flows as density estimators but require careful choice of the outlier detection metric.

- **Pawlowski, N.** et al. (2018). "Unsupervised Lesion Detection in Brain CT using Bayesian Convolutional Autoencoders". In: *Medical Imaging with Deep Learning Abstract track* – (Pawlowski et al. 2018)
- Chen, X.*, **Pawlowski, N.***, Rajchl, M., Glocker, B., and Konukoglu, E. (2018b). "Deep generative models in the real-world: An open challenge from medical imaging". In: *arXiv preprint arXiv:1806.05452* – (Chen et al. 2018b)
- **Pawlowski, N.** and Glocker, B. (2021). "Abnormality Detection in Histopathology via Density Estimation with Normalising Flows". In: *Medical Imaging with Deep Learning Short Paper Track* – (Pawlowski and Glocker 2021)

Chapter 6 seeks to imbue neural networks with causal knowledge (**Goal 3**) and uses recent advances in deep generative modelling to build deep structural causal models (DSCMs) capable of modelling imaging and non-imaging data (Pawlowski et al. 2020). Our deep structural causal model framework allows for tractable counterfactual inference with high-dimensional data. Extensive experiments on synthetic toy datasets verify that DSCMs fulfil all three rungs of Pearl's causal hierarchy (association, intervention, and imagination) (Pearl 2019). Another case study shows that the framework is capable of modelling real-world data. We model brain MR images together with non-imaging information (age, sex, brain volume, and ventricle volume). The results show that the trained model is able to generate realistic high-fidelity counterfactual medical images that preserve details relevant to subject identity.

- **Pawlowski, N.***, Castro, D. C.*, and Glocker, B. (2020). "Deep Structural Causal Models for Tractable Counterfactual Inference". In: *Advances in Neural Information Processing Systems* – (Pawlowski et al. 2020)

Chapter 7 summarises the main contributions of this thesis and frames them in the context of the original research goals. It concludes with some open questions as well as known limitations of the presented work and offers directions for future research questions.

1.3 Publications and Research Context

Throughout my PhD, I was lucky enough to be able to collaborate with a wide range of groups and individuals from within Imperial as well as outside. Some of these collaborations were closely aligned with the goals of this thesis while others offered the opportunity to learn and explore novel areas of machine learning research. Various collaborations with Xiaoran Chen, a PhD student at ETH Zürich, explored the use of variational autoencoders for brain lesion detection (Chen et al. 2019b, 2021, 2018b). An internship at Google Health allowed me to work on safety aspects of der-

matology classifiers and the question of whether "a dermatology classifier knows what it doesn't know?" (Roy et al. 2021). Monteiro et al. (2020) introduced the use of low-rank multivariate Gaussian distributions for the modelling of spatially correlated uncertainty in segmentation networks. The development of the deep learning toolkit for medical image analysis (DLTK) (Pawlowski et al. 2017c) led to various collaborations using the reference implementations of the toolkit. One example collaboration is the winning entry to the 2017 BraTS challenge (Kamnitsas et al. 2017b) included in a longer overview paper (Bakas et al. 2017). A complete list of papers from collaborations not included in this thesis is given below:

- Chen, X., **Pawlowski, N.**, Glocker, B., and Konukoglu, E. (2021). "Normative ascent with local gaussians for unsupervised lesion detection". In: *Medical Image Analysis* – (Chen et al. 2021)
- Roy, A. G. et al. (2021). "Does Your Dermatology Classifier Know What It Doesn't Know? Detecting the Long-Tail of Unseen Conditions". In: *arXiv preprint arXiv:2104.03829* – (Roy et al. 2021)
- Monteiro, M., Folgoc, L. L., Castro, D. C. de, **Pawlowski, N.**, Marques, B., Kamnitsas, K., Wilk, M. van der, and Glocker, B. (2020). "Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncertainty". In: *Advances in Neural Information Processing* – (Monteiro et al. 2020)
- Charakorn, R., Thawornwattana, Y., Itthipuripat, S., **Pawlowski, N.**, Manoonpong, P., and Dilokthanakul, N. (2020). "An explicit local and global representation disentanglement framework with applications in deep clustering and unsupervised object detection". In: *arXiv preprint arXiv:2001.08957* – (Charakorn et al. 2020)
- Chen, X., **Pawlowski, N.**, Glocker, B., and Konukoglu, E. (2019b). "Unsupervised Lesion Detection with Locally Gaussian Approximation". In: *International Workshop on Machine Learning in Medical Imaging* – (Chen et al. 2019b)
- Dilokthanakul, N., Kaplanis, C., **Pawlowski, N.**, and Shanahan, M. (2019). "Feature Control as Intrinsic Motivation for Hierarchical Reinforcement Learning". In: *IEEE Transactions on Neural Networks and Learning Systems* – (Dilokthanakul et al. 2019)
- Antoniou, A., **Pawlowski, N.**, Turner, J., Owers, J., Mellor, J., and Crowley, E. J. (2019). "Meta-meta-learning for Neural Architecture Search through arXiv Descent". In: *Proceedings of the 2019 ACH Special Interest Group on Harry Queue Bovik (SIGBOVIK)*. Association for Computational Heresy – (Antoniou et al. 2019)
- Lee, M. C., Petersen, K., **Pawlowski, N.**, Glocker, B., and Schaap, M. (2019). "TETRIS: Template Transformer Networks for Image Segmentation with Shape Priors". In: *IEEE Transactions on Medical Imaging* – (Lee et al. 2019)
- Lee, M. C., Petersen, K., **Pawlowski, N.**, Glocker, B., and Schaap, M. (2019). "Template Trans-

former Networks for Image Segmentation". In: *Medical Imaging with Deep Learning Abstract track* – (Lee et al. 2019)

- Meng, Q., **Pawlowski, N.**, Rueckert, D., and Kainz, B. (2019). "Representation Disentanglement for Multi-task Learning with application to Fetal Ultrasound". In: *arXiv preprint arXiv:1908.07885* – (Meng et al. 2019)
- Bakas, S. et al. (2018). "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge". In: *arXiv preprint arXiv:1811.02629* – (Bakas et al. 2018)
- Rajchl, M., **Pawlowski, N.**, Rueckert, D., Matthews, P. M., and Glocker, B. (2018). "NeuroNet: Fast and Robust Reproduction of Multiple Brain Image Segmentation Pipelines". In: *Medical Imaging with Deep Learning* – (Rajchl et al. 2018)
- Valindria, V. V., **Pawlowski, N.**, Rajchl, M., Lavdas, I., Aboagye, E. O., Rockall, A. G., Rueckert, D., and Glocker, B. (2018). "Multi-Modal Learning from Unpaired Images: Application to Multi-Organ Segmentation in CT and MRI". in: *IEEE Winter Conference on Applications of Computer Vision (WACV)* – (Valindria et al. 2018)
- Bocklisch, T., Faulker, J., **Pawlowski, N.**, and Nichol, A. (2017). "Rasa: Open source language understanding and dialogue management". In: *arXiv preprint arXiv:1712.05181* – (Bocklisch et al. 2017)
- Goldsborough, P., **Pawlowski, N.**, Caicedo, J. C., Singh, S., and Carpenter, A. E. (2017). "CytoGAN: Generative Modeling of Cell Images". In: *NIPS Workshop on Machine Learning for Computational Biology* – (Goldsborough et al. 2017)
- **Pawlowski, N.**, Jaques, M., and Glocker, B. (2017b). "Efficient variational Bayesian neural network ensembles for outlier detection". In: *ICLR Workshop Track* – (Pawlowski et al. 2017b)

Chapter 2

Background

This chapter provides a short introduction to a few core topics that will reoccur throughout this thesis. The chapter begins by introducing some basic concepts of probabilistic modelling[1], including Bayesian inference and variational approaches. It then explores the use of deep learning techniques for imaging applications, in particular with a focus on medical images. Lastly, we touch upon the use of neural networks to build various types of deep generative models.

2.1 Probabilistic Modelling

Most of modern deep learning aims to model probabilities of observations of one or multiple variables. In the field of imaging applications this could be the probability of an image being observed as well as the conditional probability of an image presenting a specific property. The framework for all those applications can be explained with relatively simple scenarios.

Similar to modelling the probability of an image could be the modelling of the probability of another event occurring, e.g. the outcome of a coin toss or the height of a person from the overall population. To model this unconditional probability we usually chose a class of probability distributions that we believe to describe the observed data. In the case of modelling a coin toss this could be a Bernoulli distribution, while we might model the height of people as a Gaussian distribution.

Given a set of observations $X = x_1, \ldots, x_n$ corresponding to the outcomes of the coin tosses, we would optimise the likelihood of the assumed Bernoulli distribution that models the probability

[1] While this chapter introduces some concepts of probabilistic modelling, it still relies on a basic understanding of probabilities. We refer to a comprehensive introduction such as (Bishop 2006).

of the outcome of a single coin toss:

$$p(x) = \pi^x (1-\pi)^{(1-x)}\,, \tag{2.1}$$

where π is the probability of the coin toss coming out heads. To estimate the parameter π we can apply maximum likelihood estimation (MLE) which aims to find π such that the likelihood of all observations is maximised:

$$\arg\max_{\pi} \prod_{i=1}^{n} p(x_i)\,. \tag{2.2}$$

This is equivalent to maximising the log-likelihood, which often is easier to handle because the product turns into a summation:

$$\arg\max_{\pi} \sum_{i=1}^{n} \log p(x_i) = \sum_{i=1}^{n} x_i \log \pi + (1-x_i) \log 1 - \pi\,. \tag{2.3}$$

By optimising the log-likelihood one can recover the well known equation that π is the expectation of the observed values $\mathbb{E}_{x_i \sim X}[x_i]$:

$$\frac{d}{d\pi} \sum_{i=1}^{n} \log p(x_i) = \frac{d}{d\pi} \sum_{i=1}^{n} x_i \log \pi + (1-x_i) \log 1 - \pi\,, \tag{2.4}$$

$$= \frac{\sum_{i=1}^{n} x_i}{\pi} + \frac{\sum_{i=1}^{n} 1 - x_i}{1-\pi}\,, \tag{2.5}$$

$$\pi = \frac{1}{n} \sum_{i=1}^{n} x_i\,. \tag{2.6}$$

Modelling coin tosses is relatively straight forward. However, many practical applications of statistical modelling are slightly more complicated and study the relationship between multiple variables. As such, questions about what is visible in an image deal with conditional probabilities of the type $p(y|x)$, where y is the visible property and x is the image. One simple way of modelling those conditional probabilities is the use of linear regression, where we chose to model y as the linear combination of the input variables $x = (x_1, \ldots, x_D)^T$ or transformations of them, $\phi(x)$, as well as some zero-centred Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$:

$$y(x, \theta) = \theta_0 + \sum_{j=1}^{M-1} \theta_j \phi_j(x) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)\,, \tag{2.7}$$

where θ are the M learned parameters. We note that in this formulation the regression is linear only with respect to the parameters θ but not the inputs x. For ease of notation we define a dummy function $\phi_0(x) = 1$, so that we can write:

$$y(x, w) = \sum_{j=0}^{M-1} \theta_j \phi_j(x) = \theta^T \phi(x) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \,, \tag{2.8}$$

where $\theta = (\theta_0, \ldots, \theta_{M-1})^T$ and $\phi = (\phi_0, \ldots, \phi_{M-1})^T$. The choice of ϕ depends on the problem at hand and many machine learning applications naturally apply some forms of preprocessing or feature extraction to the original variables x. A simple example could be polynomial regression in which $\phi_j(x) = x^j$. Linear regression can also be estimated using the maximum likelihood principle. The Gaussian noise model provides a Gaussian likelihood for the observations:

$$p(y \,|\, x, \theta) = \mathcal{N}(y; \theta^T \phi(x), \sigma^2) \,, \tag{2.9}$$

which can be optimised for θ and yields the solution

$$\arg\max_{\theta} p(Y \,|\, X, \theta) = \left(\phi^T(X)\phi(X)\right)^{-1} \phi^T(X) Y \,. \tag{2.10}$$

2.1.1 Bayesian Inference

So far we have considered the modelling of the probabilities of observed variables. However, how do we chose the exact model to use? Which functions ϕ should be included? How many parameters should the model have? Those questions cannot simply be answered by optimising the likelihood of the observed data as it could lead to overly complex models and overfitting. One potential solution to this problem is the use of held-out tuning data. However, this might be costly to acquire. Instead we treat our model in a Bayesian way in which we consider the distribution over the parameter θ as well.

First, we need to capture our assumptions about the parameters θ in a prior distribution $p(\theta)$. For simplicity, we choose a Gaussian prior which also is the corresponding conjugate prior to our previously used likelihood $p(y \,|\, x, \theta)$:

$$p(\theta) = \mathcal{N}(\theta; \mu_0, \Sigma_0) \,, \tag{2.11}$$

where μ_0 is the prior mean and Σ_0 is the corresponding prior covariance of the parameters.

We are interested in finding the posterior distribution $p(\theta \,|\, \mathcal{D})$ over the parameters θ given the observed data $\mathcal{D} = (x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})$. Following Bayes' rule:

$$p(\theta \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \theta) p(\theta)}{p(\mathcal{D})} = p(Y \,|\, X, \theta) p(\theta) / Z \,, \tag{2.12}$$

we can express the posterior $p(\theta|\mathcal{D})$ in terms of the likelihood $p(Y|X,\theta)$, a prior $p(\theta)$ and a normalising constant Z. Because both those distributions are Gaussian we can analytically find the solution as:

$$p(\theta|\mathcal{D}) = \mathcal{N}(\theta; \mu_N, \Sigma_N)\,, \tag{2.13}$$

with

$$\mu_N = \Sigma_N(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}\phi^T Y) \tag{2.14}$$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2}\phi^T\phi\,. \tag{2.15}$$

One special case is the maximum a posteriori (MAP) solution, which in the case of Gaussian distributions coincides with the mean $\theta_{MAP} = \mu_N$. To use the posterior weight distribution $p(\theta|\mathcal{D})$ to make predictions for a new datapoint x' we marginalise over the parameters:

$$p(y|X) = \int p(y|x',\theta)p(\theta|\mathcal{D})d\theta\,, \tag{2.16}$$

which again can be solved by the manipulation of Gaussian distributions for which we refer to (Bishop 2006).

2.1.2 Variational Inference

The previous subsection derived the solution to the posterior parameter distribution with conjugate distributions. However, let us consider the case of Bayesian Logistic regression. Instead of working with a Gaussian likelihood for $p(y|x,\theta)$, we are now using a Bernoulli likelihood to model binary outcomes. For this model there is no convenient conjugate prior to the posterior distribution of the model's parameters that we can use. Therefore we cannot easily find the posterior parameter distribution analytically. Instead we can simplify the problem by approximating the true posterior distribution with a simpler distribution. One potential method is called variational inference or variational Bayes.

Suppose we observe some data $\mathcal{D} = (x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})$ that we want to model with some likelihood $p(y|x,\theta)$, a prior on the parameters $p(\theta)$, and a simpler approximate distribution $q_\varphi(\theta)$ with some variational parameters φ. We can then derive a bound on the likelihood of the observed data as follows:

$$\log p(y|x) = \log \int p(y|x,\theta)p(\theta|\mathcal{D})d\theta \tag{2.17}$$

$$= \log \int \frac{q_\varphi(\theta)}{q_\varphi(\theta)} p(y|x,\theta)p(\theta|\mathcal{D})d\theta \tag{2.18}$$

$$= \log \int q_\varphi(\theta) \frac{p(\theta|\mathcal{D})}{q_\varphi(\theta)} p(y|x,\theta)d\theta \tag{2.19}$$

$$\geq \int q_\varphi(\theta) \log \left[\frac{p(\theta|\mathcal{D})}{q_\varphi(\theta)} p(y|x,\theta) \right] d\theta \tag{2.20}$$

$$\geq \mathbb{E}_{q_\varphi(\theta)}[p(y|x,\theta)] - \mathrm{KL}(q_\varphi(\theta)\|p(\theta|\mathcal{D}))\,, \tag{2.21}$$

where $\mathbb{E}_{q_\varphi(\theta)}[p(y|x,\theta)]$ is the expected likelihood of the observed data under the approximate posterior and $\mathrm{KL}(q_\varphi(\theta)\|p(\theta|\mathcal{D}))$ is the Kullback-Leibler divergence between the approximate posterior $q_\varphi(\theta)$ and the true posterior $p(\theta|\mathcal{D})$. The step between Eq. (2.19) and Eq. (2.20) applied the Jensen inequality (Jensen 1906). The predictive distribution can be approximated using Monte-Carlo sampling:

$$p(y'|x') = \mathbb{E}_{q_\varphi(\theta)}[p(y'|x',\theta)] \approx \frac{1}{K} \sum_{i=1}^{K} p(y'|x',\theta^{(i)})\ ; \theta^{(i)} \sim q_\varphi(\theta)\,. \tag{2.22}$$

We will find applications of variational inference throughout this thesis: applied to the estimation of the weights of neural networks in Chapter 4 and when used in the form of variationel autoencoders (see Section 2.3.2) in Chapter 5 and Chapter 6.

2.2 Deep Learning for Imaging

Predictions from images can be modelled similarly to the regression examples in the previous chapter. However, the spatial structure of images, represented as grids of pixels or voxels, as well as their high dimensionality[2] mean that simple linear models only achieve mediocre modelling performances. Instead, progress has been achieved by applying convolutional neural networks (CNNs) and most recently self-attention based architectures to imaging related tasks.

The curation of clean and easily accessible datasets facilitated the progress in developing these methods. The development of digit recognition systems on the MNIST dataset (LeCun et al. 1998a) in the 1990s led to one of the first popular CNNs (LeCun et al. 1998b) and the MNIST dataset is still used today as a testbed for early ideas. A submission to the ImageNet competition (Deng et al. 2009) marked a breakthrough for CNN-based approaches with AlexNet (Krizhevsky et al. 2012) which saw increasingly complicated network architectures developed year on year (He et al. 2016a;

[2] Images often contain thousand to millions of pixels.

Huang et al. 2017b; Simonyan and Zisserman 2015). Most recently, self-attention, with its ability to model long-range relations, has been adapted from its use in natural language processing (Vaswani et al. 2017) to imaging problems (Dosovitskiy et al. 2020) and its popularity is growing.

Throughout this thesis various common imaging datasets are used. Specifically, Chapters 3, 4 and 6 use the MNIST (LeCun et al. 1998a) dataset and Chapter 4 uses the CIFAR-10 (Krizhevsky 2009) dataset. MNIST contains images of handwritten digits with an approximately uniform distribution over all 10 digits. It consists of 60.000 training images and 10.000 test images. CIFAR-10 consists of 50.000 training and 10.000 test images, totalling 6.000 images across 10 different classes. Other less common datasets are described in the experimental sections of the following chapters.

2.2.1 Medical Images

A lot of regular computer vision research works on problems relating to natural images from clean and well defined datasets. However, the field of medical image analysis covers a wide range of imaging modalities from two dimensional images such as ultrasound or histopathology to multi-modal and temporal magnetic resonance image (MRI) sequences. Each of those modalities comes with its own intricacies and often requires expert knowledge to be correctly handled. Additionally, medical imaging datasets often only contain a limited amount of images due to privacy constraints, acquisition, or labelling costs. Naive handling of medical imaging datasets often leads to suboptimal results and many tools have been developed to deal with those special requirements – most noteworthy in the space of deep learning are NiftyNet (Gibson et al. 2018) and DLTK (Pawlowski et al. 2017c) that might have been the first general purpose software packages designed to ease the application of deep learning algorithms to medical images. Arguably, those packages built the foundation for later developments such as MONAI (MONAI Consortium 2020) or torchIO (Pérez-García et al. 2021). These packages deal with various common problems:

Size of images: Some image modalities such as computed tomography (CT) can produce very detailed images of biological structures of resolutions below $1mm^3$ that lead to volumes with more than $512 \cdot 512 \cdot 128 = 33,554,432$ voxels. Training state-of-the-art deep learning algorithms on those volumes has memory requirements that go beyond many recent hardware accelerators, mostly GPUs. Instead, one common approach is to break the task down into subtasks that can be solved on sub-volumes ("patches"). Image segmentation and object detection are tasks that lend themselves particularly well to this approach, because ground truth labels are available on a local (i.e. pixel) level rather than only a global (i.e. image) level. The splitting of an image with only global labels into subtasks would introduce label noise as there is no information as to which part of the image causes the label.

Class imbalances: Given the acquisition and labelling costs of medical images, datasets are often of small size. This also often leads to fewer samples for some pathologies – especially if they are rare. Similarly, some anatomical or pathological structures are by nature smaller than others. Both those properties can lead to class-imbalances. A naive but still useful solution to this problem is to show rare or small structures more often than they naturally occur to balance the class probabilities. This approach works well in conjunction with the patch-based training to handle the large size of the images.

Imaging coordinates: Medical images are acquired in a way to help clinicians diagnose and treat a patient's medical condition. As such the positioning and orientation of the images can hugely vary due to other conditions that might prevent the standard procedure to not harm a patient further. Additionally, different scanners or scanning protocols can lead to different resolutions – that might also vary across spatial dimensions. Common preprocessing steps include mapping the original images into a common space. Different resolutions (or voxel spacings) are often resampled into images with common isotropic spacing using classic methods such as bilinear sampling. The field of image registration deals with the problem of ensuring common orientation. Here, an image of the dataset or a set of reference images for the shown anatomical structure (a so-called atlas) is used as a reference, e.g. brain images are commonly registered to the MNI atlas (Fonov et al. 2011). The other images are then deformed to increase the similarity between the deformed and the reference image. The deformations are often constrained to be affine or rigid.

Image contrast: Different image modalities exhibit different contrasts that highlight different structures or properties depending on the underlying imaging physics. While the voxel intensities of some modalities do not have interpretable units associated with them, others such as quantitative MR or CT can be directly interpreted. CT intensities are usually measured in Hounsfield units that correspond to the radiodensity of the measured volume. As such, one can deduct possible substances or tissue types from the measured radiodensity. On the other hand, non-quantitative MRI might require sophisticated intensity normalisation routines.

The following chapters of this thesis will make use of various of these common preprocessing approaches when dealing with medical images. Specifically, we handle histopathology images in Section 3.1 and Section 5.3, brain CT images in Section 5.2, brain MRI in Section 3.2, Section 5.1 and Chapter 6.

2.3 Deep Generative Modelling for Imaging

In Section 2.1 we touched upon the unconditional modelling of a data distribution when we considered the coin toss example. The resulting Bernoulli distribution can be thought of as a generative model of the data: we can sample from it to simulate throwing a coin and will obtain a comparable

result. Now we want to apply similar techniques to building generative models of images. Simple approaches such as modelling small images as multivariate Gaussian distributions are capable of achieving baseline results. However, those approaches quickly reach their limits with increasing image sizes, multiple channels (e.g. colours), and the fact that they might not capture multiple modes of the image space. Advances in deep learning have not only led to progress in predictive modelling (see Section 2.2) but also improved generative models of high-dimensional data.

2.3.1 Normalising Flows

One of the most direct approaches of modelling a target distribution is through the use of invertible transformations on random variables and the application of the change of variable formula for probability densities:

$$p(x) = p(z)|_{z=g^{-1}(x)} \cdot |\det \nabla g^{-1}(x)| = \frac{p(z)}{|\det \nabla g(z)|}\bigg|_{z=g^{-1}(x)}, \tag{2.23}$$

where $z \in \mathcal{Z} \subseteq \mathbb{R}^D$ is a local latent variable to $x \in \mathcal{X} \subseteq \mathbb{R}^D$, $p(z)$ is the probability density of z, and we assume $g : \mathcal{Z} \to \mathcal{X}$ to be a diffeomorphic transformation from z to x.

In machine learning, g is a parametrised function and the base density $p(z)$ is often assumed to be a simple and easy to evaluate distribution. This allows the fitting of the parameters of g to maximise the likelihood of observations of x that are mapped back into the space of the base distribution $p(z)$ via the inverse transformation g^{-1}. This type of model that maps a complex distributions to a base distribution is called a normalising flow.

However, the calculation of the Jacobian determinant $\det \nabla g$ is generally computationally costly as it scales cubically in the data dimensionality D, which makes it unsuitable for high-dimensional data. To overcome this limitation, it is possible to design transformations g in a way that allow for more computationally efficient calculation of the determinant due to specific structures of the Jacobian ∇g, e.g. diagonal, triangular or block matrices offer easier computation of determinants.

Restricting the form of the Jacobian also restricts the flexibility of the functions g can model. Rather than relying on a single transformation to map from x to z we can stack multiple learnable transformations $g_1, \ldots, g_L$ as compositions of diffeomorphisms are also diffeomorphic. Analogously to neural networks, this allows for construction of more complex transformations from simple components with tractable Jacobian determinant:

$$x = g(z) = (g_L \circ \ldots g_1)(z) \tag{2.24}$$

Figure 2.1 exemplifies the likelihood of a bivariate normal transformed by the application of multiple planar flows (Rezende and Mohamed 2015). The left-most graphic shows the likelihood of bivariate

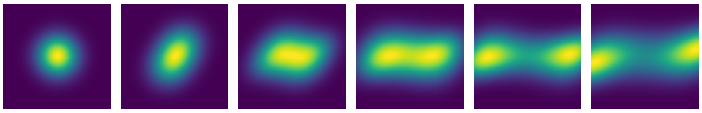


Figure 2.1: Visualisation of the likelihood of a bivariate Normal distribution (left) being transformed by the repeated application of planar flows (Rezende and Mohamed 2015). A single flow only slightly changes the shape of the base distribution, whereas the stacking of multiple flows changes the normal distribution to a bimodal distribution (right).

Normal base distribution, while the other pictures show the result of applying a planar flow to the previous distribution – by stacking multiple flows one can transform the unimodal Normal distribution into a more complex bimodal distribution in the graphics on the right.

There exists an ever growing amount of literature proposing novel functional forms as elementary transformations. We refer to a recent review on normalising flows (Papamakarios et al. 2019) for more details. In this thesis, we apply normalising flows to the task of outlier detection in Section 5.3 and causal generative modelling in Chapter 6.

2.3.2 Variational Autoencoders

Compared to normalising flows, variational autoencoders (VAEs) (Kingma and Welling 2014; Rezende et al. 2014) allow for less restrictive forms of models and therefore lose the ability to directly calculate the likelihood of observations x. VAEs assume a latent variable model in which a sample x of our distribution of interest is generated by first sampling a value z from some prior distribution $p(z)$ over the local latent variable and then sampling from the conditional distribution $p(x\,|\,z)$. To allow learning of this model, we assume both the prior $p(z)$ and the conditional likelihood $p(x\,|\,z)$ to be differentiable and of parametric form $p_{\theta^*}(z)$ and $p_{\theta^*}(x\,|\,z)$ where θ^* are the true but unknown parameters.

As seen previously in Section 2.1.2, those type of models only have analytical solutions in few special cases. Instead, VAEs approximate the true data generating process by adapting the variational Bayes algorithm. The algorithm optimises the variational lower bound

$$\log p_\theta(x) \geq \mathbb{E}_{q_\varphi(z|x)}[\log p_\theta(x\,|\,z)] - \mathrm{KL}(q_\varphi(z\,|\,x)\|\,p_\theta(z))\,, \tag{2.25}$$

where $q_\varphi(z\,|\,x)$ is an amortised variational approximation. Optimisation of this lower bound using the usual naive Monte Carlo estimate of the gradient:

$$\nabla_\varphi \mathbb{E}_{q_\varphi(z)}[f(z)] = \mathbb{E}_{q_\varphi(z)}[f(z)\nabla \log q_\varphi(z)]\,, \tag{2.26}$$

is known to be of high variance (Paisley et al. 2012) and can therefore cause issues during optimisation. Instead, it is possible to reparametrise the random variable $z \sim q_\varphi(z\,|\,x)$ using a diffeomor-

phic transformation[3] $g_\varphi(\epsilon, x)$ and some auxiliary noise variable $\epsilon \sim p(\epsilon)$. We can now replace z by $g_\varphi(\epsilon, x)$ and approximate expectations over distributions of $q_\varphi(z|x)$ as expectations over $p(\epsilon)$:

$$\mathbb{E}_{q_\varphi(z)}[f(z)] = \mathbb{E}_{p(\epsilon)}[f(g_\varphi(\epsilon, x))] \approx \frac{1}{L}\sum_{l=1}^{L} f\big(g_\varphi(\epsilon^{(l)}, x)\big);\ \epsilon^{(l)} \sim p(\epsilon)\,, \tag{2.27}$$

which also allows for simple calculation of gradients with respect to φ. This reparameterisation is often referred to as the *reparameterisation trick*. In practice, the prior $p(z)$ and variational posterior $q(z|x)$ are often chosen to be independent Normal distributions. Then, g becomes an affine transformation of a unit Normal $p(\epsilon) = \mathcal{N}(0, 1)$. We apply VAEs in Chapter 5 to detect outliers in medical images and in Chapter 6 to build causal generative models.

[3] This uses the same principle as normalising flows.

Chapter 3

Extracting Information from Small Image Regions

This chapter is based on the following publications:

(a) **Pawlowski, N.**, Bhooshan, S., Ballas, N., Ciompi, F., Glocker, B., and Drozdzal, M. (2019). "Needles in Haystacks: On Classifying Tiny Objects in Large Images". In: *arXiv preprint arXiv:1908.06037* – (Pawlowski et al. 2019)

(b) **Pawlowski, N.** and Glocker, B. (2019). "Is Texture Predictive for Age and Sex in Brain MRI?". In: *Medical Imaging with Deep Learning Abstract track* – (Pawlowski and Glocker 2019)

The work on the paper (Pawlowski et al. 2019) was conducted during a research internship at Facebook AI Research.

The code for all experiments is available at https://github.com/facebookresearch/Needles-in-Haystacks and https://github.com/pawni/MedicalBagNet.

Image analysis has largely advanced due to the application of deep learning techniques to a wide range of problems, from image classification to object detection and image captioning. Many of these works focus on natural images with clearly visible objects that are classified, detected or described. However, specialised application domains such as medical imaging (see Section 2.2.1) harbour different properties of the objects or structures of interest. In one extreme, the task of interest might require the analysis of multiple instances of similar, potentially tiny, regions within given images (a field known as multiple instance learning), whereas in other extremes, the task requires the analysis of large regions or the relation between the parts of such a region. An example for such a task is the grading of whole-slide histopathology images. The grading describes the

severity of the cancer that might be present on the images and is dependent on the total amount of individual cancerous cells.

This chapter investigates the capability of convolutional neural networks to extract information from small parts of images. The first part of the chapter tackles the problem of classifying big images where only small regions are informative about the overall class label[1]. This section systematically investigates how different choices in the design of the neural network architecture and its training impact the learning capabilities in this special scenario. The study is conducted on a synthetic dataset based on MNIST (LeCun et al. 1998a) that allows for very controlled image properties as well as a medical dataset that is derived from the real-world CAMELYON (Bejnordi et al. 2017) dataset containing histopathology images. We find that the naive application of CNNs fails in the extreme setting where we want to classify tiny objects in large images due to the low signal to noise ratio. Furthermore, we propose some countermeasures to this behaviour but do not offer a complete solution in the case, when there is no localisation information available. One of these countermeasures builds on restricting the receptive field of the neural network to ensure that the network focuses on relevant features rather than spurious ones.

The second part of this chapter studies how the restriction of the receptive field of a network impacts the performance and interpretability of a simple medical imaging task: predicting the age and biological sex from three dimensional brain MRI scans from the CamCAN dataset (Taylor et al. 2017). This approach allows us to ask questions like *"Is texture information in brain MRI scans predictive for the age and sex?"* as well as *"Do localised predictions due to restricted receptive fields improve the interpretability of neural network predictions?"*. We find that even the texture of small patches of MRI scans contain information about the age and biological sex of subjects. However, the localised predictions turn out to be hard to interpret and do not immediately improve human understanding.

3.1 Classifying small regions in big images

Convolutional Neural Networks (CNNs) are the current state-of-the-art approach for image classification (He et al. 2016a; Huang et al. 2017b; Krizhevsky et al. 2012; Simonyan and Zisserman 2015). The goal of image classification is to assign an image-level label to an image. Typically, it is assumed that an object (or concept) that correlates with the label is clearly visible and occupies a *significant* portion of the image (Deng et al. 2009; Krizhevsky 2009; LeCun et al. 1998a). Yet, in a variety of real-life applications, such as medical image or hyperspectral image analysis, only a small portion of the input correlates with the label, resulting in low signal-to-noise ratio. We define this input image signal-to-noise ratio as Object to Image (O2I) ratio. The O2I ratio range for three real-life datasets is depicted in Figure 3.1. As can be seen, there exists a distribution

[1]Specifically, we study the extreme scenario where objects occupy less than 1% of the area of an image.

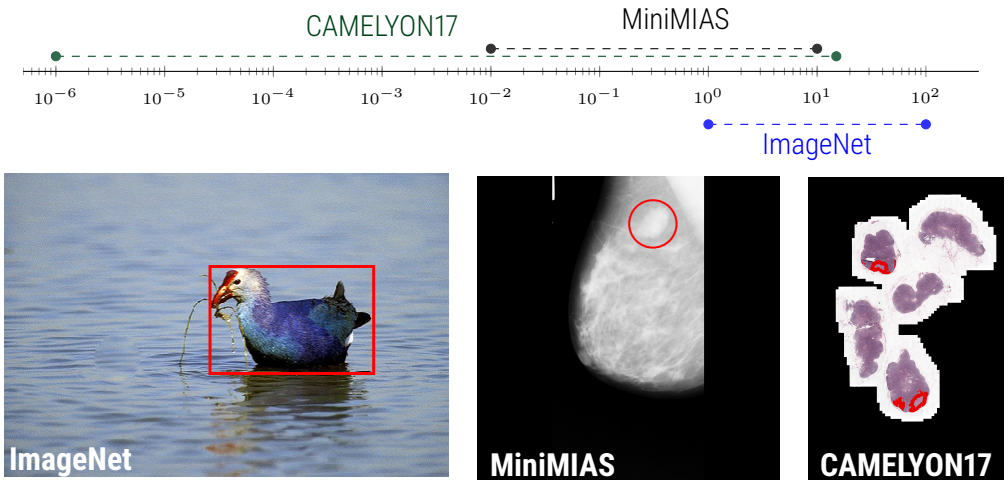


Figure 3.1: Range of Object to Image (O2I) ratios [%] for two medical imaging datasets (CAMELYON17 (Ehteshami Bejnordi et al. 2017) and MiniMIAS (Suckling 1994)) as well as one standard computer vision classification dataset (ImageNet (Deng et al. 2009)). The ratio is defined as $\text{O2I} = \frac{A_{object}}{A_{image}}$, where A_{object} and A_{image} denote the area of the object and the image, respectively. Together with O2I range, we display examples of images jointly with the object area A_{object} (in red).

shift between standard classification benchmarks and domain specific datasets. For instance, in the ImageNet dataset (Deng et al. 2009) objects fill at least 1% of the entire image, while in histopathology slices (Ehteshami Bejnordi et al. 2017) cancer cells can occupy as little as $10^{-6}\%$ of the whole image.

Recent works have studied CNNs under different noise scenarios, either by performing random input-to-label experiments (Arpit et al. 2017; Zhang et al. 2017) or by directly working with noisy annotations (Han et al. 2018; Jiang et al. 2018; Mahajan et al. 2018). While, it has been shown that large amounts of label-corruption noise hinders the CNNs generalization (Arpit et al. 2017; Zhang et al. 2017), it has been further demonstrated that CNNs can mitigate this label-corruption noise by increasing the size of training data (Mahajan et al. 2018), tuning the optimizer hyperparameters (Jastrzkebski et al. 2017) or weighting input training samples (Han et al. 2018; Jiang et al. 2018). However, all these works focus on input-to-label corruption and do not consider the case of noiseless input-to-label assignments with low and very low O2I ratios.

In this paper, we build a novel testbed allowing us to specifically study the performance of CNNs when applied to tiny object *classification* and to investigate the interplay between input signal-to-noise ratio and model generalization. We create two synthetic datasets inspired by the children's puzzle book *Where's Wally?* (Handford 1987). The first dataset is derived from MNIST digits and allows us to produce a relatively large number of datapoints with explicit control of the O2I ratio. The second dataset is extracted from histopathology imaging (Ehteshami Bejnordi et al. 2017) where we crop images around lesions and obtain a small number of datapoints with an approximate control of the O2I ratio. To the best of our knowledge these datasets are the first ones designed to

explicitly stress-test the behaviour of the CNNs in the low input image signal-to-noise ratio.

We develop a classification framework, based on CNNs, and analyze the effects of different factors affecting the model optimization and generalization. Throughout an empirical evaluation, we make the following observations:

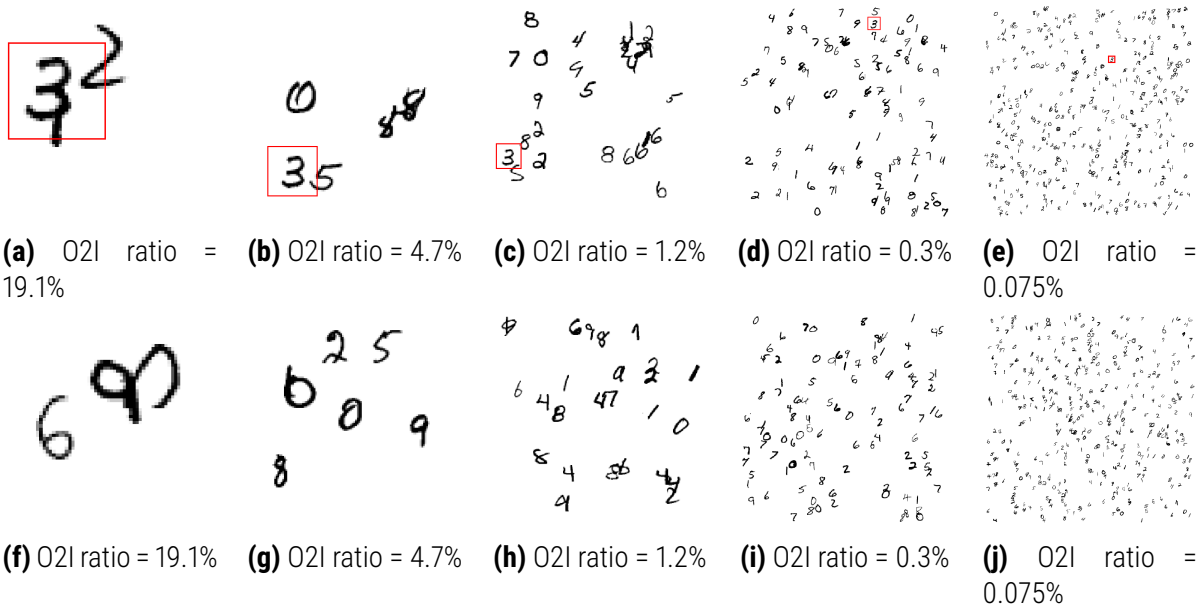
- In our experimental setup, models can be *trained in low O2I regime without using any pixel-level annotations and generalize* if we leverage enough training data. However, *the amount of training data required for the model to generalize scales rapidly with the inverse of the O2I ratio*. When considering datasets with fixed size, we observe an *O2I ratio limit* in which all tested scenarios fail to exceed random performance.
- We empirically observe that *higher capacity models show better generalization*. We hypothesize that high capacity models learn to ignore the input noise structure and, as result, achieve satisfactory generalization.
- We confirm the importance of model inductive bias — in particular, the *model's receptive field size*. Our results suggest that different pooling operations exhibit similar performance, for larger O2I ratios; however, for very small O2I ratios, the type of *pooling operation affects the optimization ease*, with max-pooling leading to fastest convergence.

3.1.1 Datasets: Is there a Wally in an image?

To study the optimization and generalization properties of CNNs, we build two datasets: one derived from the MNIST (LeCun et al. 1998a) dataset and another one produced by cropping large resolution images from the CAMELYON dataset (Ehteshami Bejnordi et al. 2017). Each dataset allows to evaluate the behaviour of a CNN-based binary classifier when altering different data-related factors of variation such as dataset size, object size, image resolution and class balance. In this subsection, we describe the data generation process.

3.1.1.1 Digits: needle MNIST (nMNIST)

Inspired by the cluttered MNIST dataset (Ba et al. 2015), we introduce a scaled up, large resolution cluttered MNIST dataset, suitable for binary image classification. In this dataset, images are obtained by randomly placing a varying number of MNIST digits on a large resolution image canvas. We keep the original 28×28 pixels digit resolution and control the O2I ratio by increasing the resolution of the canvas. Alternatively, we could fix canvas image resolution and downscale MNIST digits; however, downscaling might reduce the object quality. As result, we obtain the following O2I ratios $\{19.1, 4.8, 1.2, 0.3, \text{ and } 0.075\}\%$ that correspond to the following canvas resolutions 64×64, 128×128, 256×256, 512×512, and 1024×1024 pixels, respectively. As the object of interest, we select the digit 3. All positive images contain exactly one instance of the digit 3 randomly placed within the image canvas, while negative instances do not contain any instance. We also in-



(a) O2I ratio = 19.1% **(b)** O2I ratio = 4.7% **(c)** O2I ratio = 1.2% **(d)** O2I ratio = 0.3% **(e)** O2I ratio = 0.075%

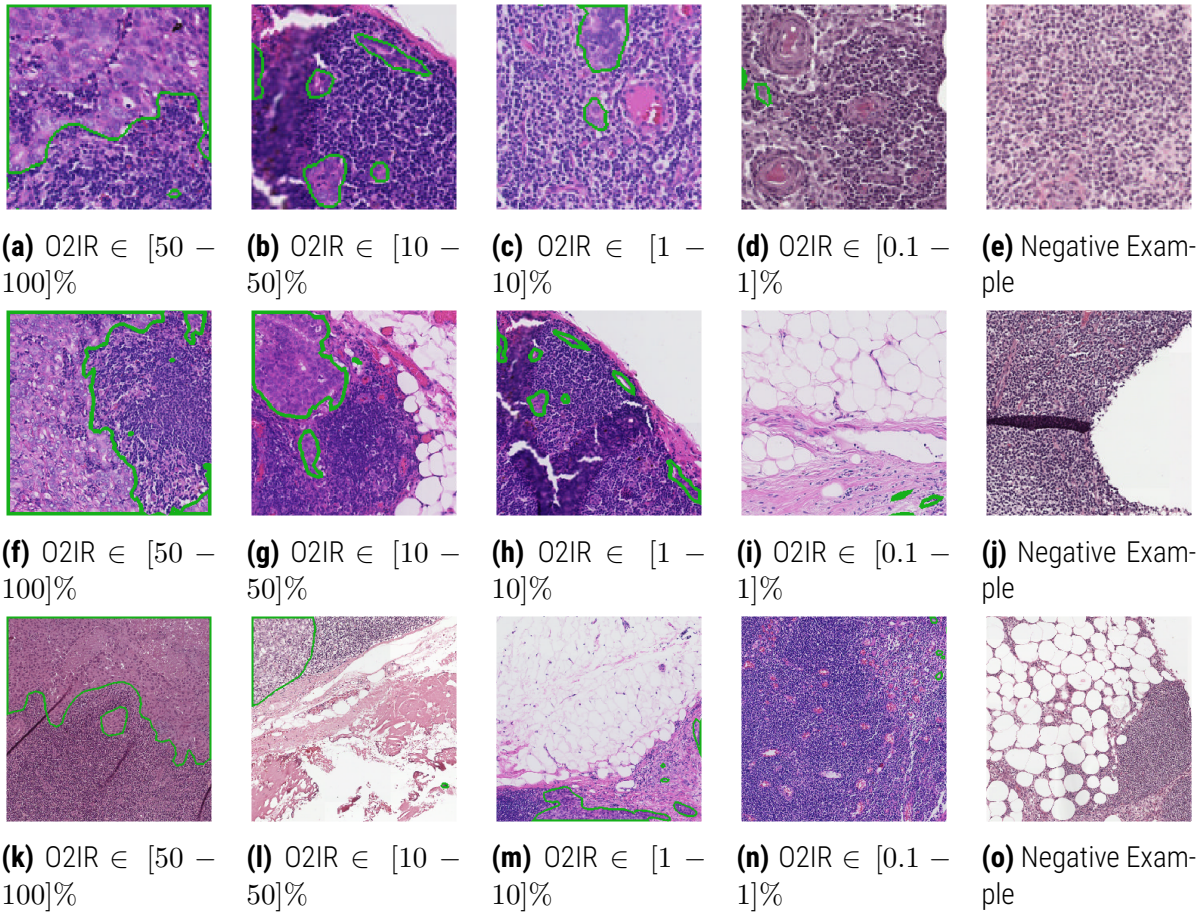
(f) O2I ratio = 19.1% **(g)** O2I ratio = 4.7% **(h)** O2I ratio = 1.2% **(i)** O2I ratio = 0.3% **(j)** O2I ratio = 0.075%

Figure 3.2: Example images from our MNIST dataset with different O2I ratios. Top row images represent positive examples — digit 3 is present (marked with red rectangle), while bottom row depicts negative images. Note that for visualization purposes all images have been rescaled to the same resolution.

clude distractors (clutter digits): any MNIST digit image sampled with replacement from a set of labels $\{0, 1, 2, 4, 5, 6, 7, 8, 9\}$. We maintain approximately constant clutter density over different O2I ratios. Thus, the following O2I ratios $\{19.1, 4.8, 1.2, 0.3, \text{and } 0.075\}\%$ correspond to $2, 5, 25, 100,$ and 400 clutter objects, respectively. For each value of O2I ratio, we obtain $11276, 1972, 4040$ of training, validation and test images. We obtain those numbers by using the original MNIST data, we use every digit 3 only once to generate positive images and we balance the dataset with negative images. We present both positive and negative samples for different O2I ratios in Fig. 3.2.

3.1.1.2 Histopathology: needle CAMELYON (nCAMELYON)

The CAMELYON (Ehteshami Bejnordi et al. 2017) dataset contains gigapixel histopathology images with pixel-level lesion annotations from 5 different acquisition sites. The needle CAMELYON (nCAMELYON) is designed as a derived binary classification task: *Are there breast cancer metastases in the image or not?*. We rely on the pixel-level annotations within CAMELYON to extract samples for nCAMELYON with controlled O2I ratios. We use downsampling level 3 from the original whole slide image using the MultiResolution Image interface released with the original CAMELYON dataset. Namely, we generate datasets for O2I ratios in the range of $(100 - 50)\%$, $(50 - 10)\%$, $(10 - 1)\%$, and $(1 - 0.1)\%$, and we crop different image resolutions with the size of $128 \times 128, 256 \times 256$, and 512×512 pixels. This results in training sets of about $20 - 235$ unique lesions per dataset configuration (see Table 3.1 and Table 3.2 for a detailed list of dataset sizes). More precisely, for positive examples, we identify contiguous regions within the annotations, and take 50 random crops around each contiguous region ensuring that the full contiguous region is



(a) O2IR $\in [50 - 100]\%$ **(b)** O2IR $\in [10 - 50]\%$ **(c)** O2IR $\in [1 - 10]\%$ **(d)** O2IR $\in [0.1 - 1]\%$ **(e)** Negative Example

(f) O2IR $\in [50 - 100]\%$ **(g)** O2IR $\in [10 - 50]\%$ **(h)** O2IR $\in [1 - 10]\%$ **(i)** O2IR $\in [0.1 - 1]\%$ **(j)** Negative Example

(k) O2IR $\in [50 - 100]\%$ **(l)** O2IR $\in [10 - 50]\%$ **(m)** O2IR $\in [1 - 10]\%$ **(n)** O2IR $\in [0.1 - 1]\%$ **(o)** Negative Example

Figure 3.3: Example images from our CAMELYON dataset for different crop sizes and O2I ratios. We show crops with size 128×128, 256×256, and 512×512 in the top, middle, and bottom row, respectively. The green outlines show the cancerous regions. Note that for visualization purposes all images have been rescaled to same resolution.

inside the crop, and total number of lesion pixels inside the crop are in the desired O2I ratio. The negative crops are taken from healthy images randomly filtering for images that are mostly background using a heuristic that the average green pixel value in the crop is below 200. Since the CAMELYON dataset contains images acquired by 5 different centers, we split training, validation and test sets center-wise to avoid any contamination of data across the three sets. All crops coming from center 3 are part of the validation set, and all crops coming from center 4 are part of the test set. We ensure the class balance by sampling an equal amount of positive and negative crops. Once the crops were extracted, no pixel-wise information is used during training. Figure 3.3 shows examples of images from nCAMELYON dataset, Table 3.1 presents number of unique lesions in each dataset, and Table 3.2 depicts number of dataset images stratified for image resolution and O2I ratios. Because center 3 does not contain lesions of suitable size for crops of with resolution 128×128 and O2I ratio $(50 - 100)\%$, we do not include those training runs in our analysis.

Table 3.1: Number of unique lesions extracted for each set of the nCAMELYON data for different O2I ratios and crop sizes.

Crop Size	128			256			512		
O2I ratio	Train	Val	Test	Train	Val	Test	Train	Val	Test
(50 - 100)%	20	0	8	27	2	13	23	5	13
(10 - 50)%	84	12	16	101	16	15	68	15	17
(1 - 10)%	176	17	18	227	17	18	235	21	15
(0.1 - 1)%	33	5	5	93	16	9	173	20	11

Table 3.2: Number of crops extracted for each set of the nCAMELYON data for different O2I ratios and crop sizes. Note that the dataset is balanced (e. g. 50% are positive images and 50% are negative). Moreover, for positive images we have relatively small number of unique cancer regions as noted in Table 3.1.

Crop Size	128			256			512		
O2I ratio	Train	Val	Test	Train	Val	Test	Train	Val	Test
(50 - 100)%	1000	0	400	1350	100	650	1150	250	650
(10 - 50)%	4200	600	800	5050	800	750	3400	750	850
(1 - 10)%	8686	850	900	11270	850	900	11750	1050	750
(0.1 - 1)%	1488	247	207	4255	800	450	8312	965	550
negative	19608	6000	6100	19595	6000	6100	19574	6000	6100

3.1.2 Models

Our classification pipelines follow the BagNet (Brendel and Bethge 2019) backbone, which allows us to explicitly control for the network receptive field size. Figure 3.4 shows a schematic of our approach. As can be seen, the pipelines are built of three components: (1) topological embedding extractor in which we can control for embedding receptive field, (2) global pooling operation that converts the topological embedding into a global embedding, and (3) a binary classifier that receives the global embedding and outputs binary classification probabilities. By varying the embedding extractor and the pooling operation, we test a set of 48 different architectures.

3.1.2.1 Topological embedding extractor

The extractor takes as input an image $\mathbf{I}$ of size $[w_{img} \times h_{img} \times c_{img}]$ and outputs a topological embedding $\mathbf{E}^t$ of shape $[w_{enc} \times h_{enc} \times c_{enc}]$, where $w_.$, $h_.$, and $c_.$ represent width, height and number of channels. Due to the relatively large image sizes, we train the pipeline with small batch sizes and, thus, we replace BagNet-used BatchNorm operation (Ioffe and Szegedy 2015) with Instance Normalization (Ulyanov et al. 2016). In our experiments, we test 12 different extractor architectures

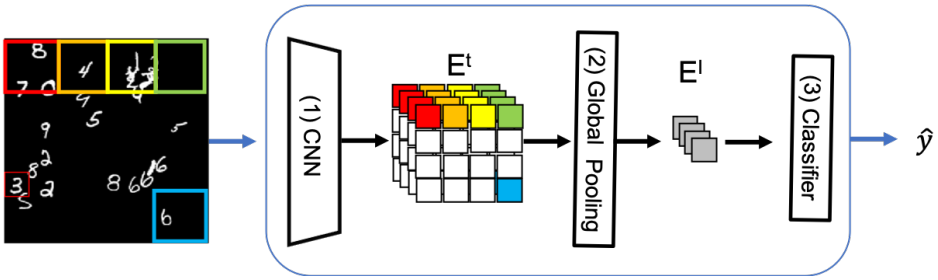


Figure 3.4: Pipeline. Our pipeline is built of three components: (1) a CNN extracting topological embedding, (2) a global pooling operation and (3) a binary classifier. See text for details.

obtained by adapting embedding extractor receptive field and capacity.

Specifically, we adapt the BagNet architecture proposed in (Brendel and Bethge 2019). An overview of the architectures for the tested three receptive field sizes is shown in Table 3.3. We depict the layers of residual blocks in brackets and perform downsampling using convolutions with stride 2 within the first residual block. Note that the architectures for different receptive fields differ in the number of 3×3 convolutions. The rightmost column shows a regular ResNet-50 model. The receptive field is decreased by replacing 3×3 convolutions with 1×1 convolutions. We increase the number of convolution filters by a factor of 2.5 if the receptive field is reduced to account for the loss of the trainable parameters. Moreover, when testing different network capacities we evenly scale the number of convolutional filters by multiplying with a constant factor of $s \in \{1/4, 1/2, 1, 2\}$.

3.1.2.2 Global pooling operation

The global pooling operation takes a topological embedding $\mathbf{E}^t$ of shape $[w_{enc} \times h_{enc} \times c_{enc}]$ as an input and outputs a global image embedding $\mathbf{E}^I$ of shape $[1 \times 1 \times c_{enc}]$. In our experiments, we are testing four different global pooling functions: max-pooling, mean-pooling, logsumexp and soft attention. The max pooling operation simply returns the maximum value per each channel in the topological embedding. This operation can be formally defined as: $\mathbf{E}^I = \max_w \max_h \mathbf{E}^t_{[w,h]}$. Note, that we use subscript notation to denote dimensions of the embedding. The max pooling operation has a spacing effect on gradient backpropagation, during the backward pass through the model all information will be propagated through the embedding position that corresponds to the maximal value. In order to improve gradient backpropagation, one could apply logsumexp pooling, a soft approximation to max pooling. This pooling operation is defined as:

$$\mathbf{E}^I = \log \sum_{w=1}^{w_{enc}} \sum_{h=1}^{h_{enc}} \exp \mathbf{E}^t_{[w,h]}. \tag{3.1}$$

Alternatively, one could use an average pooling operation that computes the mean value for each channel in the topological embedding. This pooling operation can be formally defined as fol-

lows:

$$\mathbf{E}^I = \frac{1}{w_{enc}} \frac{1}{h_{enc}} \sum_{w=1}^{w_{enc}} \sum_{h=1}^{h_{enc}} \mathbf{E}^t_{[w,h]}. \tag{3.2}$$

Finally, the attention based pooling includes an additional weighting tensor $\mathbf{a}$ of dimension $(w_{enc} \times h_{enc} \times c_{enc})$ that rescales each topological embedding before averaging them. This operation can be formally defined as:

$$\mathbf{E}^I = \sum_{w=1}^{w_{enc}} \sum_{h=1}^{h_{enc}} a_{[w,h]} \cdot \mathbf{E}^t_{[w,h]} \tag{3.3}$$

$$s.t. \sum_{w=1}^{w_{enc}} \sum_{h=1}^{h_{enc}} \mathbf{a}_{[w,h]} = 1 \tag{3.4}$$

In our experiments, following Ilse et al. (2018), we parametrize the soft-attention mechanisms as $\mathbf{a}_{[w,h]} = softmax(f(\mathbf{E}_{spat}))_{[w,h]}$, where $f(\cdot)$ is modelled by two fully connected layers with tanh-activation and 128 hidden units.

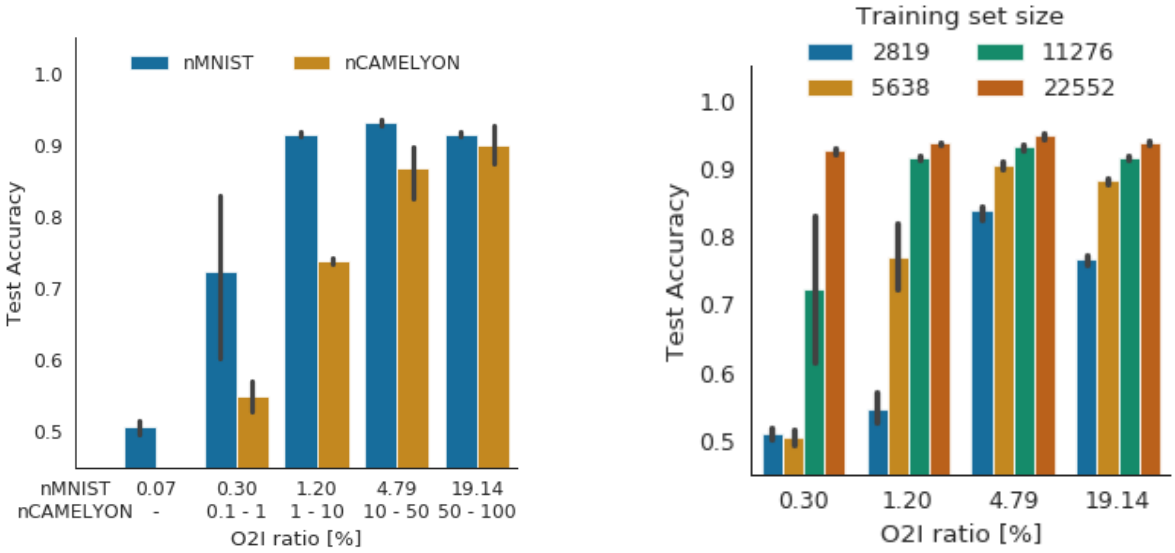
3.1.3 Experimental results

In this subsection, we experimentally test how the CNNs' optimization and generalization scale with *low* and *very low* O2I ratios. First, we provide details about our experimental setup and then we design experiments to provide empirical evidence to the following questions: (1) **Image-level annotations**: Is it possible to train classification systems that generalize well in low and very low O2I scenarios? (2) **O2I limit vs. dataset size**: Is there an O2I ratio limit below which the CNNs will experience generalization difficulties? Does this O2I limit depend on the dataset size? (3) **O2I limit vs. model capacity**: Do higher capacity models generalize better? (4) **Inductive bias - receptive field**: Is adjusting receptive field size to match (or exceed) the expected object size beneficial? (5) **Global pooling operations**: Does the choice of global pooling operation affect model generalization? Finally, we inquire about the **optimization** ease of the models trained on data with very low O2I ratios.

3.1.3.1 Experimental Setup

We adapted the published code from (Brendel and Bethge 2019) for the topological embedding extractor and trained the model with a cross entropy loss. In all our experiments, we used RMSProp (Tieleman and Hinton 2012) with a learning rate of $\eta = 5 \cdot 10^{-5}$ and decayed the learning rate multiplying it by 0.1 at 80, 120 and 160 epochs [2]. All models were trained with cross entropy loss for a maximum of 200 epochs. We used an effective batch size of 32. If the batch did not fit

[2] Before committing to a single optimization scheme, we evaluated a variety of optimizers (Adam, RMSprop and SGD with momentum), learning rates ($\eta \in \{1, 2, 3, 5, 7, 10\} \cdot 10^{-5}$), and 3 learning rate schedules.



(a) O2I ratio performance.

(b) nMNIST

Figure 3.5: a) **Image-level annotations:** Test set accuracy vs. O2I ratio for best models chosen on the validation set. See text for more details. b) **Testing the O2I limit:** Test set performance as a function of training dataset size for the nMNIST dataset.

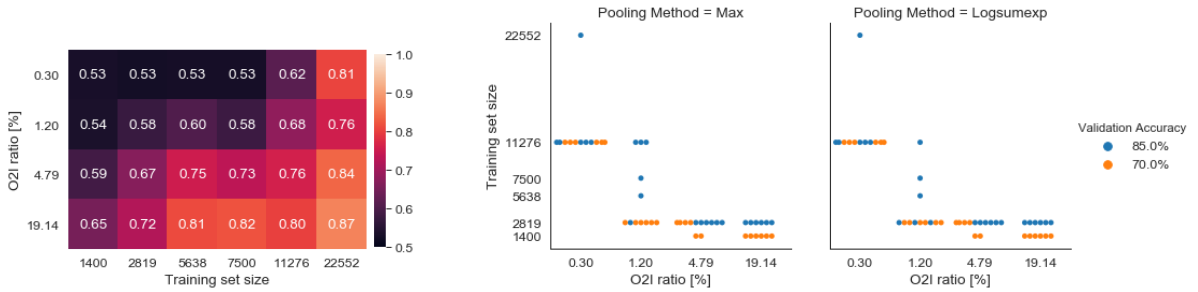
into memory we used smaller batches with gradient accumulation. To ensure robustness of our conclusions, we run every experiment with six different random seeds and report the mean and standard deviation. Throughout the training we monitored validation accuracy, and reported test set results for the model that achieved best validation set performance. Unless stated otherwise, the capacity of the ResNet-50 network is about $2.3 \cdot 10^7$ parameters.

3.1.3.2 Image-level annotations

For this experiment, we vary the O2I ratio on nMNIST and nCAMELYON to test its influence on the generalization of the network. Figure 3.5a depicts the results for the best configuration according to the validation performance: we use max-pooling and receptive field sizes of 33×33 and 9×9 pixels for the nMNIST and nCAMELYON datasets, respectively. For the nMNIST dataset, the plot represents the mean over 6 random seeds together with the standard deviation; while for the nCAMELYON dataset we report an average over both the 6 seeds and the crop sizes. We find that the tested CNNs achieve reasonable test set accuracies for the O2I ratios larger than 0.3% for the nMNIST datset and the O2I ratios above 1% for the histopathology dataset. For both datasets, smaller O2I ratios lead to poor or even random test set accuracies.

3.1.3.3 O2I limit vs. dataset size

We test the influence of the training set size on model generalization for the nMNIST data, to understand the CNNs' generalization problems for very small O2I ratios. We tested six different dataset



(a) nMNIST **(b)** nMNIST

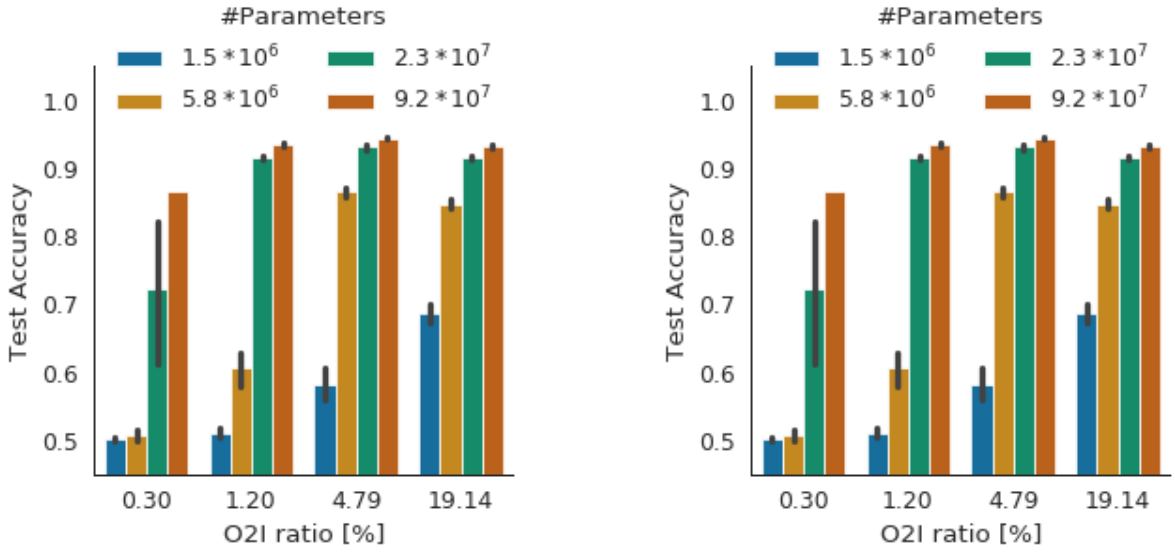
Figure 3.6: Testing the O2I limit: (a) mean validation set accuracy heatmap for max pooling operation, and (b) minimum required training set size to achieve the noted validation accuracy. We test training set sizes $\in \{1400, 2819, 5638, 7500, 11276, 22552\}$ and report the minimum amount of training examples that achieve a specific validation performance pooling over different network capacities.

sizes $(1400, 2819, 5638, 7500, 11276, 22552)$ [3]. Figure 3.5b depicts the results for max-pooling and a receptive field of 33×33 pixels. We observe that larger datasets yield better generalization and this increment is more pronounced for small O2I ratios. For further insights, we plot a heatmap representing the mean validation set results [4] for all considered O2Is and training set sizes (Fig. 3.6a) as well as the minimum number of training examples to achieve a validation accuracy of 70% and 85% (Fig. 3.6b). We observe that in order to achieve good classification generalization the required training set size rapidly increases with the decrease of the O2I ratio.

3.1.3.4 O2I limit vs. capacity

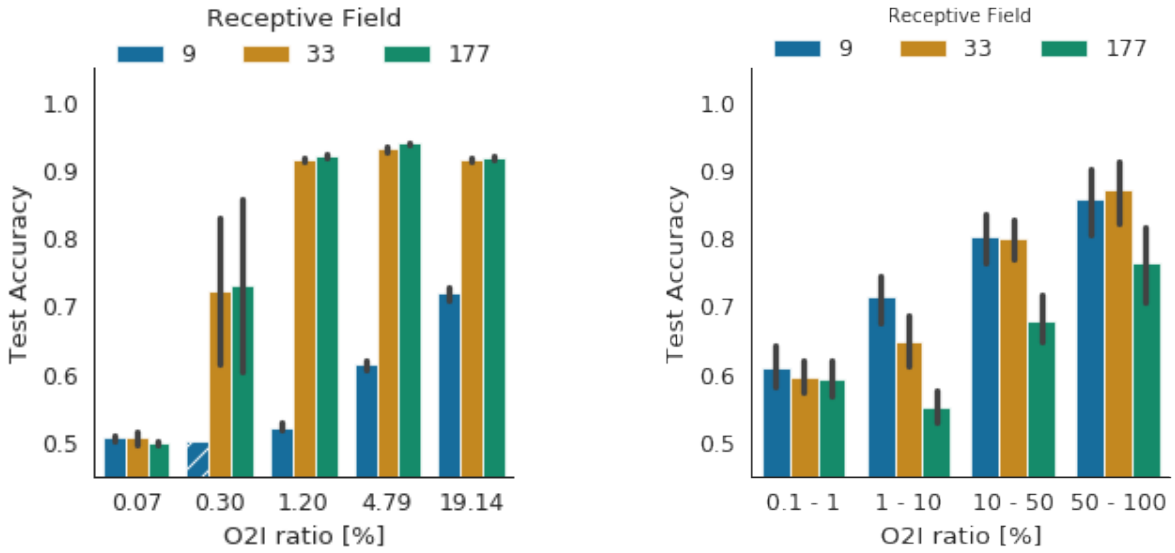
In this experiment, we train networks with different capacities — by uniformly scaling the initial number of filters in convolutional kernels by $[\frac{1}{4}, \frac{1}{2}, 1, \text{and } 2]$. We chose the maximum scaling factor so that the largest resolution images still fit in the available GPU memory. For images with O2I ratio of 0.07, the available GPU memory prevents testing networks with higher capacity. We show the CNNs test set performances as a function of the O2I ratio and the network capacity in Fig. 3.7a and Fig. 3.7b for the nMNIST (with 11k training points) and nCAMELYON data, respectively. On nMNIST, we observe a clear trend, where the model test set performance increases with capacity and this boost is larger for smaller O2Is. We hypothesize, that this generalization improvement is due to the model ability to learn-to-ignore the input data noise; with smaller O2I there is more noise to ignore and, thus, higher network capacity is required to solve the task. However, for the nCAMELYON dataset, this trend is not so pronounced and we attribute this to the limited dataset size (more precisely to the small number of unique lesions). These results suggest that collecting a very large histopathology dataset might enable training of CNN models using *only* image level annotations.

[3] We allow to reuse each digit 3 for larger training sets and select a subset for smaller training sets.
[4] More precisely, we plot the mean of all pipeline configurations that surpassed 70% training accuracy.



(a) nMNIST **(b)** nCAMELYON

Figure 3.7: Testing the O2I limit: Subfigure (a) depicts the test set performance as a function of training dataset size for the nMNIST dataset, while subfigure (b) shows the test set performance as a function of model capacity for the nMNIST dataset and the nCAMELYON dataset, respectively.

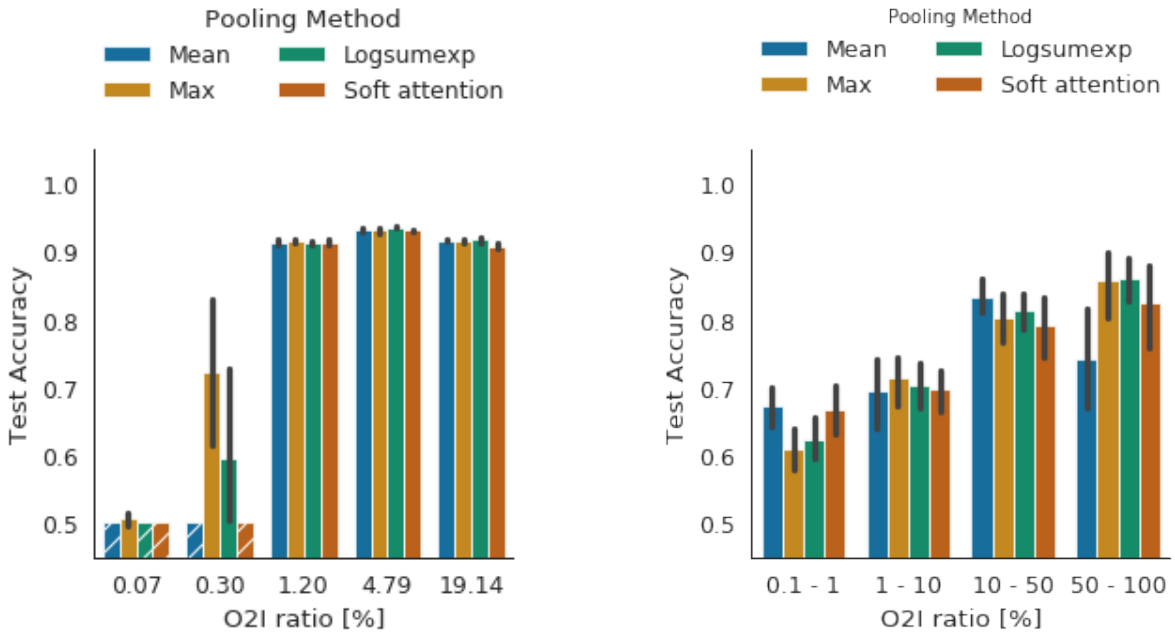


(a) nMNIST **(b)** nCAMELYON

Figure 3.8: Inductive bias: for (a) the nMNIST dataset and (b) the nCAMELYON dataset. We report only runs that fit the training data. Otherwise we report random accuracy and depict it with a texture on the bars.

3.1.3.5 Inductive bias - receptive field

We report the test accuracy as a function of the O2I ratio and the receptive field size for nMINIST in Fig. 3.8a and for nCAMELYON in Fig. 3.8b. Both plots depict results for the global max pooling operation. For nMNIST, we observe that a receptive field that is bigger than the area occupied by one



(a) nMNIST **(b)** nCAMELYON

Figure 3.9: Global pooling operations: for (a) the nMNIST dataset and (b) the nCAMELYON datset. We report only runs that fit the training data. Otherwise we report random accuracy and depict it with a texture on the bars.

single digit leads to best performances; for example, receptive fields of 33×33 and 177×177 pixels clearly outperform the smallest tested receptive field of 9×9 pixels. However, for the nCAMELYON dataset we observe that the smallest receptive field actually performs best. This suggests that most of the class-relevant information is contained in the texture and that higher receptive fields pick up more spurious correlations, because the capacity of the networks is constant.

3.1.3.6 Global pooling operations

In this experiment, we compare the performance of four different pooling approaches. We present the relation between test accuracy and pooling function for different O2I ratios with a receptive field of 33×33 pixels for nMNIST in Fig. 3.9a and 9×9 pixels for nCAMELYON in Fig. 3.9b. On the one hand, for the nMNIST dataset, we observe that for the relatively large O2I ratios, all pooling operations reach similar performance; however, for smaller O2Is we see that max-pooling is the best choice. We hypothesize that the global max pooling operation is best suited to remove nMNIST-type of structured input noise. On the other hand, when using the histopathology dataset, for the smallest O2I mean and soft attention poolings reach best performances; however, these outcomes might be affected by the relatively small nCAMELYON dataset used for training.

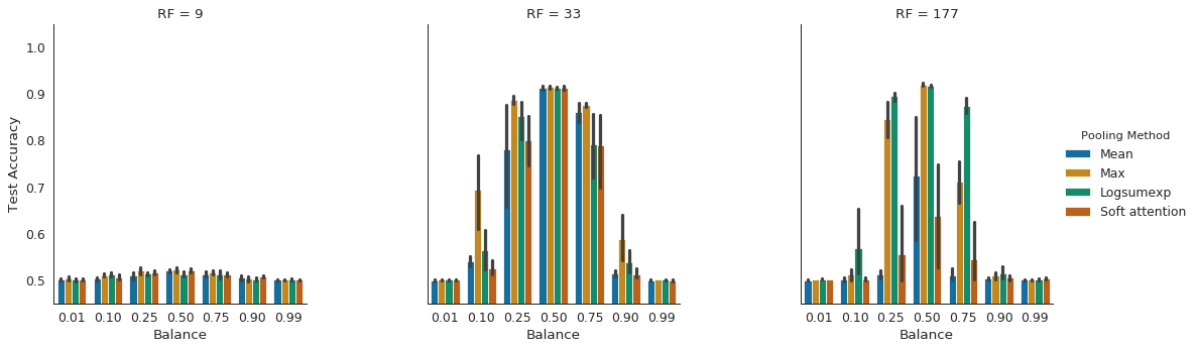


Figure 3.10: Impact of the training set balance on model accuracy for different pooling operations and receptive field sizes.

3.1.3.7 Class-imbalanced classification

In many medical imaging datasets, it is common to be faced with class-imbalanced datasets. Therefore, in this experiment, we use our nMNIST dataset and test CNNs generalization under moderate and severe class imbalanced scenario. We alter the training set class balance by altering the proportion of positive images in the training dataset and use the following balance values 0.01, 0.1, 0.25, 0.5, 0.75, 0.9 and 0.99, where a value of 0.01 means almost no positive examples and 0.99 indicates very low number of negative images available at training time. Moreover, we ensure that the dataset size is constant ($\approx$ 11k) and only the class-balance is modified. We run the experiments using the O2I ratio of 1.2%, three receptive field sizes (9×9, 33×33 and 177×177 pixels) and four pooling operations (mean, max, logsumexp and soft attention). For each balance value, we train 6 models using 6 random seeds and we oversample the underrepresented class. The results are depicted in Fig. 3.10. We observe that the model performance drops as the the training data becomes more unbalanced and that max pooling and logsumexp seem to be the most robust to the class imbalance.

3.1.3.8 Increase of model capacity for small dataset sizes.

We also tested the effect of model capacity increase while having access only to a small dataset (3k class-balanced images) and contrast it with a larger dataset of $\approx$ 11k training images. We run this experiment on the nMNIST dataset using a network with $2.3 \cdot 10^7$ parameters using global max pooling operation and there different receptive field sizes: 9×9, 33×33 and 177×177 pixels. The results are depicted in Fig. 3.11. It can be seen that the model's capacity increase does not lead to better generalization, for small size datasets of $\approx$ 3k.

3.1.3.9 Optimization

In our large scale nMNIST experiments (when using $\approx$ 11k datapoints), we observed that some configurations have problems fitting the training data [5]. In some runs, after significant efforts put

[5] We did not observe optimization problems for small dataset sizes of the nMNIST nor for nCAMELYON.

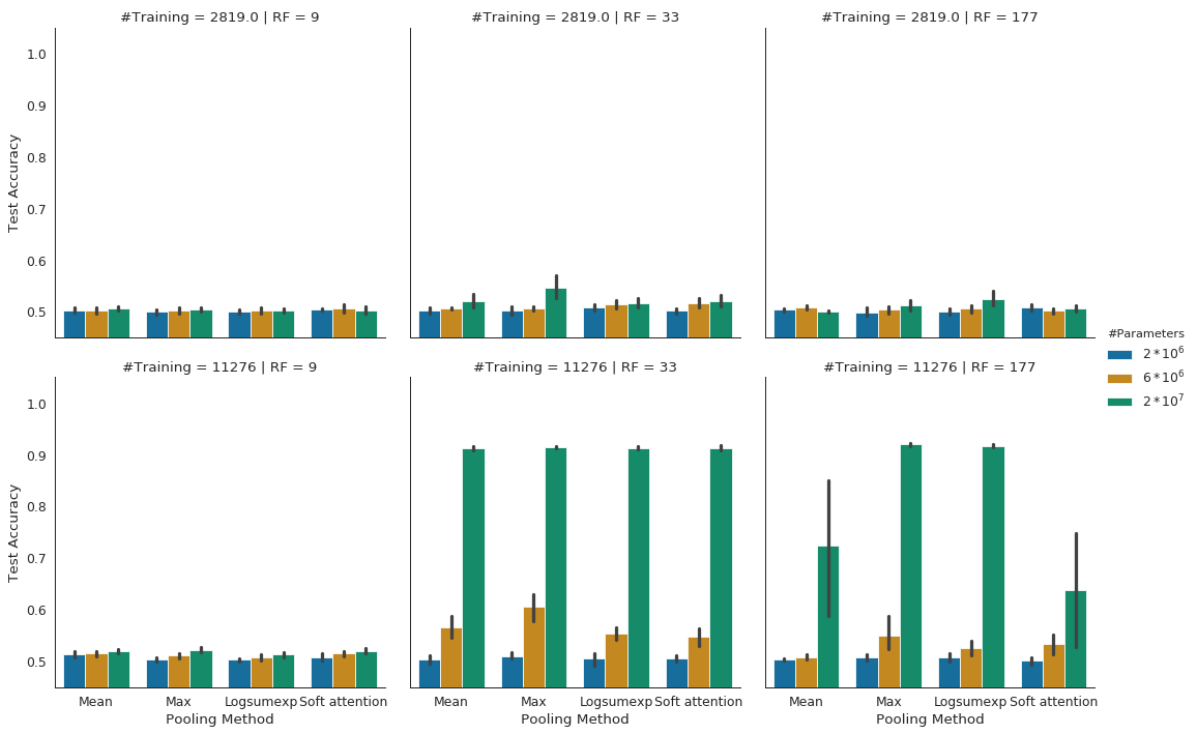
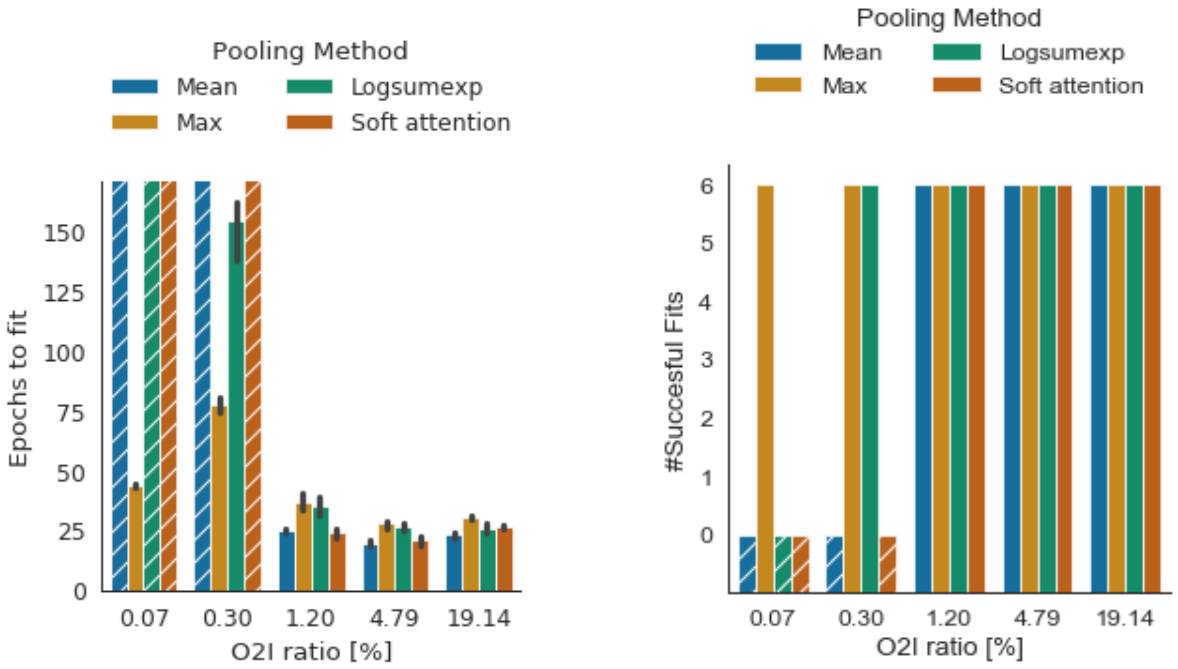


Figure 3.11: Impact of the network capacity on the generalization performance dependent on the training set size for nMNIST at O2I ratio = 1.2%. The improvement based on the increased network capacity shrinks with smaller training set.

into CNNs hyperparamenter selection, the training accuracy was close to random. To investigate this issue further, we followed the setup of randomized experiments from (Arpit et al. 2017; Zhang et al. 2017) and we substituted the nMNIST datapoints with samples from an isotropic Gaussian distribution. On the one hand, we observed that all the tested setups of our pipeline were able to memorize the Gaussian samples, while, on the other hand, most setups were failing to memorize the same-size, nMNIST datataset for small and very small O2I ratios. We argue that the nMNIST structured noise and its compositionality may be a "harder" type of noise for the CNNs than Gaussian isotropic noise. To provide further experimental evidence, we depict average time-to-fit the training data (in epochs) in Fig. 3.12a as well as number of successful optimizations in Fig. 3.12b for different O2I ratios and pooling methods. Here, we define an optimization to be successful if it the training set accuracy surpassed 99%.. We observe that the optimization gets progressively harder with decreasing O2I ratio (with max pooling being the most robust). Moreover, we note that the results are consistent across different random seeds, where all runs either succeed or fail to converge.

3.1.3.10 Weakly supervised object detection: nMNIST

We test the object localization capabilities of the trained classification models by examining their saliency maps. Figure 3.13 shows examples of the nMNIST dataset with the object bounding box in blue and the magnitude of the saliency in red. We rescale the saliency to $[0, 1]$ for better con-



(a) nMNIST **(b)** nMNIST

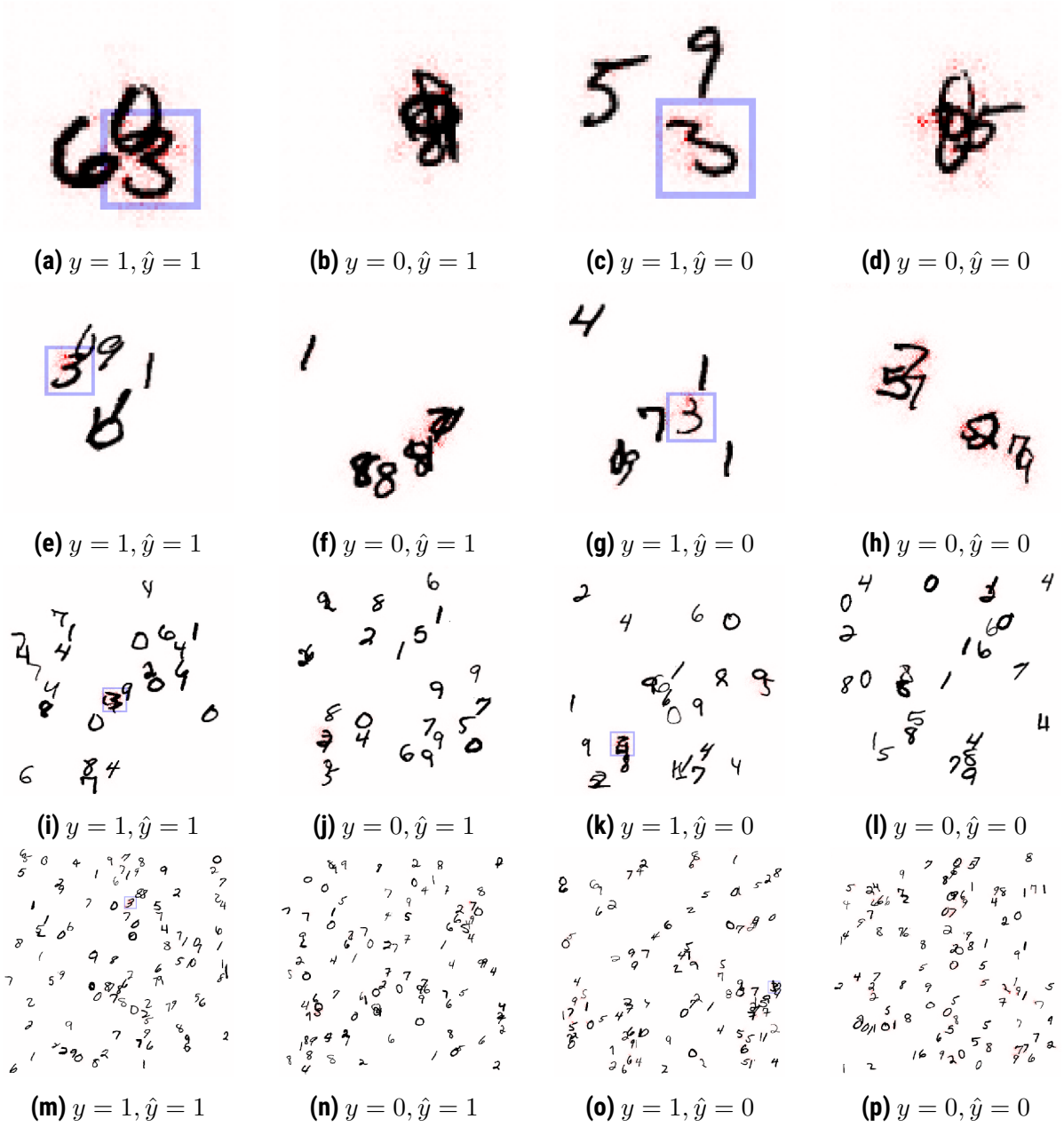
Figure 3.12: nMNIST optimization: (a) number of training epochs needed to fit the 11k training data and (b) the number of successful runs. The textured bars indicate that the model did not fit the training data for all random seeds.

trast. However, this prevents the comparison of absolute saliency values across different images. In samples containing an object of interest, the models correctly assign high saliency to the regions surrounding the relevant object. On negative examples, the network assigns homogenous importance to all objects.

We localise an object of interest as the location with maximum saliency. We follow (Oquab et al. 2015) to quantitatively examine the object detection performance using the saliency maps of the models. We plot the corresponding average precision in Fig. 3.14. We find that the detection performance deteriorates for smaller O2I ratios regardless of the method. This is aligned with the classification accuracy. For small O2I ratios, max-pooling achieves the best detection scores. On larger O2I ratios, logsumexp achieves the best scores.

3.1.3.11 Weakly supervised object detection: nCAMELYON

We qualitatively show object detection on nCAMELYON in Figs. 3.15 to 3.18, for True Positives, True Negatives, False Positives and False Negatives. We observe weak correlation between segmentation maps and saliency maps, signifying that the classifier was able to focus on the object of interest instead of looking at superficial signals in the data.



(a) $y = 1, \hat{y} = 1$ **(b)** $y = 0, \hat{y} = 1$ **(c)** $y = 1, \hat{y} = 0$ **(d)** $y = 0, \hat{y} = 0$

(e) $y = 1, \hat{y} = 1$ **(f)** $y = 0, \hat{y} = 1$ **(g)** $y = 1, \hat{y} = 0$ **(h)** $y = 0, \hat{y} = 0$

(i) $y = 1, \hat{y} = 1$ **(j)** $y = 0, \hat{y} = 1$ **(k)** $y = 1, \hat{y} = 0$ **(l)** $y = 0, \hat{y} = 0$

(m) $y = 1, \hat{y} = 1$ **(n)** $y = 0, \hat{y} = 1$ **(o)** $y = 1, \hat{y} = 0$ **(p)** $y = 0, \hat{y} = 0$

Figure 3.13: Example images from the nMNIST validation set and their corresponding saliency maps in red. We generate the saliency maps by calculating the absolute of the gradients with respect to the input image using max-pooling, a receptive field of 33, and ResNet-50 capacity. From top to bottom, we show random examples for O2I ratios of $\{19.14, 4.79, 1.20, 0.30\}\%$. We annotate the object of interest with a blue outline. The captions show the true label y and the prediction $\hat{y}$.

3.1.4 Related Work

3.1.4.1 Tiny Object Classification

Reasoning about tiny objects is of high interest in many computer vision areas, such as medical imaging (Aresta et al. 2018; Ehteshami Bejnordi et al. 2017; Setio et al. 2017; Suckling 1994; Sudre et al. 2018) and remote sensing (Pang et al. 2019; Xia et al. 2018). To overcome the low signal-

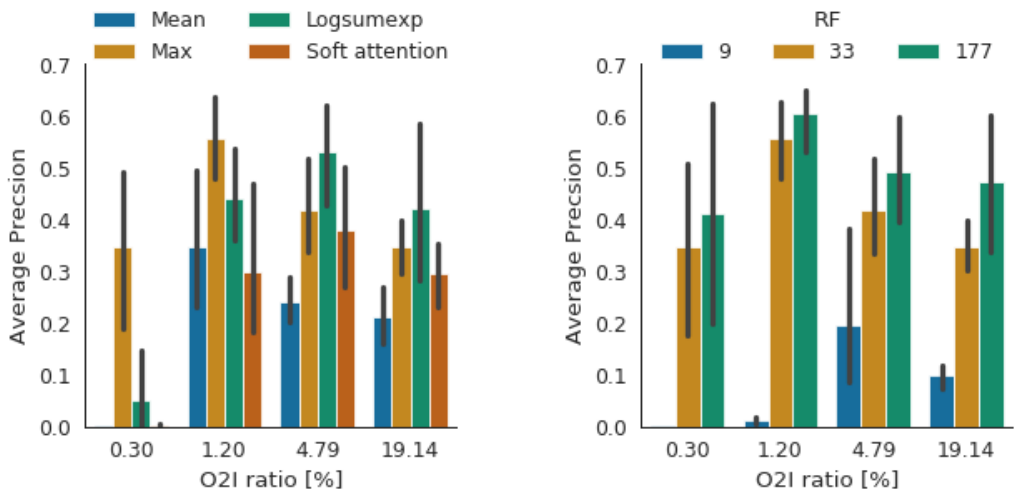


Figure 3.14: Average precision for detecting the object of interest using the saliency maps for nMNIST. We adapt (Oquab et al. 2015) and localize an object by the maximum magnitude of the saliency. We use the magnitude of the saliency as the confidence of the detection. We count wrongly localised objects both as false positive and false negative. For images without object of interest, the we increase the false positive count only. We plot results for max-pooling, a receptive field of 33, a training set with 11276 examples and ResNet-50 capacity. (a) shows the dependence of the AP on the pooling method using $RF = 33 \times 33$, (b) shows the dependence on the receptive field using max-pooling.

to-noise ratio, most approaches rely on manual dataset "curation" and collect additional *pixel-level annotations* such as landmark positions (Borovec et al. 2018), bounding boxes (Resta et al. 2011; Wei et al. 2019) or segmentation maps (Ehteshami Bejnordi et al. 2017). This additional annotation allows to transform the original needle-in-a-haystack problem into a less noisy but imbalanced classification problem (Bándi et al. 2019; Lee and Paeng 2018; Wei et al. 2019). However, collecting pixel level annotations has a significant cost and might require expert knowledge, and as such, is a bottleneck in the data collection process.

Other approaches leverage the fact that task-relevant information is often not uniformly distributed across input data, e.g. by using attention mechanisms to process very high-dimensional inputs (Almahairi et al. 2016; Ba et al. 2015; Katharopoulos and Fleuret 2019; Mnih et al. 2014). However, those approaches are mainly motivated from a computational perspective trying to reduce the computational footprint at inference time.

Some recent research has also studied attention based approaches both in the context of multi-instance learning (Ilse et al. 2018) and histopathology image classification (Tomita et al. 2018). However, neither of the works report the exact O2I ratio used in the experiments.

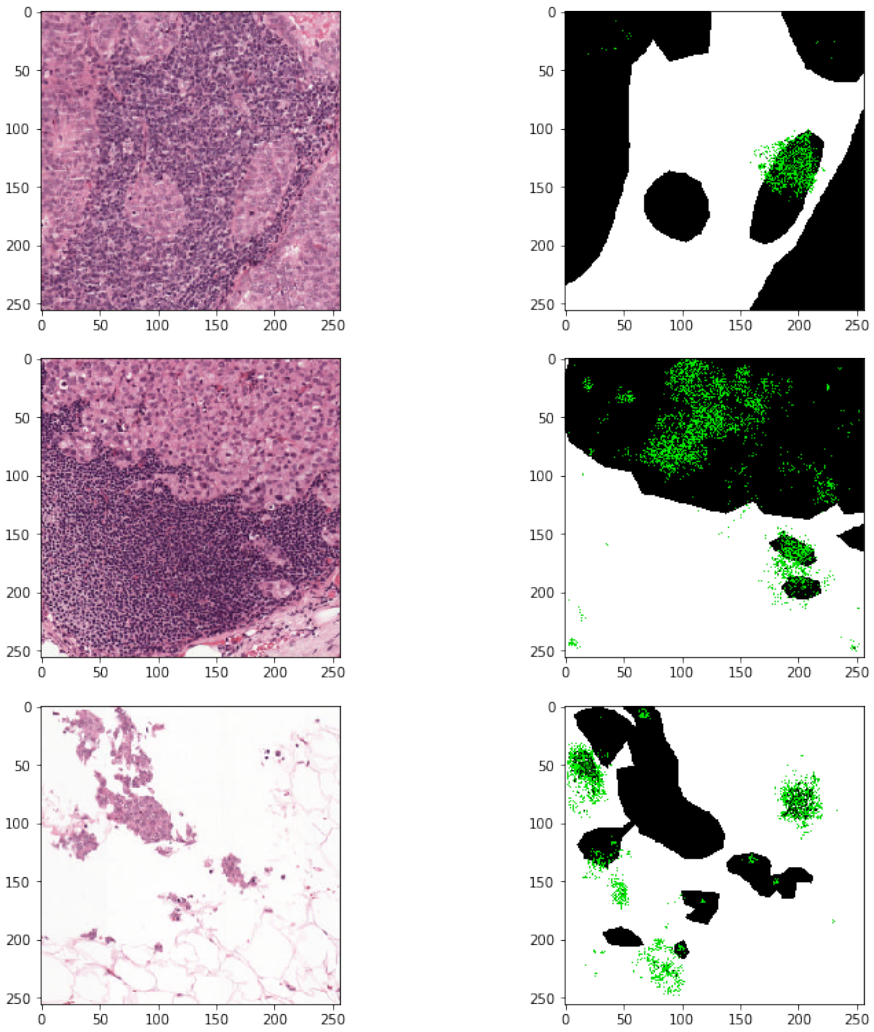


Figure 3.15: Example True Positive Images of nCAMELYON validation sets and their corresponding segmentation maps with saliencies overlaid.

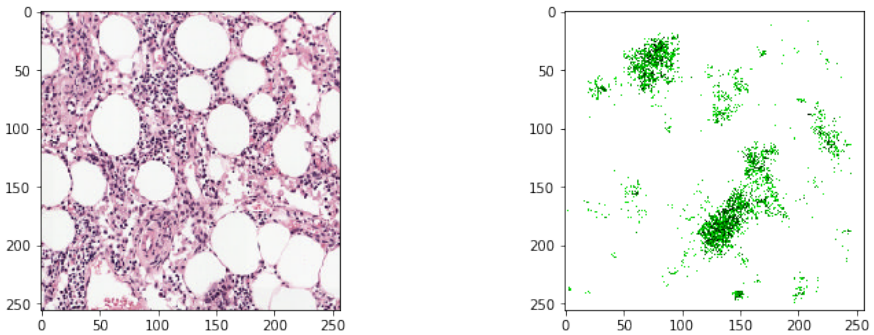


Figure 3.16: Example True Negative Image of nCAMELYON validation sets and corresponding saliency map.

3.1.4.2 Generalization of CNNs

In this subsection, we briefly highlight the dimensions of optimization and generalization of CNN that are handy in low O2I classification scenarios.

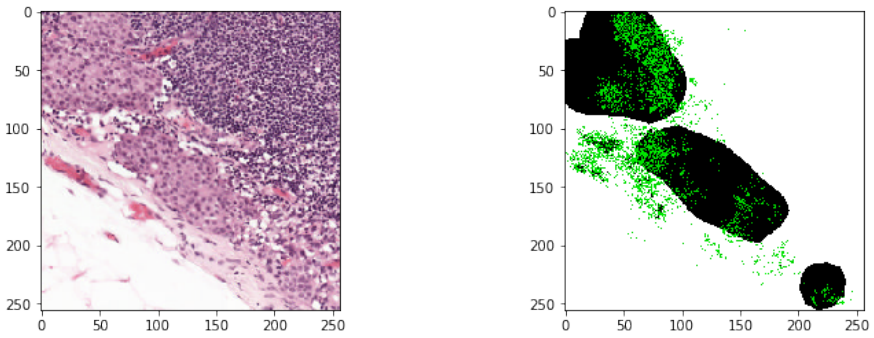


Figure 3.17: Example False Negative Image of nCAMELYON validation sets and corresponding segmentation map with saliency overlaid.

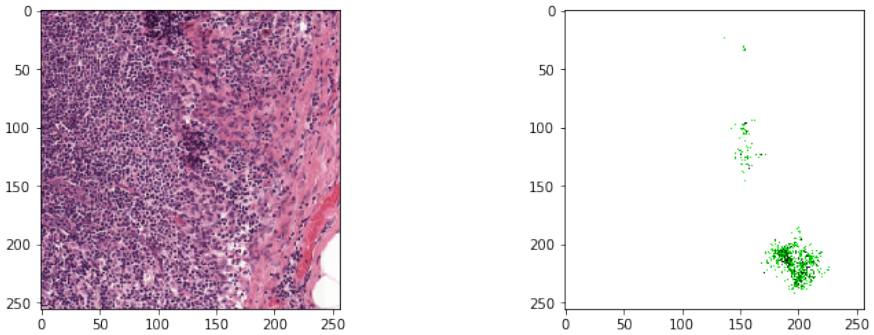


Figure 3.18: Example False Positive Image of nCAMELYON validation sets and corresponding saliency map.

Model capacity. For fixed training accuracy, over-parametrized CNNs tend to generalize better (Novak et al. 2018). In addition, when properly regularized and given a fixed size dataset, higher capacity models tend to provide better performance (He et al. 2016b; Huang et al. 2017b). However, finding proper regularization is not trivial (Goodfellow et al. 2016).

Dataset size. CNN performance improves logarithmically with dataset size (Sun et al. 2017). Moreover, in order to fully exploit the data benefit, the model capacity should scale jointly with the dataset size (Mahajan et al. 2018; Sun et al. 2017).

Model inductive biases. Inductive biases limit the space of possible solutions that a neural network can learn (Goodfellow et al. 2016). Incorporating these biases is an effective way to include data (or domain) specific knowledge in the model. Perhaps the most successful inductive bias is the use of convolutions in CNNs (LeCun and Bengio 1998). Different CNN architectures (e. g. altering network connectivity) also lead to improved model performance (He et al. 2016b; Huang et al. 2017b). Additionally, it has been shown on the ImageNet dataset that CNN accuracy scales logarithmically with the size of the receptive field (Brendel and Bethge 2019).

3.1.5 Discussion and Conclusions

Although low input image signal-to-noise scenarios have been extensively studied in the field of signal processing (e.g. in tasks such as image reconstruction), less attention has been devoted to low signal-to-noise classification scenarios. Thus, in this paper we identified an *unexplored* machine learning problem, namely image classification in *low* and *very low* signal-to-noise ratios. In order to study such scenarios, we built two datasets that allowed us to perform controlled experiments by manipulating the input image signal-to-noise ratio and highlighted that CNNs struggle to show good generalization for *low* and *very low* signal-to-noise ratios even for a relatively elementary MNIST-based dataset. Finally, we ran a *series of controlled experiments*[6] that explore both a variety of CNNs' architectural choices and the importance of training data scale for the *low* and *very low* signal-to-noise classification. One of our main observation was that properly designed CNNs can be *trained in low O2I regime without using any pixel-level annotations and generalize* if we leverage enough training data; however, *the amount of training data required for the model to generalize scales rapidly with the inverse of the O2I ratio*. Thus, with our paper (and the code release) we invite the community to work on data-efficient solutions to *low* and *very low* signal-to-noise classification.

Our experimental study exhibits limitations: First, due to the lack of large scale datasets that allow for explicit control of the input signal-to-noise ratios, we were forced to use the synthetically built nMNIST dataset for most of our analysis. As a real life dataset, we used crops from the histopathology CAMELYON dataset; however, due to relatively a small number of unique lesions we were unable to scale the histopathology experiments to the extent as the nMNIST experiments, and, as result, some conclusions might be affected by the limited dataset size. Other large scale computer vision datasets like MS COCO (Lin et al. 2014) exhibit correlations of the object of interest with the image background. For MS COCO, the smallest O2I ratios are for the object category "sports ball" which on average occupies between 0.3% and 0.4% of an image and its presence tends to be correlated with the image background (e. g. presence of sports fields and players). However, future research could examine a setup in which negative images contain objects of the categories "person" and "baseball bat" and positive images also contain "sports ball". Second, all the tested models improve the generalization with larger dataset sizes; however, scaling datasets such as CAMELYON to tens of thousands of samples might be prohibitively expensive. Instead, further research should be devoted to developing computationally-scalable, data-efficient inductive biases that can handle very low signal-to-noise ratios with limited dataset sizes. Future work, could explore the knowledge of the low O2I ratio and therefore sparse signal as an inductive bias. Finally, we studied low signal-to-noise scenarios only for binary classification scenarios; further investigation should be devoted to multi-class problems. We hope that this study will stimulate the research in image classification for low signal-to-noise input scenarios which are highly relevant for biomedical imaging applications.

[6] We ran more than 750 experiments each with 6 different seeds.

Table 3.3: Schematic of the architecture of the different topological embedding encoders used in this paper. The operations and their corresponding parameters of the residual blocks are denoted in brackets. The first block within each section performs downsampling using convolutions with stride 2. We use InstanceNorm instead of BatchNorm and test different pooling methods after the topological embeddings.

RF=9	RF=33	RF=177
	conv, $1 \times 1, 64$	
	conv, $3 \times 3, 64$	
$\begin{bmatrix}\text{conv}, 1 \times 1, 64\\ \text{conv}, 3 \times 3, 64\\ \text{conv}, 1 \times 1, 256\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 64\\ \text{conv}, 3 \times 3, 64\\ \text{conv}, 1 \times 1, 256\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 64\\ \text{conv}, 3 \times 3, 64\\ \text{conv}, 1 \times 1, 256\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 64\\ \text{conv}, 1 \times 1, 160\\ \text{conv}, 1 \times 1, 256\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 64\\ \text{conv}, 1 \times 1, 160\\ \text{conv}, 1 \times 1, 256\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 64\\ \text{conv}, 3 \times 3, 64\\ \text{conv}, 1 \times 1, 256\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 64\\ \text{conv}, 1 \times 1, 160\\ \text{conv}, 1 \times 1, 256\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 64\\ \text{conv}, 1 \times 1, 160\\ \text{conv}, 1 \times 1, 256\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 64\\ \text{conv}, 3 \times 3, 64\\ \text{conv}, 1 \times 1, 256\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 3 \times 3, 128\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 3 \times 3, 128\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 3 \times 3, 128\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 1 \times 1, 320\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 1 \times 1, 320\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 3 \times 3, 128\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 1 \times 1, 320\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 1 \times 1, 320\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 3 \times 3, 128\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 1 \times 1, 320\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 1 \times 1, 320\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 128\\ \text{conv}, 3 \times 3, 128\\ \text{conv}, 1 \times 1, 512\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 3 \times 3, 256\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 3 \times 3, 256\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 3 \times 3, 256\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 3 \times 3, 256\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 3 \times 3, 256\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 3 \times 3, 256\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 1 \times 1, 640\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 256\\ \text{conv}, 3 \times 3, 256\\ \text{conv}, 1 \times 1, 1024\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 1 \times 1, 1280\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 3 \times 3, 512\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 3 \times 3, 512\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 1 \times 1, 1280\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 1 \times 1, 1280\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 3 \times 3, 512\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 1 \times 1, 1280\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 1 \times 1, 1280\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 3 \times 3, 512\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$
$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 1 \times 1, 1280\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 1 \times 1, 1280\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$	$\begin{bmatrix}\text{conv}, 1 \times 1, 512\\ \text{conv}, 3 \times 3, 512\\ \text{conv}, 1 \times 1, 2048\end{bmatrix}$

3.2 Extracting information from patches in brain scans

We explore whether texture information is sufficient for certain tasks in medical image analysis. For this, we generalise BagNets to arbitrary regression tasks and 3D images and examine the performance of different receptive fields. We apply BagNets to age regression and sex classification tasks on T1-weighted brain MRI to examine the dependency of modern deep learning architectures on local texture in these medical imaging tasks. We find that the bag-of-local-features approach yields comparable results to larger receptive fields.

3.2.1 Method

BagNets (Brendel and Bethge 2019) are adaptations of the ResNet-50 architecture (He et al. 2016a), that restrict the receptive field by replacing 3×3 convolutional kernels with 1×1 kernels. A regular ResNet-50 has a receptive field of 177 pixels, whereas BagNets explore receptive fields of 9, 17 and 33 pixels. The use of small receptive fields enforces locality in the extracted features. After extracting local features a global spatial average builds the bag-of-local-features and enforces the invariance to spatial relations. The bag of features is then processed by a linear layer to provide the final prediction. Because of the linearity of the average operation and the final linear layer, it is possible to exchange the order of those operations, which enables the extraction of localised prediction maps.

3.2.2 Experiments & Results

We test the BagNets on the public Cambridge Centre for Ageing and Neuroscience (CamCAN) dataset (Taylor et al. 2017). The dataset contains T1- and T2-weighted brain MRI of 652 healthy subjects within an age range of 18 to 87. The subjects are approximately uniformly distributed across age and sex. The standard-deviation of the age across all subjects is 18.6 years. We only use the T1-weighted scans for our experiments and randomly split the scans into training, validation and test sets with 456, 65 and 131 subjects each. All scans have an isotropic resolution of 1 mm. We use skull-stripped, bias-fiel corrected images and extract random crops of shape $[128 \times 160 \times 160]$ during training. We whiten the images with statistics extracted from within the brain mask.

We use the architecture from (Brendel and Bethge 2019) but replace 2D with 3D convolutions and half the number of feature maps. We train the network with batch size 1 and accumulate gradients over 16 batches. To alleviate the effects of small batches we use instance normalization (Ulyanov et al. 2016) instead of batch normalization (Ioffe and Szegedy 2015). We use a cross-entropy loss for the sex classification and MSE loss for the age regression. We use the Adam optimizer (Kingma and Ba 2015) with a learning rate of $\eta = 0.001$, $\epsilon = 10^{-5}$ and employ an L_2-regularization of $\lambda = 0.0001$. We train the network for 500 epochs and decay the learning rate by a factor of 10

every 100 epochs. We use the checkpoint with the best validation performance for evaluation on the test data.

Table 3.4 shows the mean absolute error (MAE) and accuracy for the age and sex prediction for different receptive fields. We achieve a MAE between $3.86 - 5.53$ years for age and an accuracy between $80.9 - 84.0\%$ for sex. Age regression has a stronger dependency on the receptive field than the sex classification. However, we find that the larger receptive field exhibits better training performance and might be prone to overfitting.

Table 3.4: Results for the age regression and sex classification task for different receptive fields. We report the mean absolute error for age and classification accuracy for sex. We find that small receptive fields yield comparable results on those tasks.

Receptive Field	Age	Sex
$(9mm)^3$	5.53	83.2%
$(17mm)^3$	5.32	84.0%
$(33mm)^3$	4.98	84.0%
$(177mm)^3$	3.86	80.9%

We examine the local predictions on two examples from the test set in Fig. 3.19 for age regression and Fig. 3.20 for sex classification. The sex classification predicts 0 for male and 1 for female. The first row shows a 20 year old male subject, the second row shows an 80 year old female. The columns respectively show the middle slice of the T1-weighted MRI and the local predictions with receptive fields 9, 17, 33 and 177. Similarly to (Brendel and Bethge 2019), we find that small receptive fields lead to more localised predictions, whereas larger receptive fields show more spread out predictions. Interestingly, the age regression exhibits very high variance predictions, where only few very high values contribute to the mean prediction of the volume. Generally, we find that the local predictions we get from our model do not seem as interpretable as in (Brendel and Bethge 2019).

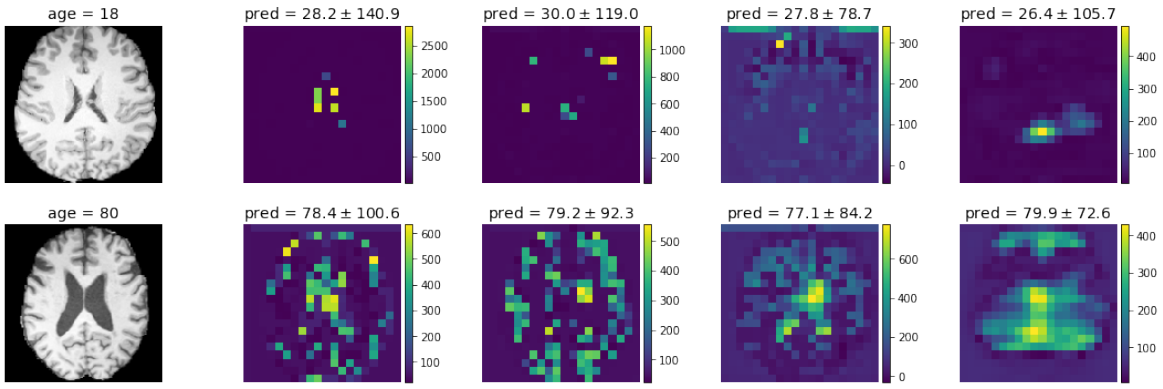


Figure 3.19: Localised age prediction on a 20 year old male subject (first row) and 80 year old female subject (second row). The columns show the middle slice of the T1-weighted MRI and the localised predictions for receptive fields 9, 17, 33 and 177.

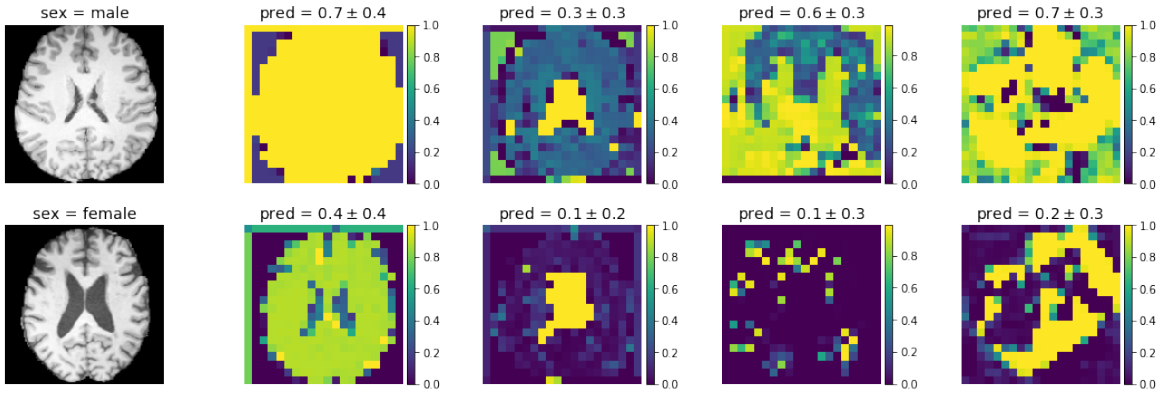


Figure 3.20: Localised sex classification on a 20 year old male subject (first row) and 80 year old female subject (second row). The different columns show the middle slice of the T1-weighted MRI and the localised predictions for receptive fields 9, 17, 33 and 177. The network predicts male as 0 and female as 1.

3.2.3 Discussion & Conclusion

We have generalised the concept of BagNets (Brendel and Bethge 2019) to the setting of 3D images and general regression tasks. We have shown that a BagNet with a receptive field of $(9mm)^3$ yields surprisingly accurate predictions of age and sex from T1-weight MRI scans. However, we find that localised predictions of age and sex do not yield easily interpretable insights into the workings of the neural network which will be subject of future work. Further, we believe that more accurate localised predictions could lead to advanced clinical insights similar to (Becker et al. 2018; Cole et al. 2019).

Chapter 4

Quantifying the Uncertainty of Deep Learning Models

This chapter is based on the following publication:

(a) **Pawlowski, N.**, Brock, A., Lee, M. C., Rajchl, M., and Glocker, B. (2017a). "Implicit Weight Uncertainty in Neural Networks". In: *NeurIPS Workshop on Bayesian Deep Learning* – (Pawlowski et al. 2017a)

Code is available at https://github.com/pawni/BayesByHypernet.

While deep learning methods achieve state-of-the-art predictive performances across various domains, there is no conclusive approach of enabling AI-augmented human decision making. In human-human interactions decisions are often made by form of discussion or argumentation that include the mutual explanation of how an opinion was formed and which factors contributed to the decision as well as sharing insights about potential gaps in the individual's knowledge. The previous chapter explored constraints to the neural network architecture as a way to offer crude explanations based on the ability to only produce localised predictions and combining them in simple ways. However, this leaves the question of how to equip neural networks with the capability of saying *"I don't know"* which is necessary to convey knowledge about the lack of knowledge. AI models that know their limits and convey their certainty could lead to more trust in their predictions and the output of a prediction's certainty might even be necessary in safety-critical scenarios such as medical imaging.

Estimating the uncertainty of predictions of deep learning models is an active area of research and calibrated prediction probabilities are critical for the safe and trustworthy deployment of deep

learning systems to real world applications. One particular approach to enable reliable predictions has been to rely on Bayesian methods to account for model uncertainty due to finite data by modelling the full distributions of neural network weights rather than only using point estimates. This approach either uses complex sampling techniques (Welling and Teh 2011) that are very computationally expensive or relies on approximations that use variational Bayes (see Section 2.1.2). Most variational approaches only build very crude approximations to the true posterior distribution due to the choice of variational distribution (Blundell et al. 2015) and therefore limit the reliability and accuracy of the resulting deep learning methods.

This chapter proposes the usage of a more complex variational distribution that makes use of recent advances in deep learning and approximate Bayesian computation: inspired by work on hypernetworks (Brock et al. 2018; Ha et al. 2017) we use generative adversarial networks (GANs (Goodfellow et al. 2014)) to implicitly model the posterior distribution of the weights of neural networks. We show with our experiments that this approach allows for highly complex posterior distributions that yield competitive predictive performance while providing insights into what the neural network does not know.

4.1 Introduction

Neural networks achieve state of the art results on a wide variety of tasks (LeCun et al. 2015), with applications spanning image recognition (Hu et al. 2018), machine translation (Lample et al. 2018) and reinforcement learning (Silver et al. 2017). Such success is often mitigated by the need for vast troves of data, and a tendency towards poorly calibrated and overconfident predictions (Guo et al. 2017). However, real-world decision making processes that aim to leverage neural networks (*e.g.* medical applications, self-driving cars, *etc.*) are frequently faced with a dearth of data and the need for reliable uncertainty estimates, as overconfidence in the wrong situation could prove dangerous (Amodei et al. 2016).

To address the issue of overconfident predictions, recent works proposed approaches based on calibration methods (Guo et al. 2017), frequentist interpretations of ensembles (Lakshminarayanan et al. 2017; Osband et al. 2016), and approximate Bayesian inference (MacKay 1992; Welling and Teh 2011). Of those approaches, Bayesian deep learning (BDL) offers a particularly principled approach to enable uncertainty estimates within the existing deep learning framework by aiming to marginalise the model parameters.

The current research in BDL is primarily divided into variational methods (Blundell et al. 2015; Gal and Ghahramani 2015; Louizos and Welling 2017) and Monte Carlo methods (Chen et al. 2014; Lu et al. 2017; Welling and Teh 2011). We conducted a toy experiment (Fig. 4.1) that illustrates the issue of low predictive uncertainty in unseen regions for regular deep learning methods (e.g. MAP estimate and ensembles) as well as the more reliable uncertainty of Bayesian approximations.

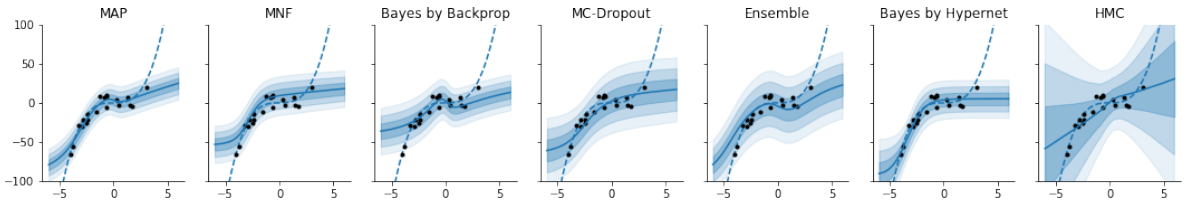


Figure 4.1: Toy experiment inspired by Hernández-Lobato and Adams (2015): Real function (dashed line) with sampled data points (black dots). Hamiltonian Monte Carlo (HMC) is seen as the 'gold standard' in finding the true distribution. The proposed *Bayes by Hypernet* finds a good trade-off between data fit and predictive uncertainty. MC-Dropout (Gal and Ghahramani 2015), deep ensembles (Lakshminarayanan et al. 2017) and MAP produce a good fit but underestimate the predictive uncertainty. Multiplicative Normalizing Flows (MNF) (Louizos and Welling 2017) and Bayes by Backprop (BbB) (Blundell et al. 2015) show a high predictive uncertainty but underfit the data.

Applying approximate Bayesian inference with neural networks was first studied by MacKay (1992) and Neal (1995). Both remain relevant today with the Laplace approximation (MacKay 1992) being one of the easiest to use Bayesian approximations to date and Hamiltonian Monte Carlo (Neal 1995) being one of the most widely employed Monte-Carlo methods. However, both methods scale poorly to current neural network architectures due to the computational burden induced by the high dimensionality of the weight space.

4.1.1 Related Work

More recent methods improve the scalability of approximate inference methods and the complexity of the approximations, introducing sampling-based methods like SGLD (Welling and Teh 2011) or other minibatch-based sampling methods (Ma et al. 2015). Graves (2011) proposed a simple, but biased method to perform variational inference with a fully factorized posterior distribution. This work was extended by Blundell et al. (2015) using the reparametrisation trick from Kingma and Welling (2014) and scale-mixture priors to build a Gaussian variational approximation to the true posterior. An Expectation Propagation (Minka 2001) based approach using a fully factorized posterior approximation was proposed by Hernández-Lobato and Adams (2015). Dropout-based (Srivastava et al. 2014) approximate inference methods have been proposed employing Gaussian Dropout (Blum et al. 2015) and Bernoulli Dropout (Gal and Ghahramani 2015). Further, Louizos and Welling (2016) introduced structured posterior approximations, using matrix Gaussians rather than fully-factorized Gaussians as in (Blundell et al. 2015).

Several studies (Krueger et al. 2017; Louizos and Welling 2017) proposed to employ normalising flows to further increase the flexibility of the variational approximation[1]. However, both approaches only employ the high-fidelity approximations as multiplicative factors on otherwise factorised Gaussians (Louizos and Welling 2017) or single delta peaks (Krueger et al. 2017). Only

[1]Krueger et al. (2017) use the term *hypernetwork* to refer to normalising flows rather than the more general concept of weight generating networks from (Ha et al. 2017).

recently, implicit models have been studied under the framework of variational inference (Huszár 2017; Mescheder et al. 2017; Tran et al. 2017), but only Shi et al. (2018) have used them to model weight uncertainty by parametrising weight matrices as outer product of two vectors.

In contrast to Bayesian inference methods, frequentist approaches have recently been proposed. A bootstrap-based approach was proposed by Osband et al. (2016), whereas Lakshminarayanan et al. (2017) use ensembles of deep networks to calculate predictive uncertainties based on the sample difference due to different initialisation and noise in the stochastic gradients. Even though deep ensembles are straightforward to train, their main disadvantages are that the computational cost scales linearly with the amount of networks in the ensemble and that their uncertainty estimation solely relies on noise during training rather than estimating the full posterior distribution of the weights of the neural network.

4.1.2 Contributions

We introduce the concept of hypernetworks (Ha et al. 2017) into the framework of implicit variational inference (Huszár 2017; Mescheder et al. 2017; Shi et al. 2018; Tran et al. 2017). Our method, *Bayes by Hypernet* (*BbH*) reinterprets hypernetworks (Ha et al. 2017) as implicit distributions similar to generators in generative adversarial networks (Goodfellow et al. 2014) and use them to approximate the posterior distribution of the weights of a neural network. Hypernetworks are able to model a wide range of distributions and can therefore provide rich variational approximations. Furthermore, the hypernetworks are inherently able to learn complex correlations between the weights as they generate samples of multiple weights at the same time. *BbH* avoids hand-crafted strategies of building variational approximations and instead exploits the inherent capabilities of learned approximations to model rich, varied distributions. We show that compared to other Bayesian methods, *BbH* achieves competitive performance: *BbH* demonstrates comparable predictive accuracy without compromising predictive uncertainty, while being the least vulnerable against adversarial attacks. We note that concurrent works (Henning et al. 2018; Ratzlaff and Li 2019; Ukai et al. 2018) propose similar methods that employ hypernetworks to model implicit distributions of neural network weights. However, none of the mentioned approaches deals with the direct approximation of the KL divergence between the posterior weight distribution and the prior weight distribution.

4.2 Bayes by Hypernet

Given a dataset $\mathcal{D}$ with data points $(x_1, y_1), \ldots, (x_n, y_n)$ variational inference for Bayesian neural networks aims to approximate the posterior distribution $p(\mathbf{w} \mid \mathcal{D})$ of the weights $\mathbf{w}$ of a neural network. Given this distribution we can estimate the posterior prediction $\hat{y}$ of a new data point $\hat{x}$ as $p(\hat{y} \mid \hat{x}, \mathcal{D}) = \mathbb{E}_{w \sim p(\mathbf{w}|\mathcal{D})}[p(\hat{y} \mid \hat{x}, \mathbf{w})]$. Because exact Bayesian inference is usually intractable in neural networks we find a variational approximation $q(\mathbf{w} \mid \theta)$ with parameters θ that minimises

the evidence lower bound (ELBO):

$$\begin{aligned}\theta^* &= \arg\min_\theta KL(q(\mathbf{w} \mid \theta) \| \, p(\mathbf{w} \mid \mathcal{D})) \\ &= \arg\min_\theta KL(q(\mathbf{w} \mid \theta) \| \, p(\mathbf{w})) - \mathbb{E}_{w \sim q(\mathbf{w}|\theta)}[\log p(\mathcal{D} \mid \mathbf{w})] \\ &= \arg\min_\theta \mathbb{E}_{w \sim q(\mathbf{w}|\theta)} \left[\log \frac{q(\mathbf{w} \mid \theta)}{p(\mathbf{w})} - \log p(\mathcal{D} \mid \mathbf{w}) \right] \end{aligned} \tag{4.1}$$

Recent works improved upon the Laplace approximation (MacKay 1992) by using the reparametrisation trick (Kingma and Welling 2014) or stochastic backpropagation (Rezende et al. 2014). One of the first works to combine the reparametrisation trick with variational inference for Bayesian neural networks used fully factorised Gaussians (Blundell et al. 2015) to model the approximative distribution. This allows for straightforward optimisation but can only model unimodal distributions in the high dimensional weight space of neural networks.

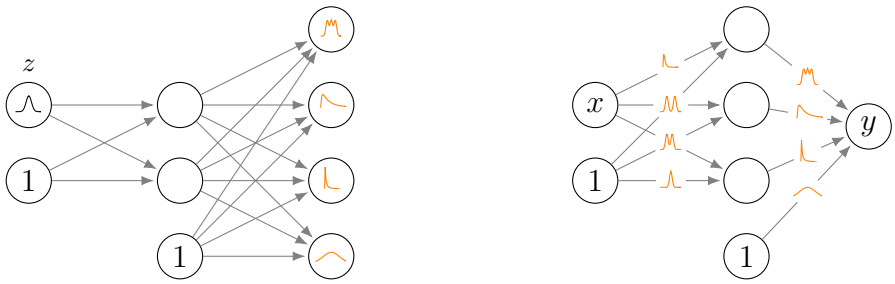
4.2.1 Complex variational approximations

Various works have since proposed different extensions to allow for richer approximations such as mixture of delta peaks (Gal and Ghahramani 2015) or Matrix Gaussians (Louizos and Welling 2016). Nevertheless, those approximations are far from optimal as the true posterior will likely be more complex than delta peaks or correlated Gaussians (Louizos and Welling 2016). Recently, normalising flows have been proposed to allow for more complex approximative distributions (Krueger et al. 2017; Louizos and Welling 2017). Normalising flows use bijective functions with learnable parameters and simple Jacobians to transform samples from simple densities into more complex distributions. By stacking multiple of those similar to the layers of neural networks it is possible to resemble highly complex distributions (Rezende and Mohamed 2015). However, both Multiplicative Normalising Flows (MNF) and Bayesian Hypernetworks (Krueger et al. 2017) only use normalising flows as multiplicative factors of variation and model the majority of the weights with factorised Gaussians (Louizos and Welling 2017) or regular point estimates (Krueger et al. 2017). This parametrisation limits the relations between weights that are able to be modelled while the parametrisation of Louizos and Welling (2017) requires an auxiliary inference network.

4.2.2 Hypernetworks as implicit distributions

Implicit distributions are distributions that may have intractable probability densities but allow for easy sampling. They enable simple calculation of approximate expectations and their corresponding gradients (Huszár 2017). Probably the most well-known group of deep implicit distributions are generative adversarial networks (Goodfellow et al. 2014) that can transform a sample from a simple noise distribution into high-fidelity images.

Using an implicit distribution to model the weights of a neural network requires a generator that is able to capture inherent complexity of neural network weights. Hypernetworks (Ha et al. 2017)



(a) *Hypernetwork G* **(b)** Main Network

Figure 4.2: Illustration of the components of Bayes by Hypernet: The *hypernetwork* G takes a sample $z \sim p(z)$ and converts it into a sample of the weights w of the main network. The *hypernetwork* in Fig. 4.2a generates samples of the weights of the second layer of the main network. The main network takes a data sample x and generates an output y using the weight samples generated by the *hypernetworks*.

are shown to be able to generate weights of networks like ResNets or RNNs while still achieving competitive state-of-the-art performances. Let G be a hypernetwork with parameters θ. Further, let z be an input vector to the hypernetwork G that contains information about the weight w to generate. Then weights w of the main network are generated as $w = G(z \mid \theta)$.

When z is a sample from a simple auxiliary random variable the hypernetwork resembles a generator within the GAN framework. Rather than generating high-fidelity image samples, our generator predicts samples of the weight distribution of the main network. An illustration of the combination of hypernetwork and main network is shown in Fig. 4.2. The graphs shown resemble distributions the auxiliary variable z or the weight samples w could be drawn from.

In the original work (Ha et al. 2017), hypernetworks were introduced as means of weight sharing and therefore network compression. Here, we do not focus on compression, but use arbitrary neural networks as hypernetworks. This is different from the terminology presented in (Krueger et al. 2017) where stacked normalising flows are called hypernetworks.

4.2.3 Estimating the Evidence Lower Bound

Because implicit distributions do not have tractable probability densities, the prior matching term $KL(q(w \mid \theta)||p(w))$ of the ELBO becomes intractable. Previous works (Huszár 2017; Mescheder et al. 2017; Tran et al. 2017) describe how to perform variational inference with implicit distributions. The proposed approaches closely follow the structure of adversarial training, where a generator $w = G(z \mid \theta)$ models the variational distribution $q(\mathbf{w} \mid \theta)$ and a discriminator D estimates the log density ratio from Eq. (4.1). Here, z is an auxiliary noise variable $z \sim p(z)$ which may also contain additional conditioning information.

Specifically, we follow the notion of prior-contrastive adversarial variational inference as formulated by Huszár (2017) and estimate the density ratio in Equation 4.1 using logistic regression. This

enables a two-step update procedure with

$$\mathcal{L}(D \mid G) = \mathbb{E}_{w \sim G(z|\theta)} \log D(\mathbf{w}) + \mathbb{E}_{w \sim p(\mathbf{w})} \log(1 - D(\mathbf{w})) \tag{4.2}$$

$$\mathcal{L}(G \mid D) = \mathbb{E}_z \log \frac{D(G(z \mid \theta))}{1 - D(G(z \mid \theta))} - \mathbb{E}_z \log p(\mathcal{D} \mid G(z \mid \theta)) \tag{4.3}$$

where Equation 4.2 and Equation 4.3 are being used to update the discriminator and the generator, respectively[2]. In theory, the discriminator only gives exact gradients when it has converged to an optimal solution, but GAN training shows us that non-converged discriminators can still provide useful gradients.

However, it is not straightforward to employ adversarial training in high-dimensional spaces such as neural network weights which can accrue to hundreds of thousands or millions of parameters[3]. This number of input dimensions raises computational issues as it spans huge weight matrices when dense layers are employed. We therefore propose to treat all weights independently and estimate the density ratio by a single discriminator. We compare this approximation to the analytical form of Bayes by Backprop (BbB) (Blundell et al. 2015) and find that the single discriminator is not capable of estimating the density ratios correctly. Instead we find that estimating the density ratio using a kernel method (Pérez-Cruz 2008) yields results that are close to those of the analytical method. Specifically, we use the formulation from Jiang (2018), approximating the KL divergence $KL(q(w \mid \theta)||p(w))$ as

$$KL(q(w \mid \theta)||p(w)) = \frac{d}{n} \sum_{i=1}^{n} \log \frac{\min_j ||w_q^i - w_p^j||}{\min_{j \neq i} ||w_q^i - w_q^j||} + \log \frac{m}{n-1}, \tag{4.4}$$

where d is the dimensionality of the samples w, n is the number of approximate samples, m is the number of prior samples, and w_q and w_p are samples from the approximate posterior and prior respectively. This resembles a ratio of the nearest neighbour distance between samples from the variational and the prior distribution and the nearest neighbour distances between samples of from the variational distribution.

4.3 Experiments

We aim to assess the predictive accuracy of our method, and also its ability to estimate the predictive uncertainty. We closely follow established benchmarks (Louizos and Welling 2017) by comparing the performance of *BbH* on the MNIST digit classification task and an adaptation on CIFAR10 classification. Additionally to accuracy, we test the capability of the entropy of the softmax outputs

[2] See *Variational Inference and Density Ratios* section on https://www.inference.vc/variational-inference-with-implicit-probabilistic-models-part-1-2/

[3] The LeNet that we use in a later experimental section already has more than 400,000 weights. In comparison, most image datasets used in deep learning do not have more than 196,608 ($256 \times 256 \times 3$) dimensions.

to detect outliers and evaluate the method's robustness against adversarial examples. Methods are compared in their predictive uncertainty on test set (in-dataset examples) and on similar, yet unseen data (out-of-dataset examples). An optimal method would predict low entropy and correct predictions for the in-dataset examples and high entropy for out-of-dataset examples. The high entropy predictions on unseen data can be important in real-life decision processes as they can be used to trigger a request for human support. Similarly, by testing the robustness against adversarial attacks, we expect to see the degradation of accuracy and a simultaneous increase in predictive uncertainty. In contrast to previous works (Lakshminarayanan et al. 2017; Louizos and Welling 2017), we do not rely on (cumulative) density plots of the entropy, but rather calculate the area under the receiver operation characteristic (AUROC) to provide a quantitative measure of the ability to detect when the model does not know the correct prediction.

We generate the weights of each layer using a unique hypernetwork. This allows for enough capacity to fit complex weight distribution. Each hypernetwork is implemented as a 3-layer fully-connected network with $[64, 256, 512]$ units, as we did not find a general improvement by adding more layers or units. Further, we employ a standard normal prior for all methods (excluding ensembles) and treat all weights as independent. The experiments are implemented in Tensorflow (Abadi et al. 2016) and optimisation is performed with Adam (Kingma and Ba 2015). We follow the code from (Louizos and Welling 2017) and use a learning rate of $\eta = 0.001$ for all methods apart from *BbH*. For *BbH* we use $\eta = 0.0001$. We compare our method to MC-Dropout (Gal and Ghahramani 2015) (dropout rate $\pi = 0.5$), Bayes by Backprop (BbB) (Blundell et al. 2015), deep ensembles (Lakshminarayanan et al. 2017), multiplicative normalizing flows (MNF) (Louizos and Welling 2017), and maximum a posteriori (MAP) training. We do not compare to Bayesian Hypernetworks (Krueger et al. 2017), because they are very similar to MNF. We train the deep ensembles without predictive uncertainty as we found it to sometimes result in numerically unstable training. MNF and *BbH* anneal the KL term during training which follows the original MNF implementation. All methods use 100 posterior samples to estimate the predictive distribution and we use 5 samples to estimate the KL from Eq. (4.4).

4.3.1 MNIST Digit Classification

We reproduce the setup from (Louizos and Welling 2017), employing a LeNet for MNIST classification and notMNIST as outlier dataset. Additionally to the comparison with other methods, we use MNIST to test the newly introduced hyperparameters of *Bayes by Hypernet*: The shape of the auxiliary variable z as input to the hypernetwork, and whether or not to use the same sample $z \sim p(z)$ across different weights that are generated at the same time. Lastly, we use the recently proposed MorphoMNIST dataset (Castro et al. 2019) to generate data with a controlled difference to the original dataset and test the outlier detection on that.

Approximating the KL divergence: We run Bayes by Backprop to compare the results of approx-

Table 4.1: Comparing Adversarial Variational Bayes (AVB) and a kernel-based estimation of the KL divergence: AVB produces not only worse accuracies but also the most overconfident results. The kernel estimates achieve results close to the analytical solution.

	Error [%]	AUROC [%]
Analytical	0.72	99.2
AVB	0.97	85.6
Kernel Estimate	0.88	98.7

Table 4.2: Comparison of different auxiliary noise configurations: A higher degree of noise increases the predictive uncertainty (lower outlier AUC) but does not demonstrate a trend in the corresponding accuracy.

	Error [%]	AUROC [%]
Shared, $d = 1$	1.11	97.2
Independent, $d = 1$	1.41	92.2
Shared, $d = 8$	0.90	98.6
Independent, $d = 8$	0.81	98.7
Shared, $d = 64$	1.71	97.2
Independent, $d = 64$	1.45	96.8

imating the KL divergence using an adversarial approach or kernel-based approach with the analytical solution. We use the same settings for BbB and train the discriminator for 100 steps before starting training of the main network and then train it for 5 steps for every step we train the main network. The results are shown in Table 4.1. We find that the kernel estimation achieves results closest to the analytical ones. Adversarial Variational Bayes (AVB), however, provides the worst accuracy and most overconfident predictions. The multi-step training procedure of AVB also causes the longest runtime.

Auxiliary Noise: The hypernetworks take an auxiliary noise variable z as an input and transform it into a sample of the weight distribution. In all experiments, we draw samples from unit-variate Gaussians as z. However, the dimensionality d of z can influence the capacity of the hypernetwork. Further, hypernetworks can be coupled by drawing the same sample z for each weight or can be decoupled by drawing a different sample for each weight. The former enables the hypernetworks to learn more complicated relations across different parts of the weights, whereas the latter leads to higher variability across the generated weights. Table 4.2 shows the results of different noise configurations. We find that a too low or too high dimensionality of z deteriorates both the accuracy and outlier detection. Coupling the weights of multiple layers by sharing z leads to more accurate outlier detection. This could be explained by the capability of modelling more complex relations between the weights. The independent noise requires the different layers to be resilient

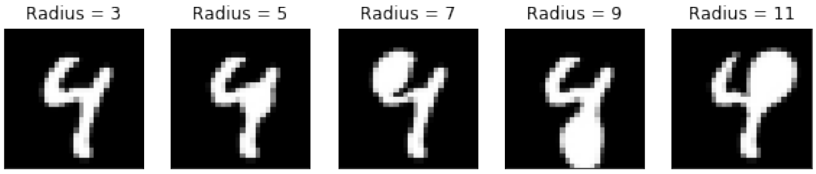


Figure 4.3: Example images of the morphologically altered MNIST digits. A higher radius of the swelling leads to a bigger deformation.

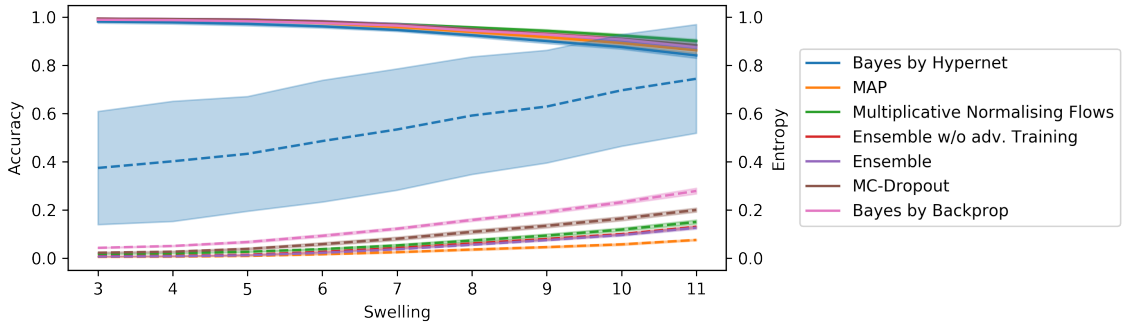


Figure 4.4: Performance of compared methods on MorphoMNIST with swelling: The solid line depicts the accuracy and the dashed line the predictive entropy relative to the maximum entropy as a function of the swelling.

to different noise values and therefore acts as a regulariser. However, the exact relation between the dimensionality of the noise and the performance metrics remains to be explored by further experimentation.

Classification and outlier detection performance: We decide to use *BbH* with $z \in \mathcal{R}^8$ and coupled weights for the following experiments. We compare the performance of *BbH* with several established variational Bayesian deep learning techniques and frequentist approaches. The results are shown in Table 4.3. All methods achieve comparable accuracy, with MC-Dropout and deep ensembles being the best. However, *BbH* exhibits a good trade-off between predictive accuracy and predictive uncertainty. Its outlier detection performance is only outperformed by Bayes by Backprop and deep ensembles.

We use MorphoMNIST (see example images in Fig. 4.3) to better understand the outlier detection capabilities of each methods. We generate outlier images with controlled morphological deformations. By increasing the strength of the deformation we expect the methods to drop a little in accuracy but acknowledge the difference in data by increasing the predictive entropy. We use the swelling deformation with a strength of 3 and increasing radii within $[3, 11]$. We plot the accuracy and entropy relative to the maximum entropy in Fig. 4.4. We see that all methods perform as expected. The loss in accuracy is comparable across all methods, with *BbH* having the highest decrease. However, *Bbh* exhibits an increase in predictive entropy that is significantly higher than the other methods.

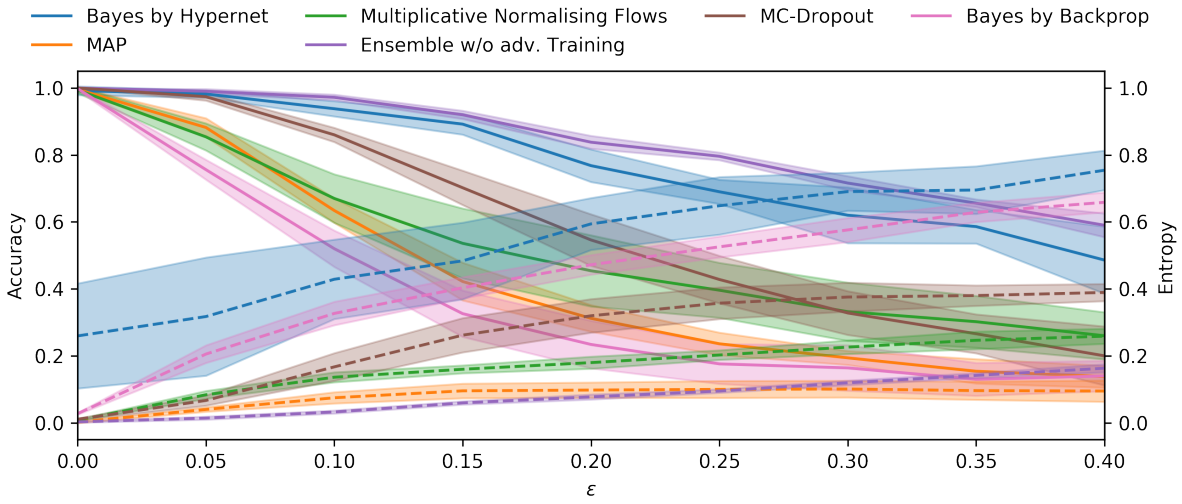


Figure 4.5: Performance of compared methods exposed to adversarial attacks on the MNIST dataset: The solid line depicts the accuracy and the dashed line the predictive entropy relative to the maximum entropy as a function of the adversarial perturbation.

Robustness against adversarial attacks: We employ the fast sign method (Goodfellow et al. 2015) on the first 1000 samples of the MNIST test set. In this experiment we compare to deep ensembles without adversarial training for a fair comparison. For variational methods, we generate the adversarial samples as an average of 100 posterior samples to account for the variation during predictions. We plot the accuracy and entropy relative to the maximum entropy in Fig. 4.5. *BbH* performs the best with the slowest decrease in performance coupled with the highest predictive uncertainty. The other methods exhibit varying degrees of decay and predictive uncertainty with BbB having the steepest decay but also a high entropy.

4.3.2 Scalability To Deep Architectures

To test the scalability of *BbH*, we run experiments using a ResNet-32 (He et al. 2016a) on the same CIFAR5 task as in (Louizos and Welling 2017). We train on the first five classes of CIFAR10 and use the remaining as outlier dataset. The results in Table 4.3 show that deep ensembles achieve the best accuracy and outlier detection performance, but are also the most compute intensive. *BbH* achieves a good accuracy that is only outperformed by the non-Bayesian methods, whereas its outlier detection performance is only subpar to ensembles and MNF.

Performing the same adversarial robustness experiment as with MNIST (see Fig. 4.6), we find that *BbH* is the most robust against adversarial attacks. However, the entropy seems relatively constant for all methods and MC-Dropout and BbB exhibits higher predictive entropies than *BbH*.

Table 4.3: Results on classification task on CIFAR5 (first 5 CIFAR10 classes) and MNIST datasets: Error [%] shows the classification error on the test set. The AUROC is the area under the receiver operation characteristic detecting outliers given the entropy of the softmax output on the corresponding data set (higher is better). Compared methods: maximum a posteriori training (MAP), deep ensembles (Lakshminarayanan et al. 2017) (Ensembles), Bayes by Backprop (Blundell et al. 2015) (BbB), MC-Dropout (Gal and Ghahramani 2015) (Dropout), Multiplicative Normalizing Flows (Louizos and Welling 2017) (MNF) and Bayes by Hypernet (*BbH*).

	CIFAR5		MNIST	
	Error [%]	AUROC [%]	Error [%]	AUROC [%]
MAP	12.36	71.4	0.70	98.5
Ensemble	10.06	73.9	0.42	98.9
BbB	14.12	70.8	0.88	99.5
Dropout	16.76	69.3	0.42	97.9
MNF	13.36	72.6	0.84	98.4
BbH (Ours)	12.90	72.0	0.90	98.6

4.3.3 Examining The Posterior Distributions

We examine the posterior distributions fitted to the the toy task from Fig. 4.1 and the corresponding covariance matrices in Fig. 4.7. This allows us to compare the variational approximations of *BbH* and MNF to samples drawn by using HMC. We find that HMC fits very Gaussian-like distributions, whereas MNF and *BbH* fit more complex, multi-modal distributions. We argue that HMC is underfitting the data and mostly resembling the prior. We find that HMC finds much more complex distributions when more data points are given. We find that even though normalising flows are capable of modelling highly complex distributions, the variational posterior still closely resembles a Gaussian. We attribute this to the multiplicative nature of MNF with the underlying Gaussian distributions. Weight distributions of all methods are shown in Appendix A.1.

MNF and *BbH* find relatively simple distributions that resemble skewed Gaussians for the MNIST and CIFAR tasks (see Appendix A.2 and Appendix A.3). This might be caused by the higher dimensionality of the space, where the variational inference methods only fit one mode. Nevertheless, both methods still model covariances between the weights. However, MNF exhibits limitations on which covariances it can model because of its multiplicative nature.

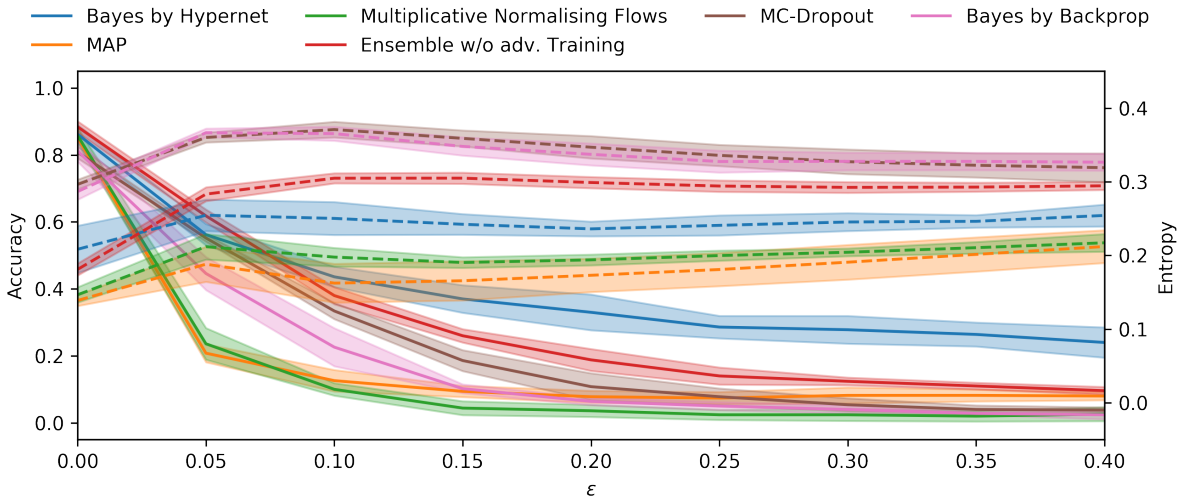
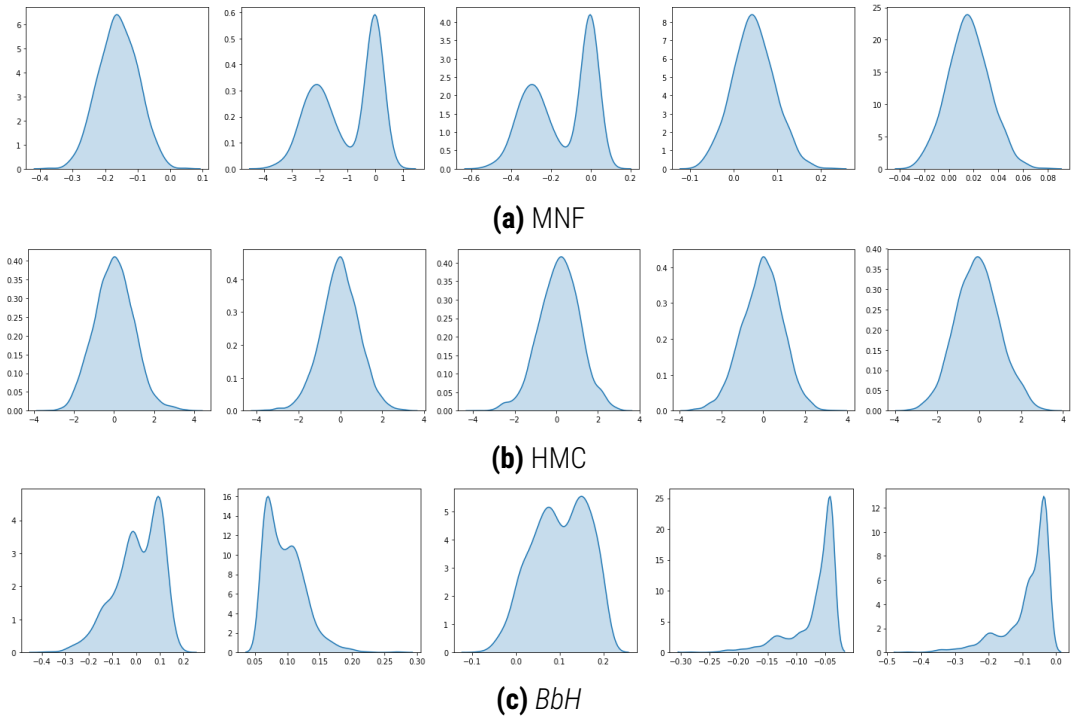


Figure 4.6: Performance of the methods exposed to adversarial attacks in the CIFAR5 domain: The solid line shows the accuracy and the dashed line the predictive entropy relative to the maximum entropy as a function of the adversarial perturbation.



(a) MNF

(b) HMC

(c) *BbH*

Figure 4.7: Illustration of the posterior distributions approximated by MNF Louizos and Welling (2017) (a), HMC (b), and BbH (c) for the 2-layer network used for the toy regression task in Fig. 4.1. BbH and MNF generate complex, multi-modal approximations whereas HMC's resemble simple Gaussian-like distributions.

4.4 Discussion & Conclusion

BbH interprets *hypernetworks* (Ha et al. 2017) as an implicit distribution, which we employ as approximate distributions within variational inference. In our experiments, we demonstrate that *BbH* yields strong predictive performance with competitive uncertainties. *BbH* finds a good trade-off between accuracy and predictive uncertainty on MNIST and CIFAR and is the most robust method against adversarial examples. Compared to other Bayesian methods *BbH* yields comparable or better accuracy and comparable uncertainties.

This work is the first to report an extensive comparison on Bayesian methods for ResNet architectures. It demonstrates competitive accuracy and predictive uncertainty of *BbH* paired with the best robustness against adversarial attacks. *BbH* qualitatively produces more complex approximate posterior distributions (Fig. 4.7) compared to MNF, even though MNF should enable similarly complex distributions. This translates to the complex correlations of the weights that are modelled by *BbH*. However, both those methods approximate the posterior to be more complex than HMC which lends itself to a deeper investigation into the 'true' shape of posterior distributions of the weights of neural networks.

Interestingly, we find that our baseline implementations perform better on MNIST than reported in (Louizos and Welling 2017). This suggests that many of those methods require more careful hyperparameter tuning to offer reliable comparisons. Furthermore, even though *BbH* clearly models more complicated posterior distributions than MNF or *BbB*, it does not always yield the highest predictive uncertainties. This raises questions whether richer variational approximations always lead to better results and should be further investigated. Related results have been shown by Rainforth et al. (2018) who explore tighter variational bounds on the autoencoder setting, or Farquhar et al. (2020) who explore a trade-off between depth of the main neural network and the complexity of the variational posterior.

Furthermore, we notice that all methods exhibit almost constant entropy values for the adversarial example task on CIFAR5, which questions the reliability of this dataset. Previous works have either used CIFAR5 as in-dataset examples and the other 5 classes as outliers (Louizos and Welling 2017) or CIFAR10 as in-dataset examples and SVHN as outliers (Lakshminarayanan et al. 2017). We believe that both settings are not perfect as SVHN seems obviously different from CIFAR10, whereas the classes within CIFAR10 share a lot of similarities that make the task to distinguish them harder. We argue that a carefully curated dataset might be needed to explicitly test different 'degrees' of differences between in-distribution and out-of-distribution samples. As an example for that we show that a controlled degree of distribution-shift using MorphoMNIST (Castro et al. 2019) enables better interpretable experiments (see Section 4.3.1).

Our reported experiments only use a single architecture to build an implicit distribution using hypernetworks. Even though we run some initial experiments with different settings we quickly settled

on the layer-wise generation of the weights. We argue that a wider investigation into different architectures could yield further performance gains and insights, such as (Ratzlaff and Li 2019). However, the size of weight matrices in modern neural networks restricts the design choices due to limited GPU memory.

Future directions: We believe this work opens a variety of directions for future work: The parametrisation of the weights of the main network could be extended to find more efficient and richer forms. This might extend to *dynamic hypernetworks*, which generate weights conditioned on the input to the main network. *BbH* enables the use of highly complex priors as it is merely required to sample from them (*e.g.* task-specific priors instead of Gaussian priors, which are subject to known limitations (Neal 1995), can be examined). This idea can be extended to transfer learning, where not only the weights, but also previously trained posterior distributions can be transferred and used as a new prior. It would also be interesting to combine *BbH* with neural architecture search methods like SMASH (Brock et al. 2018) to build Bayesian approximations of the posterior over the architectures.

Conclusions: In this chapter, we proposed and extensively evaluated *Bayes by Hypernet*, a new approach for obtaining uncertainty estimates with neural networks. The appropriation of *hypernetworks* to generate weight distributions allows for modelling arbitrary complex distributions and the proposed method naturally integrates with modern deep learning, addressing the need for certainty measures in real-world applications.

Chapter 5

Deep Generative Models for Outlier Detection

This chapter is based on the following publications:

(a) **Pawlowski, N.** et al. (2018). "Unsupervised Lesion Detection in Brain CT using Bayesian Convolutional Autoencoders". In: *Medical Imaging with Deep Learning Abstract track* – (Pawlowski et al. 2018)

(b) Chen, X.*, **Pawlowski, N.***, Rajchl, M., Glocker, B., and Konukoglu, E. (2018b). "Deep generative models in the real-world: An open challenge from medical imaging". In: *arXiv preprint arXiv:1806.05452* – (Chen et al. 2018b)

(c) **Pawlowski, N.** and Glocker, B. (2021). "Abnormality Detection in Histopathology via Density Estimation with Normalising Flows". In: *Medical Imaging with Deep Learning Short Paper Track* – (Pawlowski and Glocker 2021)

The work for (b) was completed with shared first-authorship with equal contribution from Nick Pawlowski and Xiaoran Chen. Both authors contributed to all aspects of the work with main responsibilities as follows: N.P. led the implementation of the Bayesian models as well as the probabilistic formalisation of the problem; X.C. had main responsibilities on the experimentation with the regular deep learning models as well as the visualisation of the results. Both ideated the approach independently in (Pawlowski et al. 2018) and (Chen and Konukoglu 2018) and had equal responsibility in conceptualisation, and writing of the manuscript.

The previous chapter tackled the idea of *"knowing what one does not know"* and introduced a Bayesian deep learning method (*Bayes by Hypernet*) to equip neural networks with the ability to

provide uncertainty estimates for the prediction they are making. This approach relied on the availability of labels to perform a task which allows for saying *"I don't know the answer"*. However, there can be different reasons for lack of an answer: it could be that the sample is ambiguous, a lack of examples during training can lead to conservative estimates, or, the sample could be completely different from anything that was observed during training. This last scenario is a special one, as it does not require a predictive task to tell whether something is different to previous observations and depending on the exact setting can be called abnormality detection, outlier detection, novelty detection, or out-of-distribution detection.

In a very rough approximation, the task of diagnosing an illness is very similar to identifying an image or a part of an image to be unusual – *e.g.*in a generally healthy population one will mostly observe healthy brain scans and very few unhealthy ones. If one knows what the variations in healthy brains looks like, one is capable of flagging abnormal areas and can direct an expert to investigate further. This is especially powerful in the scenario of rare or heterogeneous diseases where there is too little data to train a supervised model but it is possible to refer suspicious cases.

This chapter frames different medical imaging tasks as outlier detection problems and studies the use of generative models to differentiating healthy from unhealthy cases. The first two sections tackle the detection of lesions on 3D brain scans using variational autoencoders (VAEs; see Section 2.3.2). The first section aims to detect lesions in brain MR scans using VAES. Due to their nature, MR images suffer from inconsistent appearances between different imaging sites, protocols, and scanners. The approach of using deep generative models for detecting outliers in brain MRI is limited due to the domain shift between the datasets for healthy and unhealthy samples. We find that relatively simple non-deep learning methods achieve the best performance on this task. The second section applies the same methodology as the first and shows that VAEs can find lesions in brain CT scans. However, VAEs are often outperformed by very simple baseline methods due to the consistent intensity profile of CT scans. The last section investigates the classification of histopathology patches into healthy or unhealthy tissue based on normalising flows (see Section 2.3.1). Again, we find that the naive application of deep learning models yields subpar results. However, this can be addressed by the use of more complicated measures to detect whether a sample is abnormal or not.

5.1 Detecting Outliers in MR images

Learning high-dimensional data distributions from finite number of examples and being able to generate new samples from such distributions is a challenging task. Developments in deep learning based techniques and unsupervised learning in the last five years set a new standard for this problem, especially for imaging data. Generative adversarial networks (GANs) (Goodfellow et al. 2014), variational auto-encoders (Kingma and Welling 2014; Rezende et al. 2014) and variants of these models (Arjovsky et al. 2017; Karras et al. 2018; Makhzani et al. 2015; Radford et al. 2015)

demonstrate that it is possible to train networks that can approximate distributions of images well enough to sample realistic looking sharp images. Such models have already been successfully applied in various vision tasks, such as generating data samples (Karras et al. 2018), domain adaptation (Chen et al. 2018c; Tzeng et al. 2017) and image in-painting (Yeh et al. 2017).

Unsupervised learning and generative modeling have numerous clinically relevant applications in medical image computing. One particular application, *unsupervised abnormality detection*, is scientifically interesting and technically challenging. The task is simple to state: given an image acquired from a patient, detect the regions in the image that should not be there in the 'normal' case, if any. This is one of the routine tasks of a radiologist that they need to perform for every image they assess and a critical first step in diagnosis. For complicated cases, years of experience are necessary to distinguish normal from abnormal. However, for a large set of problems, such as brain tumours, even non-experts can perform the task after seeing a handful of 'normal' looking images. Despite the simplicity of its description and the clear separation of abnormal from normal tissue appearance, unsupervised abnormality detection remains as a huge challenge for machine learning.

Deep learning based generative modeling approaches provide new opportunities for developing automatic algorithms for unsupervised abnormality detection. In this work, we empirically investigate feasibility of such approaches using relatively large, publicly available datasets. We use magnetic resonance images (MRI) of the brain acquired from healthy individuals at different age groups to train different auto-encoder based generative models to learn the distribution of 'normal' brain MRI. Then we apply the trained models on two other datasets of brain MRI bearing tumours and stroke lesions to detect the abnormal lesions in an unsupervised manner. Detection performance is a good quality indicator that assesses how well the models learn the distribution of 'normal' images. We describe the datasets and present empirical evaluation comparing different deep learning based models as well as non-deep learning methods, which have been used in the medical image computing community. The evaluation provides a benchmark showing the state-of-research for this difficult problem, indicating that unsupervised detection of abnormalities remains an open challenge and demands further research.

5.1.1 Related work

Related work on deep generative models: Literature on generative modeling with neural networks date back to MacKay's work (MacKay 1995). More recent, generative adversarial networks (GANs) (Goodfellow et al. 2014) and variational autoencoders (VAEs) (Kingma and Welling 2014; Rezende et al. 2014) have shown the feasibility of generative modeling with deep models. Further modifications suggested in (Arjovsky et al. 2017; Gulrajani et al. 2017; Radford et al. 2015) improved the performance and stability of GAN training, and achieve realistic data generation such as (Karras et al. 2018). Alternatively, VAEs mostly rely on a reconstruction loss that is known to lead to blurry

reconstructions (Larsen et al. 2016). In contrast to GANs, VAE-based works mainly focus on the latent variable model (Gregor et al. 2015; Yan et al. 2016) and how to disentangle the latent variables (Chen et al. 2018a). Many publications propose methods to improve the expressibility of this model (Burda et al. 2015; Kingma et al. 2016; Rezende and Mohamed 2015).

Lesion detection and segmentation: Detection of brain lesions is a critical step to diagnose diseases such as cranial trauma, abscesses and cancer. Traditionally, radiologists manually detect and segment lesions slice by slice. However, the large resolution of the 3D images and high level of required expertise have made it a time-consuming and expensive task to accomplish. Studies such as (Prastawa et al. 2004), (Ayachi and Amor 2009), and (Zikic et al. 2012) have suggested supervised methods for automatic detection of brain lesions. Due to the importance of the application, the medical image computing community has been hosting public challenges specifically for lesion detection, the Multi-modal Brain tumour Image Segmentation (BRATS) and Ischemic Stroke Lesion Segmentation (ISLES). A benchmark (Bauer et al. 2012) was released to evaluate and compare existing models. With the introduction of fully convolutional neural networks (FCNs) (Long et al. 2015), DeepMedic (Kamnitsas et al. 2017a) and U-Nets (Ronneberger et al. 2015), the field has since then been dominated by deep learning approaches achieving the highest accuracy on most if not all imaging challenges.

Supervised methods however, are specific to certain lesions and need to be trained with respective example data. Furthermore, their application on previously unseen lesions is not straightforward.

On unsupervised lesion detection: Unsupervised detection of abnormalities has been an important topic in medical image computing. Non-deep learning based methods have been proposed over the last two decades that use mixture modeling and expectation maximization (Van Leemput et al. 2001), atlas-registration (Prastawa et al. 2004) and probabilistic models that utilize image registration (Tomas-Fernandez and Warfield 2015; Zeng et al. 2016).

Deep-learning based models have also been recently applied to abnormality detection following related developments in computer vision (An and Cho 2015; Chalapathy et al. 2017). Schlegl et al. used GANs to detect abnormalities in (Schlegl et al. 2017). Their method was based on determining the best latent space representation of a given image with an abnormality and then computing the difference between reconstruction from this representation and the image. The underlying idea was that GANs trained on healthy images should not be able to reconstruct abnormal lesions. Based on related ideas, more recent work explored the use of autoencoder-based models and detected abnormal regions through high reconstruction errors (Baur et al. 2018; Chen and Konukoglu 2018; Pawlowski et al. 2018; Sato et al. 2018).

5.1.2 Methodology

We approach the lesion detection problem in a way similar to one-class classification, where firstly we model the pixel-wise probability using healthy brain MRI images, then detect lesion regions as pixels with low probability according to the model learned on healthy data. Assume we have a dataset of healthy images $\{X_H^1, \ldots, X_H^N\}$ where each image is a set of pixels $X_H^n \in R^{d \times d} = \{x_H^{(n,1)}, \ldots, x_H^{(n,M)}\}$. Given this dataset, we aim to estimate the distribution of healthy data p_H to evaluate the probability of an unseen image and its pixels $p_H(x^{\{m\}})$. Now, suppose we have another test image $X_A \in R^{d \times d}$ with both abnormal R_A and healthy regions R_H. Because p_H only models the distribution of healthy image, the probability of the pixels in abnormal regions $p_H(x_A^{(p \in R_A)})$ are expected to be low, whereas pixels in healthy regions will have high probability $p_H(x_A^{(p \in R_H)})$.

A naive approach could be to model p_H as a location dependent function with a Gaussian per pixel $p_H(\hat{x}) = p_H(\hat{x} \mid i, j) = \mathcal{N}(\hat{x} \mid \mu_{(i,j)}, \sigma^2_{(i,j)})$, where i and j denote the location of the pixel in the image and $(\mu_{(i,j)}, \sigma_{(i,j)})$ are the parameters of the Gaussian corresponding the pixel (i, j).

A more advanced approach would be to model pixel intensities at each location with a Gaussian Mixture Model (GMM) where the probability $p_H(\hat{x})$ of an unseen pixel is modeled as $p_H(\hat{x}) = \sum_{i=1}^{K} \Phi_i^j \mathcal{N}(\hat{x} \mid \mu_i, \sigma_i^2)$ with the number of mixture components K, the mixture weights $\{\Phi_i^j\}$ that depends on location j, and the parameters of each mixture component (μ_i, σ_i). Given an image, the parameters of this model can be estimated using the Expectation-Maximization algorithm and an atlas image. Furthermore, abnormality detection can be performed by adding an additional component whose $\Phi_o^j p(\hat{x}|o) = \lambda$ is a constant for all pixels and intensities similar to what is proposed in (Van Leemput et al. 2001).

Autoencoder-based models consist of two deterministic mappings, the encoder f_{enc} and the decoder f_{dec}. An input $\hat{x}$ goes through f_{enc} to be encoded into a lower dimensional latent variable z, and then goes through f_{dec} to be decoded back into an reconstruction $\hat{X}' = AE(\hat{X})$ that is based on the latent encoding z. The functions f_{enc} and f_{dec} are then optimized to minimize the reconstruction loss $L(\hat{X}, \hat{X}')$. We chose to employ the frequently used L_2 loss as reconstruction loss $L(\hat{X}, \hat{X}') = ||\hat{X} - \hat{X}'||_2$. Due to the lower dimensionality of z, Autoencoder-based methods are forced to learn a compression of the data that is related to learning a lower manifold representation.

We argue that because the autoencoder relies on this lower dimensional representation it is not capable of reconstructing variations in the data that it has not seen during training. Therefore the reconstruction loss $L(\hat{X}, \hat{X}')$ can be interpreted as an unnormalized probability of a sample belonging to the data distribution $P_H(\hat{X}) = L(\hat{X}, AE(\hat{X}))/Z$ where Z is an unknown normalization constant.

Various adaptations of the basic autoencoder have been suggested. Denoising autoencoders

(DAEs) follow the same concept as regular autoencoders but aim to reconstruct clean images $\hat{X}$ from corrupted images $\hat{X} + \epsilon$. By trying to remove noise from the images, a DAE distinguishes noise from structure in $\hat{X}$, thus better capturing the information in them. However, AE and DAE are not generative models as they do not approximate p_H and merely serve as dimensionality reduction methods. variational AEs (VAEs) and adversarial AEs (AAEs) integrate stochastic inference into the AE framework and enable to approximately model p_H via variational inference. The deterministic mappings f_{enc} and f_{dec} in regular AEs, become probabilistic mappings and model the inference network $q(z \mid X)$ and generative process $p(X \mid z)$ respectively. The distribution is learned in such a way that,

$$p_H(X_H) = \int p(X_H \mid z_H) p(z_H) dz, \tag{5.1}$$

where $p(z_H)$ describes a prior on the latent encoding that constrains z_H to lie in a structured latent space. The structured latent space is imposed on $Q(z \mid X)$ by minimizing a Kullback-Leibler (KL) divergence $KL[q(z|X)\|p(z)]$. The overall model is optimised by maximizing the evidence lower bound (ELBO) (Doersch 2016; Kingma and Welling 2014),

$$\log p_H(X_H) \geq \mathbb{E}_{z \sim q(z_H|X_H)}[\log p(X_H \mid z_H)] - KL[q(z_H \mid X_H)\|p(z)] \tag{5.2}$$

Several studies found VAEs to generate blurry reconstructions (Bousquet et al. 2017; Larsen et al. 2016). We seek alternatives to mitigate the blurriness, as good quality reconstructions are potentially a prerequisite for satisfactory detection outcomes. Bousquet et al. (2017) suggests that AAEs resolve the blurriness by improving the encoder using adversarial learning to match $q(z) = E_{X_H}[q(z|X_H)]$ and $p(z)$ by optimizing a GAN loss,

$$\min_G \max_D \mathbb{E}_{z \sim p(z)}[\log D(z)] + \mathbb{E}_{z \sim Q(z)}[\log(1 - D(z))] \tag{5.3}$$

where the generator G corresponds to the encoder in AAE. To stabilize GAN training, we modify the loss of the original AAE to use the recently proposed Wasserstein distance from WGANs with gradient penalty (WGAN-GP) (Gulrajani et al. 2017).

Another approach to achieve sharper images with AE-based methods is to improve how the image is compared to its reconstruction. In the work of α-GAN, the model matches $Q(z)$ and $P(z)$ in the same way as AAE and adds one more discriminator D_{rec} to distinguish between X_H and X'_H. Again, the addition of D_{rec} introduces an adversarial loss that can be written in a similar form as Eq. (5.3). Here, the decoder acts as the generator in the GAN formulation. The optimization is more complicated due to the modifications above. To train the model, we follow the optimization provided in (Rosca et al. 2017).

Lastly, we employ variational inference to approximate the posterior distributions over the model parameters with factorized Gaussians (Blundell et al. 2015). This allows us to not only marginalize

the latent encoding but also the model parameters when estimating the reconstruction loss of a new image. This might enable more robust reconstructions as it is less reliant on specific model parameters, as shown by (Pawlowski et al. 2018).

5.1.3 Experiments

To give a comprehensive overview of the current state of unsupervised lesion detection, we also include baselines like GMMs and mean image difference. Furthermore, we provide a supervised segmentation baseline using an U-Net (Ronneberger et al. 2015).

5.1.3.1 Data

Cam-CAN[1](Taylor et al. 2017) We use The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) dataset for training which contains T1- and T2-weighted brain MRI of 652 subjects from a uniform age range from 18–87. All subjects are confirmed healthy after radiological assessment.

BraTS-T2w[2](Bakas et al. 2017; Menze et al. 2015) We utilize the T2-weighted images of 285 patients from the Brain tumour Segmentation Challenge (BraTS). The images show high-grade (210) and low-grade (75) glioblastomas which are visible as brighter regions in the images.

ATLAS-T1w[3](Liew et al. 2018) We also make use of the Anatomical Tracings of Lesions After Stroke (ATLAS) dataset containing T1-weighted images of 220 stroke patients. Lesions are visible as darker regions in the images and identified using location information as appearance is similar to normal structures.

5.1.3.2 Preprocessing

Throughout this paper, we consider each volume to be a set of 2D slices and apply all methods to 2D slices rather than full volumes. To reduce the variability across subjects and datasets, each scan is normalized as follows: First, empty slices with no brain are removed, and then, the images are cropped within the maximal boundary computed across the dataset to ensure the same image size, lastly, the images are normalized across all remaining slices to have zero-mean and unit-variance within the brain masks obtained from a skull-stripping process. The models are trained with datasets of two different sizes, 128×128 and 256×256. Resizing is implemented using the `scipy.misc.imresize` with nearest interpolation.

[1] http://www.cam-can.org/
[2] https://www.med.upenn.edu/sbia/brats2018.html
[3] http://fcon_1000.projects.nitrc.org/indi/retro/atlas.html

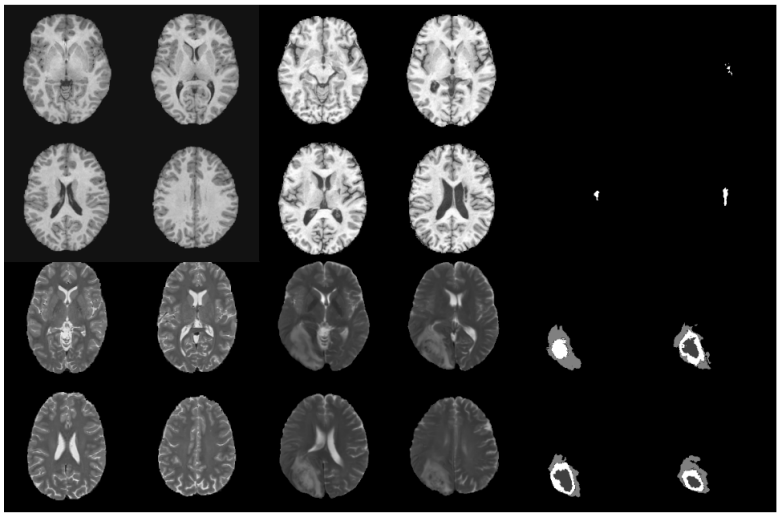


Figure 5.1: Thumbnails of example images from the different datasets. From left to the right, columns 1 and 2 show the CamCAN dataset with the top two rows being T1-weighted images and bottom two rows being T2-weighted images; columns 3 and 4 present examples from the outlier datasets with the top two rows being images from ATLAS-T1w dataset and bottom two rows being images from BraTS-T2w dataset; columns 5 and 6 are ground truth segmentations for the corresponding images in columns 3 and 4.

5.1.3.3 Evaluation

For all difference methods we calculate the difference of a new image $\hat{X}$ and its reconstruction $\hat{X}'$ as the absolute error $\hat{X}_{dif} = |\hat{X} - \hat{X}'|$ instead of the squared error. Note, that this does not change the outcome of the predictions as the ordering of the errors does not change. Then, we rely on thresholding to find abnormal regions. We evaluate $\hat{X}_{dif}$ using the ground truth annotations of lesions. Let the ground truth be Y. We use the following metrics for detection performance evaluation of the trained models,

1. **Area Under the Receiver Operation Characteristics curve (AUC)**. The AUC is calculated as the area under ROC curve due its insensitivity to label imbalance that occurs in our dataset. In particular, we compute the true positive rate (TPR) $TPR = \frac{TP}{TP+FN}$ and false positive rate (FPR) $FPR = \frac{FP}{FP+TN}$.

2. **Maximal Dice Score (mDSC)**. The dice score is commonly used to report segmentation results. To calculate the dice score in our case, it is required to set a threshold t for $\hat{X}_{dif}$ that predicts lesions as $\hat{Y} = \hat{X}_{dif} > t$. While it can be formed into a new question, we use a range of thresholds and calculate a dice score using $\hat{X}_{dif}$ and Y for each threshold. As we do not further explore thresholding, we assume there is an optimal threshold that achieves the maximal dice score (mDSC) on the model. As such, we chose the threshold using brute force optimisation to find the maximal achievable average dice score of a model on the used test set. This is purely for evaluation purposes and also measures the separability of the distributions of the reconstruction error of healthy and abnormal tissue. This method of finding

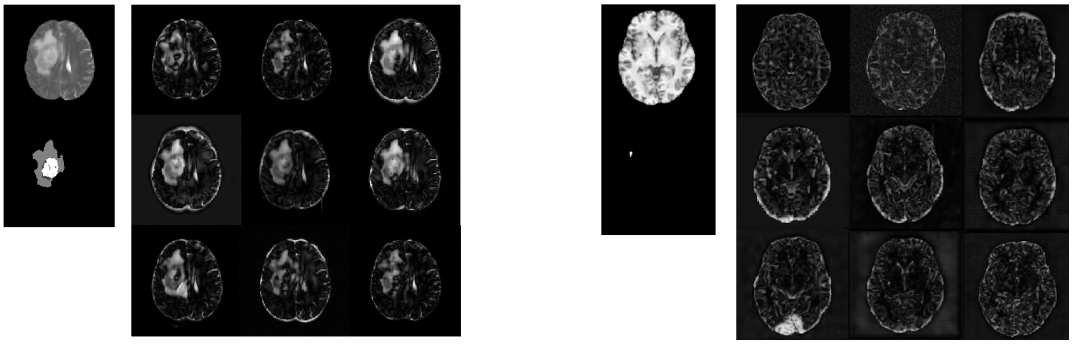


Figure 5.2: Difference maps obtained on BraTS-T2w dataset (left) and ATLAS-T1w (right). For each dataset, the leftmost column shows the original image (top) and ground truth segmentation (bottom). Columns 2-4 show difference maps obtained with autoencoder models as follows, top row: AE, DAE, VAE-128, middle row: VAE-BBB-128, VAE-256, AAE-128, bottom row: AAE-256, α-GAN-128, α-GAN-256. See the main text for further explanations.

a threshold is not applicable in practice and we explicitly leave the question of finding suitable thresholds to future work.

5.1.4 Results

We demonstrate our results by firstly presenting the difference maps as in Fig. 5.2 for visual inspection of the models and then providing quantitative results using the metric mentioned in Section 5.1.2 for detailed comparison. We denote the various methods according to their abbreviations, indicate the image size as -128 or -256, and use "BBB" to refer to the use of Bayes by Backprop (Blundell et al. 2015) to approximate the posterior distribution over the parameters. Here, we show difference maps to represent the reconstruction error X_{dif}. Due to a known domain gap, models are trained respectively for the two available modalities, T1-weighted and T2-weighted on healthy brain images in CamCAN datasets. Trained models are later tested on its corresponding modality on BraTS-T2w and ATLAS-T1w. Thus in this work, we consider two independent detection tasks, 1) detection of tumours on BraTS-T2w, 2) detection of lesions on ATLAS-T1w. All models are trained until convergence.To calculate the metric of mDSC, we iterate the conventional dice score calculation through an arbitrary range of possible thresholds on the reconstruction error with $t \in [0.0, 6.0]$ with 1001 intervals. As GMM models output normalised probability maps with values between 0.0 and 1.0, the range is changed to $t \in [0.0, 1.0]$ with 400 intervals. Note that, mDSC is more of a numerical approximation of the maximal dice score through brute force search. The approximation gets better with more intervals whereas optimization may be needed to efficiently approximate the optimal value.

For the convolution version of VAE and AAE, latent variables are obtained from the previous convolutional layer instead of a dense layer to avoid possible loss of spatial information. Following the theory of variational autoencoders, we still assume each latent variable is independent and

compute the KL divergence between $q(z_H|X_H)$ and $\mathcal{N}(0,1)$.

As autoencoder-based methods usually have difficulties reconstructing large images, such as images of size $256\text{px} \times 256\text{px}$, we reduce the challenge by training our models also on downsampled datasets. The downsampled datasets are obtained for both, training and test datasets, by resizing the original images to the size of $128\text{px} \times 128\text{px}$ as described in Section 5.1.3.1. However, experimental results achieved by training and testing on downsampled datasets are not significantly better, but rather similar to the results on the original datasets. Although downsampling has little advantage in terms of outlier detection performance, it has a larger impact on training as it requires shorter runtimes and less GPU memory compared to the original datasets.

As the tumour shows relatively high intensity on T2-weighted images, the resulting intensity differences between the healthy and abnormal images can be more obvious than the intensity differences of lesions viewed in T1-weighted images. Additionally, the size of tumour in the BraTS dataset is often larger than that of the lesions in the ATLAS dataset. This property may, to some extent, facilitate the tumour detection in the BraTS-T2w setup. When comparing the difference maps obtained on BraTS-T2w and ATLAS-T1w, we can confirm this statement. Figure 5.2 shows that the tumour can often be fully or partially highlighted by the autoencoding methods while the detection appears worse on ATLAS-T1w, indicating a worse detection outcome on BraTS-T2w.

Table 5.1 shows the quantitative results in terms of AUC and mDSC for all methods on the BraTS-T2w and ATLAS-T1w datasets. Additionally, Section 5.1.4 shows the ROC curves corresponding to the results presented in Table 5.1. Our baseline methods appear to achieve strong performances. GMM produces the highest AUC on both datasets, although its mDSC on BraTS-T2w appears to expose some limitations. The superior performance of the GMM model might be cause by the fact that it is not entirely unsupervised as the number of components are predefined based on anatomical knowledge. The tested U-Net, as a widely used supervised method, achieves the highest dice score on both datasets. The various autoencoder-based models show similar performances on the BraTS-T2w dataset, and similarly achieve comparable performances on the ATLAS-T1w dataset. The results are consistent with the difference maps as shown in Fig. 5.2. The detection of tumours on the BraTS-T2w dataset indicates that autoencoder-based models are capable of detecting the large-size abnormalities, although the separation, as measured in mDSC, can be further improved. In contrast, the results on the ATLAS-T1w dataset imply difficulties of detection lesions for unsupervised and supervised methods. In terms of autoencoder-based models, although none of them has a significant advantage over the rest on both tasks – adversarial autoencoders (AAE), variational autoencoders (VAE) and the fully Bayesian VAE (VAE-BBB) are the most effective ones achieving higher performance metrics than the others.

Table 5.1: Summary of the different evaluation metrics, AUC and mDSC, for the various tested methods. Specifically, we train the denoising autoencoder (DAE) with Gaussian noise $\mathcal{N}(\mu = 0, \sigma = 0.5)$. Methods with "-128" and "-256" refer to variational autoencoders trained on datasets with images of size $128\text{px} \times 128\text{px}$ and $256\text{px} \times 256\text{px}$, respectively. We test the GMM baseline owith two parameter settings, λ_{out}=0.01 and λ_{out}=0.001. The dimensionality of the latent variables is shown as tensors for convolutions autoencoder models.

	Latent variables	BraTS-T2w (whole tumour)		ATLAS-T1w	
Models	z	AUC	mDSC	AUC	mDSC
mean	-	0.65	0.20	0.46	0.02
AE	256	0.63	0.41	0.49	0.03
DAE (σ=0.5)	256	0.59	0.29	0.41	0.06
VAE-128	(2,2,64)	0.69	0.42	0.64	0.08
VAE-BBB-128	(2,2,64)	0.69	0.40	0.67	0.05
VAE-256	(4,4,64)	0.67	0.40	0.66	0.08
AAE-128	(2,2,64)	0.70	0.41	0.63	0.06
AAE-256	(4,4,64)	0.67	0.38	0.60	0.04
α-GAN-128	128	0.66	0.35	0.60	0.05
α-GAN-256	256	0.67	0.37	0.60	0.04
GMM (λ_{out}=0.01)	-	0.80	0.22	0.78	0.17
GMM (λ_{out}=0.001)	-	0.79	0.21	0.77	0.17
U-Net (supervised)	-	-	0.80	-	0.50

5.1.4.1 BraTS-T2w

Comparing the various fully unsupervised methods, the convolutional VAE and the VAE-BBB trained on the downsampled $128\text{px} \times 128\text{px}$ images yield the highest value, in terms of AUC score. Interestingly, the convolutional VAE outperforms the VAE-BBB on mDSC marginally by 2%, achieving the highest score in this metric. The second highest mDSC is achieved by the convolutional AAE trained on the downsampled dataset with a value of 0.41, which is 1% lower than the convolution VAE. The denoising autoencoder has significantly inferior performance compared to the other autoencoder models. Although the α-GAN theoretically produces realistic and sharp images, the models trained on downsampled and original datasets do not yield advantages in this detection task: the α-GAN models have similar – but slightly lower – AUC scores than the best performing models, while its maximum dice score performance is in the mid-range compared to the other methods.

5.1.4.2 ATLAS-T1w

The autoencoder-based models achieved worse performance on the ATLAS-T1w dataset than on the BraTS-T2w dataset. Although the AUC metric does not significantly decrease, the maximum dice score reveals the weak performance of these models when detecting lesions in T1-weighted

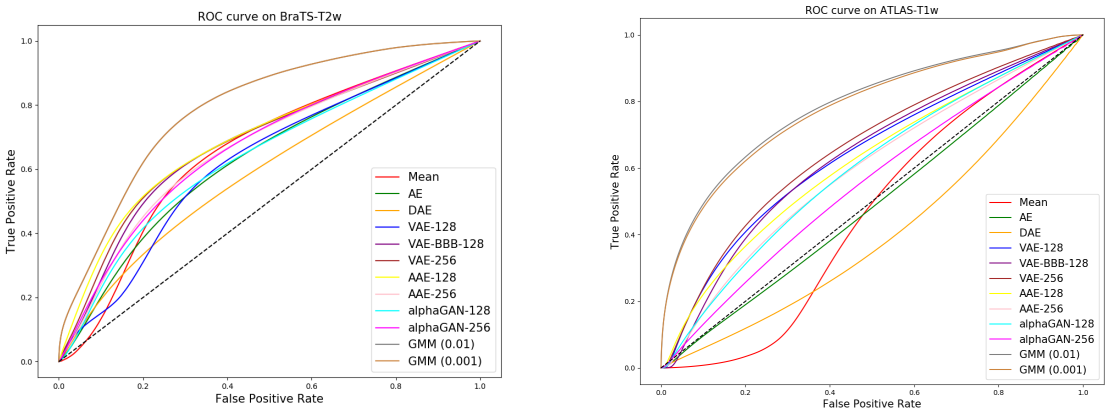


Figure 5.3: ROC curves corresponding to reported AUC in Table 5.1. The plot on the left shows ROC curves for models trained on the BraTS-T2w dataset, while the left plot shows ROC curves for models models trained on the ATLAS-T1w dataset. Black dashed lines mark the performance corresponding to random classification. Effective models should yield curves above this dashed line. Note that GMM(0.01) and GMM(0.001) have almost indistinguishable performance on the BraTS-T2w dataset which causes their curves to overlap.

images. None of the models achieved comparable results to the performances obtained on the detection task on T2-weighted images. Every model achieves mDSC scores of below 0.1, indicating that the lesions cannot sufficiently be distinguished from the normal structures by any threshold we have applied to X_{dif}. In comparison, the dice score of the supervised U-Net and that of the GMM are significantly higher than that of the autoencoder based methods. In spite of this large performance gap, supervised segmentation with U-Net achieves a dice score of only 0.50, which is relatively low for supervised segmentation. Given the results on the ALTAS-T1w dataset, we conclude that the unsupervised detection of lesions on the ATLAS-T1w dataset remains a difficult task, where even supervised segmentation has difficulties.

5.1.5 Discussion

In this work, we provided an overview of the current state of unsupervised outlier detection on brain MR image, which are fairly standard in medical image analysis. We evaluated autoencoder-based unsupervised models in terms of the area under the receiver operation characteristics (AUC) and maximum achievable dice score (mDSC) to describe their strength on this new application. Our results indicate that convolutional VAEs, Bayesian VAEs and AAEs have great potential to be further studied and developed to gain higher detection accuracy. We also identify that detection of lesions or tumours on T2-weighted datasets may be an easier first step to explore, while detection tasks on T1-weighted datasets remain more challenging. Moreover, it is easily noticed that the performance achieved by current available autoencoding models is worse than popular supervised methods, such as U-Nets. We suggest some possible directions to bring improvements.

Improvement in reconstruction quality: As the detection is based on absolute reconstruction error, it is straight-forward that higher accuracy can be achieved if the model is able to obtain sharp and accurate reconstructions. One of many approaches to achieve this is by combining VAEs and GANs to produce sharp images as several works have suggested (Larsen et al. 2016).

Estimation of pixel-wise probability: In our approach, the pixel-wise probability is approximated by calculating the reconstruction error. As shown in the difference maps, the models manage to reconstruct an image with abnormalities as a healthy-looking image, which is in line with our expectations. Although the reconstructions appears to be healthy-looking, taking absolute intensity differences might be too constrained, because it ignores structural differences. Assume there exists a pixel X_A^a which is abnormal and the pixel X_A^h which is a high-intensity normal pixel. Even if $X_A'^a$ is reconstructed into a normal pixel, it can occur that $X_A^a - X_A'^a \leq X_A^h - X_A'^h$. This behaviour largely lowers the performance even if the reconstruction is of good quality, when such cases are prevalent in the dataset. This is to say, that proper pixel-wise probability estimation with structural awareness can be helpful to improve performance.

Thresholding: Another question lies in selection of a threshold. In this work, we leave the selection of thresholds open and instead evaluate the models within a range of potential thresholds. One may argue that Dice scores can be calculated using a statistically chosen threshold, such as using the 90% percentile of the reconstruction errors as a threshold. Thresholding according to a given percentile can be valid, whereas the percentile may not be optimal for the data. The use of a proper and adaptive threshold can also help to distinguish outliers from normal structure.

5.2 Detecting Outliers in CT images

Deep learning is arguably now one of the most widely used machine learning methods for medical imaging (Litjens et al. 2017). A common task is the segmentation of lesions and other pathologies. However, most methods are based on *supervised* learning, which means they require large amounts of carefully annotated training data. Our work here focuses on *unsupervised* lesion detection that resembles pixel-wise outlier detection related to (An and Cho 2015; Carrera et al. 2015; Schlegl et al. 2017; Van Leemput et al. 2001). Most of those methods build on generative models that capture the normal distribution and detect outliers by checking their likelihood.

We introduce the use of Bayesian autoencoders to model the data distribution and interpret the reconstruction error as a measure of abnormality. Applying this method to CT mid-axial slices we show that our approach achieves superior performance to various baselines.

5.2.1 Using Autoencoders to find Anomalous Regions

We are interested in building an autoencoder AE for healthy data points $x \in \mathcal{D}_{healthy}$ so that the probability of the data point given the autoencoder reconstruction $\mathcal{N}(x|AE(x), 1)$ is maximised.

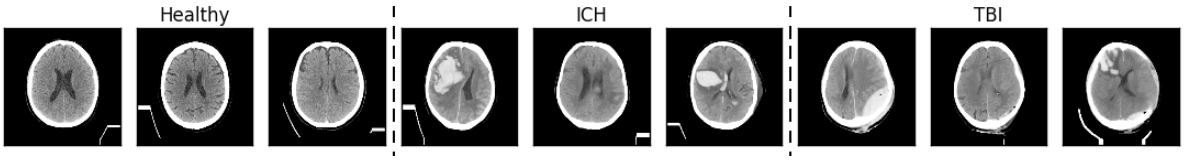


Figure 5.4: Examples of mid-axial slices of the data used. The healthy data is used for learning a normative model to detect lesions in ICH and TBI.

Rather than only learning point estimates of the weights w of the autoencoder we use dropout (Srivastava et al. 2014) and an uninformative prior to model the weight uncertainty of the autoencoder with MC-Dropout (Gal and Ghahramani 2015). This allows us to build a Bayesian autoencoder and we estimate $AE(x)$ as the Monte-Carlo (MC) estimate:

$$AE(x) = \int AE(x|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta \approx \frac{1}{N}\sum_{i=1}^{N} AE(x|\theta_i)\,,\ \theta_i \sim p(\theta|\mathcal{D}). \tag{5.4}$$

In general, autoencoders learn to compress the data and thus find a lower dimensional manifold that the training data lies on. An optimal autoencoder would therefore have a zero reconstruction error $\delta_{rec}(x) = |x - AE(x)| = 0$ for any $x \in \mathcal{D}_{training}$. For data samples different from the training manifold $x' \notin \mathcal{D}_{training}$ we argue that the autoencoder generates a reconstruction that projects the data towards the manifold, because the lower-dimensional latent space forms a bottleneck that prevents the autoencoder from learning an identity mapping. Therefore, we interpret $\delta_{rec}(x)$ as a distance to the found manifold that is related to the probability of the new sample being part of the same manifold $p(x' \in \mathcal{D}|\mathcal{D}) \propto \delta_{rec}(x')$. Here, $p(x' \in \mathcal{D}|\mathcal{D})$ describes the inverse of the probability of x' being an outlier. We use this estimate $p(x' \notin \mathcal{D}|\mathcal{D}) \propto |x - AE(x)|$ to find localised outliers by thresholding.

5.2.2 Experiments & Results

As a proof of concept, we test our Bayesian autoencoder on mid-axial slices of registered CT images. The 3D images are registered with an affine transformation to a CT atlas in canonical MNI space. We use 102 healthy cases to train the autoencoder and 107 cases with intracranial haemorrhages (ICH) and 98 cases with traumatic brain injuries (TBI) to test. We only evaluate the performance of outlier detection within a brain mask that was derived from the atlas used for pre-registration. We set the pixel intensities outside of that mask to -20, clip the intensities to the HU range of $[-20, 300]$ and rescale to $[-1, 1]$. Examples of the raw mid-axial slices before masking are shown in Fig. 5.4. The blood lesions are clearly visible as bright spots in the images, whereas oedema are not trivial to identify due to the low contrast differences compared to normal tissue.

We train convolutional autoencoders based on the implementation in DLTK (Pawlowski et al. 2017c)

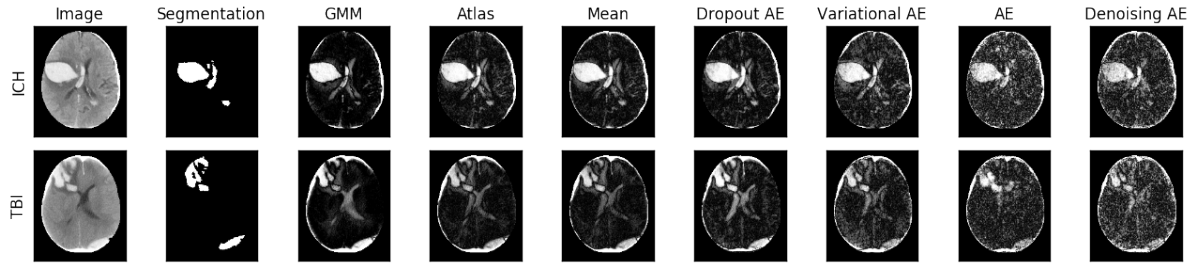
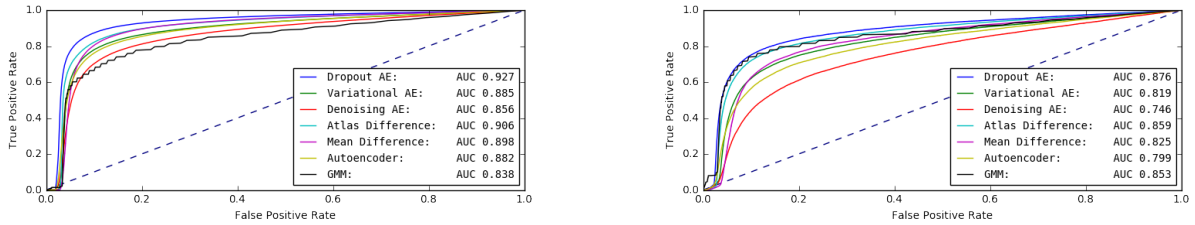


Figure 5.5: Comparison of the difference maps generated by the different methods. Brighter spots correspond to a higher difference.



(a) ICH **(b)** TBI

Figure 5.6: ROC curves for the segmentation of blood using thresholding of difference maps.

using Adam (Kingma and Ba 2015).We compare our approach to regular autoencoders, variational autoencoders (Kingma and Welling 2014) and denoising autoencoders as well as simple baselines such as the difference to the mean training image and the difference to the CT atlas. Further, we trained a Gaussian mixture model (GMM) on the intensities within the brain mask and used the fitted model to score the probability of unseen pixel values within the brain mask. For the denoising, variational and Bayesian autoencoders we use 100 MC estimates. Here, the denoising autoencoder requires a MC estimate as we also apply additional noise during testing.

Figure 5.5 shows an overview of the difference maps generated by the different methods for an example from ICH (first row) and TBI (second row). All methods have issues with imperfect skull stripping as the intensity differences are quite high. The GMM provides the least noisy difference maps, however fails to capture some lesion regions. The atlas and mean image based differences achieve good performances, but have difficulties dealing with structural information. The autoencoder-based methods should be able to capture the structural information better. However, the regular and denoising autoencoder exhibit noisy reconstruction errors. The dropout and variational autoencoder smooth this noise as they combine multiple samples from the weights or latent code.

We show receiver operation characteristics (ROC) as well as their area under the ROC curve (AUC) for quantitative results in Figure 5.6. Those curves show the true and false positive rates across all pixels given the values from the concatenated masked difference maps. The MC-Dropout-based Bayesian autoencoder achieves the best performance on the task of detecting blood in the masked CT mid-axial slice.

5.2.3 Discussion & Conclusion

We show that autoencoders are able to perform outlier detection because they fit a lower dimensional manifold for data compression and project unseen data onto this manifold. Therefore, outliers will be lost and are visible as reconstruction error. The Bayesian variant of this approach enables the autoencoder to smooth out uncertainties in the weight space and outperforms the baselines.

GANs (Goodfellow et al. 2014) are shown to be able to perform similar tasks (Schlegl et al. 2017) but have training instabilities. Because of this, we were not able to train a GAN with sufficient fidelity to the data at hand. Future work should evaluate the method on full 3D volumes and use improved approximations for Bayesian deep learning (Louizos and Welling 2017; Pawlowski et al. 2017a). Lastly, better tuned GAN training should be used for another baseline that might fit the manifold better.

5.3 Detecting Outliers on Histopathology Images

According to the World Health Organization, cancer is one of the leading causes of mortality worldwide. The diagnosis of cancer relies on the examination of tissue samples by expert pathologists which is a difficult and time-consuming task (World Health Organization 2018). Recent advances in machine learning promise to decrease the time necessary to obtain an accurate diagnosis (Komura and Ishikawa 2018). Current state-of-the-art methods in machine learning for histopathology employ deep learning which requires large annotated datasets for model training. Further, most methods rely on pixel-wise annotations of the whole-slide images (WSI) and work with patches extracted from the WSI (Bejnordi et al. 2017). Methods that work with image-level labels only have to overcome challenges which arise from large image size, as well as the low ratio of objects of interest (cancerous cells) to background in those images (Katharopoulos and Fleuret 2019; Pawlowski et al. 2019). We aim to reframe this task as an out-of-distribution (OOD) detection task that detects pathologies as outliers under a statistical model of healthy data (Chen et al. 2018b). We show that recent deep learning-based density estimation methods achieve competitive performance to fully supervised methods.

5.3.1 Background & Method

Recent work on normalising flows (Dinh et al. 2017; Papamakarios et al. 2017) allows for density estimation on high dimensional image data. Normalising flows model a complex probability density $p(x)$ using a bijective transformation f of a base distribution $\pi(u)$ as $x = f(u) \mid u \sim \pi(u)$. The base distribution π can be chosen at will, allowing for the choice of simple distributions such as

the Gaussian distribution. Because $f(\cdot)$ is invertible, the density $p(x)$ can be calculated as

$$p_\phi(x) = \pi\big(f_\phi^{-1}(x)\big)\left|\det \frac{\partial f_\phi^{-1}}{\partial x}\right|, \tag{5.5}$$

using the change of variable formula. Maximum likelihood estimation can then be used to learn the parametrised transformation $f_\phi(\cdot)$, where ϕ represents the parameters of the transformation.

However, it has been shown that the estimated likelihood is not guaranteed to be a reliable estimate for detecting OOD samples (Kirichenko et al. 2020; Le Lan and Dinh 2020; Nalisnick et al. 2019b) and various other OOD scoring metrics have been proposed (Choi et al. 2018; Nalisnick et al. 2019a; Ren et al. 2019). However, these methods either require to train multiple density estimation models (Choi et al. 2018; Ren et al. 2019) or can only handle batch-wise OOD detection (Nalisnick et al. 2019a). Instead we propose to cut down compute requirements during training by interpreting different points along the training trajectory as different models, similar to (Huang et al. 2017a; Maddox et al. 2019; Pawlowski et al. 2017b).

Given multiple density estimators $p_{\phi_1}, \dots, p_{\phi_n}$, we consider the following OOD scores:

- The log-likelihood: $\log p_{\phi_i}$
- The expected log-likelihood: $\mathbb{E}_i[\log p_{\phi_i}]$
- The Watanabe-Akaike Information Criterion (WAIC) (Choi et al. 2018): $\mathbb{E}_i[\log p_{\phi_i}(x)] - \mathrm{Var}_i[\log p_{\phi_i}(x)]$
- A variation on the typicality test from (Nalisnick et al. 2019a): $|\mathbb{E}_i[-\log p_{\phi_i}(x) - \mathbb{E}_{x' \sim X_{train}}[-\log p_{\phi_i}(x')]]|$
- The variance of the log-likelihood $\mathrm{Var}_i[\log p_{\phi_i}]$

Note that, different to the other scores, we expect the variance of inliers to be higher than that of outliers as we expect training of the models to mainly impact the behaviour for inlier samples, whereas the likelihood of outlier samples will mainly depend on the inductive biases of the model (Kirichenko et al. 2020).

5.3.2 Experiments & Discussion

We use the PatchCamelyon (PCam) dataset (Veeling et al. 2018) to test our concept of using normalising flows for OOD detection on histopathology images. PCam consists of $327,680$ patches extracted from the CAMELYON16 dataset (Bejnordi et al. 2017). Each 96×96 px patch is labelled as positive or negative to indicate whether its center 32×32 px patch contains cancerous cells or not. We use the original train, validation, and test splits. We train our density estimator on all negative examples from the training set. We then calculate the area under the ROC curve (AUROC)

to estimate the separability and classification performance of positive and negative patches. We train Residual Flows (Chen et al. 2019a) using the original code[4] as a density estimator on the 32×32 px centre patches for 60 epochs and use the checkpoints at epochs 52-60 as the different density estimators. We compare our proposed method to a statistical baseline as well as a fully supervised learning method. The statistical baseline estimates the probability of an inlier as $p(x) = \mathcal{N}(x[:,1] \mid \mu_1, \sigma_1)\mathcal{N}(x[:,2] \mid \mu_2, \sigma_2)\mathcal{N}(x[:,3] \mid \mu_3, \sigma_3)$, where $x[:,i]$ denotes the ith colour channel of the patch x and μ_i, σ_i the corresponding empirical mean and variance.

Table 5.2: Comparison of AUROCs for the task of correctly classifying patches from the PCam test set. The single log-likelihood result is computed using the last model checkpoint. Typ. refers to our variation on the typicality test introduced by (Nalisnick et al. 2019a). GDensenet refers to the official supervised PatchCAM baseline (Veeling et al. 2018).

Method	$\log p_\phi$	$\mathbb{E}_i[\log p_{\phi_i}]$	$\mathrm{Var}_i[\log p_{\phi_i}]$	WAIC	Typ.	Gaussian	GDensenet
AUROC [%]	53.4	81.6	92.4	25.3	61.8	31.8	96.3

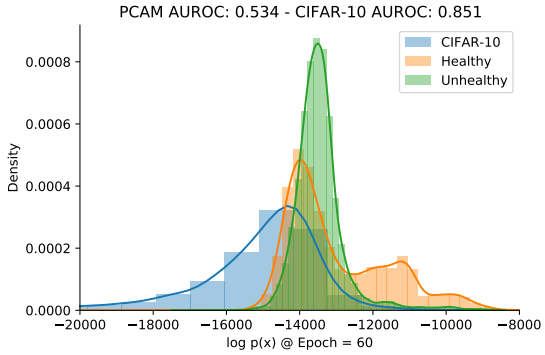
Table 5.2 summarizes the results on the PCam dataset. Consistent with previous work we find that density estimation alone is not a reliable OOD detection metric, as seen with the performance of the Gaussian estimator and the regular log-likelihood. However, more sophisticated OOD scoring metrics achieve superior performance. Specifically, using the variance of the log-likelihood achieves an AUROC of 92.4%, being competitive compared to fully supervised methods such as GDensenet. We investigate the distribution of the different outlier detection metrics on the test set[5] of healthy and unhealthy PatchCamelyon images in Fig. 5.7. We also include the distribution of images from the CIFAR10 dataset to study the behaviour of the models on far-OOD detection. The plots confirm that the regular likelihood is unable to separate healthy and unhealthy images. More complex outlier metrics can improve the separation of the different datasets. However, only the variance metric clearly separates all datasets from each other. We hypothesise that CIFAR-10 samples are assigned similar likelihoods as the training samples due to inductive biases of the used density estimation methods as found in Kirichenko et al. (2020). Summarising, we have shown that there is evidence that deep OOD detection methods are capable of identifying cancerous histopathology images without the need of annotated cancerous training data. We argue that deep density estimation with normalising flows should be further explored as it may have a significant impact on the throughput of pathological analysis avoiding the need for costly pixel-level annotations of cancerous cells.

The current work is limited as it lacks thorough tuning of the Residual Flow and relies on the PatchCamelyon dataset which is derived from WSI that all contain regions with lesions. Future evaluations will therefore look into training on crops from the CAMELYON17 dataset and examine the

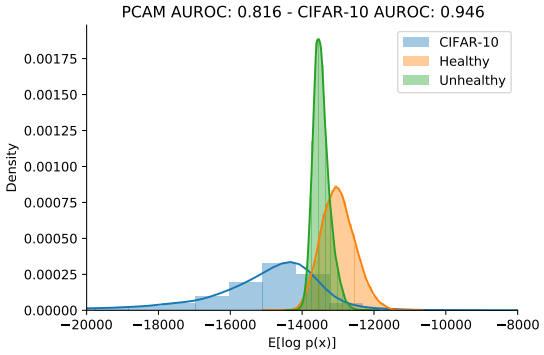
[4] See https://github.com/rtqichen/residual-flows for the original code.
[5] We show the same plots for the validation set in Fig. B.7.

performance of methods on whole-slide histopathology images to showcase their real-world applicability. Furthermore, it currently is not clear whether the suggested OOD scoring metric of the likelihood variance during training generalises to other problem domains or is specific to this particular dataset. Initial experiments on synthetic data as well as more common computer vision datasets [6] suggest that this point requires more investigation as the computer vision experiments showed little separation using this metric. Nevertheless, we believe that observing the model behaviour over the course of training warrants future research into new ways of constructing OOD metrics. Lastly, we believe that future studies should examine further the low positive data regime for supervised and semi-supervised methods as they could provide better value per annotation time than fully unsupervised ones.

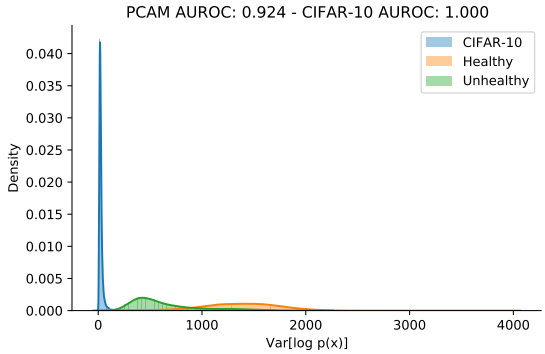
[6] We used CIFAR-10 as inlier and SVHN as OOD data.



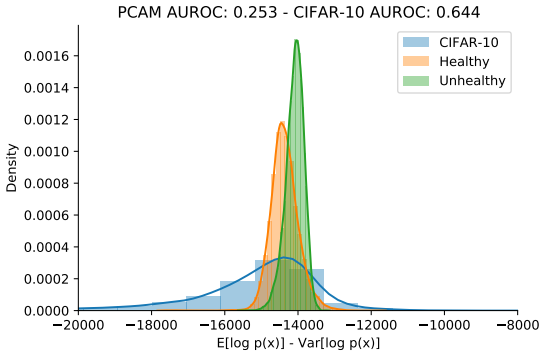
(a) Using the regular log likelihood, $\log p(x)$, for OOD detection.



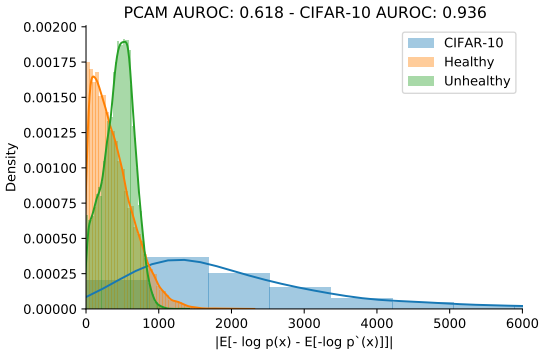
(b) Using the expected log likelihood, $\mathbb{E}[\log p(x)]$, for OOD detection.



(c) Using the variance of log likelihood, $\mathrm{Var}[\log p(x)]$, for OOD detection.



(d) Using the WAIC, $\mathbb{E}[\log p(x)] - \mathrm{Var}[\log p(x)]$, for OOD detection.



(e) Using the expected typicality, $|\mathbb{E}_{epochs}[-\log p(x)] - \mathbb{E}_x[-\log p_{train}(x)]|$, for OOD detection.

Figure 5.7: Comparison of the distribution of the different outlier metrics on the test set of healthy and unhealthy PatchCamelyon images as well as on CIFAR10. Note that the likelihood-only metrics do not separate the different dataset. More complicated outlier metrics are capable of showing some separation. However, only the variance metric fully separates all three datasets.

Chapter 6

Modelling Causal Relationships with Deep Learning

This chapter is based on the following publication:

(a) **Pawlowski, N.***, Castro, D. C.*, and Glocker, B. (2020). "Deep Structural Causal Models for Tractable Counterfactual Inference". In: *Advances in Neural Information Processing Systems* – (Pawlowski et al. 2020)

This work was completed with shared first-authorship with equal contribution from Nick Pawlowski and Daniel C. Castro. Both authors contributed to all aspects of the work with main responsibilities as follows: N.P. led the implementation of specific models, data generation and processing, and experimentation; D.C.C. had main responsibilities on the framework design, analysis, and background research. Both took equal responsibility on ideation, conceptualisation, framework implementation, and writing of the manuscript. N.P. worked on Section 6.2.5 and Section 6.6 independently after the original work was published (Pawlowski et al. 2020).

The code for all experiments and interactive demonstrations are available at https://github.com/biomedia-mira/deepscm and some extensions can be found at https://github.com/pawni/deepscm/tree/correlation.

The previous chapters have focussed on modelling the density of observed samples (Chapter 5) or the correlations between various random variables – the probability of an image belonging to a certain class, either as a measure of the certainty of its predictions based on the probability of the model (the weights) given the training dataset (Chapter 4), or simply for its class prediction

given the patches of the image (Chapter 3). This chapter aims to move beyond the modelling of (conditional) probabilities and studies the concept of causality at all of its levels, including questions of counterfactual nature like: *What would this brain image look like if the subject had a bigger brain?*[1]

The ability to answer counterfactual questions requires assumptions about the causal structure of the modelled variables as well as their mechanistic relationships. Here, we propose a framework that uses deep learning components to allow the training of deep structural causal models (deep SCMs or DSCMs). Specifically, deep SCMs use normalising flows and variational inference to allow for tractable inference of exogenous variables – the first out of three steps in the process of calculating counterfactuals in the SCM framework. We verify the capabilities of this framework on three case studies. In the first case study, we use a synthetic dataset with known causal structure based on Morpho-MNIST (Castro et al. 2019) and show that our model can perform all three rungs of Pearl's ladder of causation (Pearl 2019) by investigating the observational, interventional and counterfactual distributions. The second case study models a more complex dataset of brain MR images and various covariates. We use this example to show how the framework can be used to gain insights in real world datasets once a causal structure is assumed. Lastly, we use another synthetic Morpho-MNIST dataset to assess the limits of the deep SCM framework. Our set of experiments illustrates that our framework is capable of answering complicated causal queries in various application domains.

6.1 Introduction

Many questions in everyday life as well as in scientific inquiry are causal in nature: "How would the climate have changed if we'd had less emissions in the '80s?", "How fast could I run if I hadn't been smoking?", or "Will my headache be gone if I take that pill?". None of those questions can be answered with statistical tools alone, but require methods from causality to analyse interactions with our environment (interventions) and hypothetical alternate worlds (counterfactuals), going beyond joint, marginal, and conditional probabilities (Peters et al. 2017). Even though these are natural lines of reasoning, their mathematical formalisation under a unified theory is relatively recent (Pearl 2009).

In some statistics-based research fields, such as econometrics or epidemiology, the use of causal inference methods has been established for some time (Greenland et al. 1999; Wold 1954). However, causal approaches have been introduced into deep learning (DL) only very recently (Schölkopf 2019). For example, research has studied the use of causality for disentanglement (Parascandolo et al. 2018; Yang et al. 2020), causal discovery (Bengio et al. 2020; Goudet et al. 2018), and for deriving causality-inspired explanations (Martinez and Marca 2019; Singla et al. 2020) or data aug-

[1] Alternatively an interesting counterfactual question this year could be *"Who would have won the European Football Championship in 2021 if it would not have been delayed due to COVID-19?"*.

mentations (Kaushik et al. 2020). Causal DL models could be capable of learning relationships from complex high-dimensional data and of providing answers to interventional and counterfactual questions, although current work on deep counterfactuals is limited by modelling only direct cause-effect relationships (Singla et al. 2020) or instrumental-variable scenarios (Hartford et al. 2017), or by not providing a full recipe for tractable counterfactual inference (Kocaoglu et al. 2018).

The integration of causality into DL research promises to enable novel scientific advances as well as to tackle known shortcomings of DL methods: DL is known to be susceptible to learning spurious correlations and amplifying biases (e.g. Zhao et al. 2017), and to be exceptionally vulnerable to changes in the input distribution (Szegedy et al. 2014). By explicitly modelling causal relationships and acknowledging the difference between causation and correlation, causality becomes a natural field of study for improving the transparency, fairness, and robustness of DL-based systems (Kusner et al. 2017; Subbaswamy et al. 2019). Further, the tractable inference of deep counterfactuals enables novel research avenues that aim to study causal reasoning on a per-instance rather than population level, which could lead to advances in personalised medicine as well as in decision-support systems, more generally.

In this context, our work studies the use of DL-based causal mechanisms and establishes effective ways of performing counterfactual inference with fully specified causal models with no unobserved confounding. Our main contributions are: 1) a unified framework for structural causal models using modular deep mechanisms; 2) an efficient approach to estimating counterfactuals by inferring exogenous noise via variational inference or normalising flows; 3) case studies exemplifying how to apply deep structural causal models and perform counterfactual inference. The paper is organised as follows: we first review structural causal models and discuss how to leverage deep mechanisms and enable tractable counterfactual inference. Second, we compare our work to recent progress in deep causal learning in light of Pearl's ladder of causation (Pearl 2019). Finally, we apply deep structural causal models to synthetic experiments as well as to modelling brain MRI scans, demonstrating the practical utility of our framework in answering counterfactual questions.

6.2 Deep Structural Causal Models

We consider the problem of modelling a collection of K random variables $\mathbf{x} = (x_1, \dots, x_K)$. By considering causal relationships between them, we aim to build a model that not only is capable of generating convincing novel samples, but also satisfies all three rungs of the causation ladder (Pearl 2019). The first level, **association**, describes reasoning about passively observed data. This level deals with correlations in the data and questions of the type *"What are the odds that I observe…?"*, which relates purely to marginal, joint, and conditional probabilities. **Intervention** concerns interactions with the environment. It requires knowledge beyond just observations, as it relies on structural assumptions about the underlying data-generating process. Characteristic questions ask about the effects of certain actions: *"What happens if I do…?"*. Lastly, **counterfac-**

tuals deal with retrospective hypothetical scenarios. Counterfactual queries leverage functional models of the generative processes to imagine alternative outcomes for individual data points, answering *"What if I had done A instead of B?"*. Arguably, such questions are at the heart of scientific reasoning (and beyond), yet are less well-studied in the field of machine learning. The three levels of causation can be operationalised by employing structural causal models (SCMs)[2], recapitulated in the next section.

6.2.1 Background on structural causal models

A structural causal model $\mathfrak{G} := (\mathbf{S}, P(\boldsymbol{\epsilon}))$ consists of a collection $\mathbf{S} = (f_1, \ldots, f_K)$ of structural assignments $x_k := f_k(\epsilon_k; \mathbf{pa}_k)$ (called *mechanisms*), where $\mathbf{pa}_k$ is the set of parents of x_k (its *direct causes*), and a joint distribution $P(\boldsymbol{\epsilon}) = \prod_{k=1}^{K} P(\epsilon_k)$ over mutually independent exogenous noise variables (i.e. unaccounted sources of variation). As assignments are assumed acyclic, relationships can be represented by a directed acyclic graph (DAG) with edges pointing from causes to effects, called the *causal graph* induced by $\mathfrak{G}$. Every SCM $\mathfrak{G}$ entails a unique joint observational distribution $P_{\mathfrak{G}}(\mathbf{x})$, satisfying the causal Markov assumption: each variable is independent of its non-effects given its direct causes. It therefore factorises as $P_{\mathfrak{G}}(\mathbf{x}) = \prod_{k=1}^{K} P_{\mathfrak{G}}(x_k \,|\, \mathbf{pa}_k)$, where each conditional distribution $P_{\mathfrak{G}}(x_k \,|\, \mathbf{pa}_k)$ is determined by the corresponding mechanism and noise distribution (Peters et al. 2017).

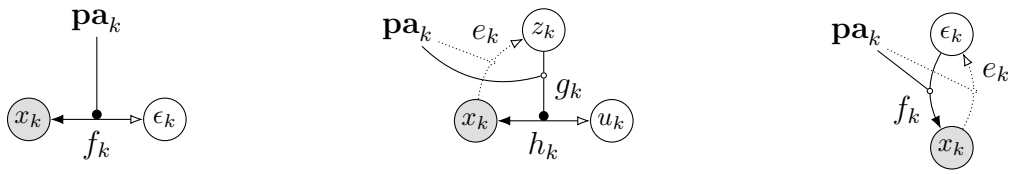
Crucially, unlike conventional Bayesian networks, the conditional factors above are imbued with a causal interpretation. This enables $\mathfrak{G}$ to be used to predict the effects of *interventions*, defined as substituting one or multiple of its structural assignments, written as '$\mathrm{do}(\cdots)$'. In particular, a constant reassignment of the form $\mathrm{do}(x_k := a)$ is called an atomic intervention, which disconnects x_k from all its parents and represents a direct manipulation disregarding its natural causes.

While the observational distribution relates to statistical associations and interventions can predict causal effects, SCMs further enable reasoning about *counterfactuals*. In contrast to interventions, which operate at the population level—providing aggregate statistics about the effects of actions (i.e. noise sampled from the prior, $P(\boldsymbol{\epsilon})$)—a counterfactual is a query at the unit level, where the structural assignments ('mechanisms') are changed but the exogenous noise is identical to that of the observed datum ($P(\boldsymbol{\epsilon} \,|\, \mathbf{x})$) (Pearl 2009; Peters et al. 2017).

These are hypothetical retrospective interventions (cf. potential outcome), given an observed outcome: 'What would x_i have been if x_j were different, given that we observed $\mathbf{x}$?'. This type of question effectively offers explanations of the data, since we can analyse the changes resulting from manipulating each variable. Counterfactual queries can be mathematically formulated as a three-step procedure (Pearl 2009, Ch. 7):

1. **Abduction:** Predict the 'state of the world' (the exogenous noise, $\boldsymbol{\epsilon}$) that is compatible with the observations, $\mathbf{x}$, i.e. infer $P_{\mathfrak{G}}(\boldsymbol{\epsilon} \,|\, \mathbf{x})$.

[2] SCMs are also known as (nonlinear) structural equation models or functional causal models.



(a) Invertible explicit likelihood **(b)** Amortised explicit likelihood **(c)** Amortised implicit likelihood

Figure 6.1: Classes of deep causal mechanisms considered in this work. Bi-directional arrows indicate invertible transformations, optionally conditioned on other inputs (edges ending in black circles). Black and white arrowheads refer resp. to the generative and abductive directions, while dotted arrows depict an amortised variational approximation. Here, f_k is the forward model, e_k is an encoder that amortises abduction in non-invertible mechanisms, g_k is a 'high-level' non-invertible branch (e.g. a probabilistic decoder), and h_k is a 'low-level' invertible mapping (e.g. reparametrisation).

2. **Action:** Perform an intervention (e.g. $\mathrm{do}(x_k := \widetilde{x}_k)$) corresponding to the desired manipulation, resulting in a modified SCM $\widetilde{\mathfrak{G}} = \mathfrak{G}_{\mathbf{x};\mathrm{do}(\widetilde{x}_k)} = (\widetilde{\mathbf{S}}, P_{\mathfrak{G}}(\boldsymbol{\epsilon}|\mathbf{x}))$ (Peters et al. 2017, Sec. 6.4).
3. **Prediction:** Compute the quantity of interest based on the distribution entailed by the counterfactual SCM, $P_{\widetilde{\mathfrak{G}}}(\mathbf{x})$.

With these operations in mind, the next section explores a few options for building flexible, expressive, and counterfactual-capable functional mechanisms for highly structured data.

6.2.2 Deep mechanisms

In statistical literature (e.g. epidemiology, econometrics, sociology), SCMs are typically employed with simple linear mechanisms (or generalised linear models, involving an output non-linearity). Analysts attach great importance to the regression weights, as under certain conditions these may be readily interpreted as estimates of the causal effects between variables. While this approach generally works well for scalar variables and can be useful for decision-making, it is not flexible enough to model higher-dimensional data such as images. Solutions to this limitation have been proposed by introducing deep-learning techniques into causal inference (Goudet et al. 2018; Kocaoglu et al. 2018).

We call an SCM that uses deep-learning components to model the structural assignments a *deep structural causal model* (DSCM). In DSCMs, the inference of counterfactual queries becomes more complex due to the potentially intractable abduction step (inferring the posterior noise distribution, as defined above). To overcome this, we propose to use recent advances in normalising flows and variational inference to model mechanisms for composable DSCMs that enable tractable counterfactual inference. While here we focus on continuous data, DSCMs also fully support discrete variables without the need for relaxations (see Section 6.2.4). We consider three types of mechanisms that differ mainly in their invertibility, illustrated in Fig. 6.1.

Invertible, explicit: Normalising flows model complex probability distributions using transformations from simpler base distributions with same dimensionality (Tabak and Turner 2013). For an observed variable x, diffeomorphic transformation f, and base variable $\epsilon \sim P(\epsilon)$ such that $x = f(\epsilon)$, the output density $p(x)$ can be computed as $p(x) = p(\epsilon)|\det \nabla f(\epsilon)|^{-1}$, evaluated at $\epsilon = f^{-1}(x)$ (Papamakarios et al. 2019; Rezende and Mohamed 2015). For judicious choices of f, the Jacobian ∇f may take special forms with efficiently computable determinant, providing a flexible and tractable probabilistic model whose parameters can be trained via exact maximum likelihood. Furthermore, flows can be made as expressive as needed by composing sequences of simple transformations. For more information on flow-based models, refer to the comprehensive survey by (Papamakarios et al. 2019). Note that this class of models also subsumes the typical location-scale and inverse cumulative distribution function transformations used in the reparametrisation trick (Kingma and Welling 2014; Rezende et al. 2014), as well as the Gumbel trick for discrete variable relaxations (Jang et al. 2017; Maddison et al. 2017).

Although normalising flows were originally proposed for unconditional distributions, they have been extended to conditional densities (Trippe and Turner 2017), including in high dimensions (Lu and Huang 2020; Winkler et al. 2019), by parametrising the transformation as $x = f(\epsilon; \mathbf{pa}_X)$, assumed invertible in the first argument. In particular, conditional flows can be adopted in DSCMs to represent invertible, explicit-likelihood mechanisms (Fig. 6.1a):

$$x_i \coloneqq f_i(\epsilon_i; \mathbf{pa}_i), \qquad p(x_i|\mathbf{pa}_i) = p(\epsilon_i) \cdot |\det \nabla_{\epsilon_i} f_i(\epsilon_i; \mathbf{pa}_i)|^{-1}\big|_{\epsilon_i = f_i^{-1}(x_i; \mathbf{pa}_i)} . \tag{6.1}$$

Amortised, explicit: Such invertible architectures typically come with heavy computational requirements when modelling high-dimensional observations, because all intermediate operations act in the space of the data. Instead, it is possible to use arbitrary functional forms for the structural assignments, at the cost of losing invertibility and tractable likelihoods $p(x_k|\mathbf{pa}_k)$. Here, we propose to separate the assignment f_k into a 'low-level', invertible component h_k and a 'high-level', non-invertible part g_k—with a corresponding noise decomposition $\epsilon_k = (u_k, z_k)$—such that

$$x_k \coloneqq f_k(\epsilon_k; \mathbf{pa}_k) = h_k(u_k; g_k(z_k; \mathbf{pa}_k), \mathbf{pa}_k), \qquad P(\epsilon_k) = P(u_k)P(z_k) . \tag{6.2}$$

In such a decomposition, the invertible transformation h_k can be made shallower, while the upstream non-invertible g_k maps from a lower-dimensional space and is expected to capture more of the high-level structure of the data. Indeed, a common implementation of this type of model for images would involve a probabilistic decoder, where g_k may be a convolutional neural network, predicting the parameters of a simple location-scale transformation performed by h_k (Kingma and Welling 2014).

As the conditional likelihood $p(x_k|\mathbf{pa}_k)$ in this class of models is no longer tractable because z_k cannot be marginalised out, it may alternatively be trained with amortised variational inference.

Specifically, we can introduce a variational distribution $Q(z_k \mid x_k, \mathbf{pa}_k)$ to formulate a lower bound on the true marginal conditional log-likelihood, which will be maximised instead:

$$\log p(x_k \mid \mathbf{pa}_k) \geq \mathbb{E}_{Q(z_k \mid x_k, \mathbf{pa}_k)}[\log p(x_k \mid z_k, \mathbf{pa}_k)] - D_{\mathrm{KL}}[Q(z_k \mid x_k, \mathbf{pa}_k) \| P(z_k)] \,. \tag{6.3}$$

The argument of the expectation in this lower bound can be calculated similarly to Eq. (6.1):

$$p(x_k \mid z_k, \mathbf{pa}_k) = p(u_k) \cdot |\det \nabla_{u_k} h_k(u_k; g_k(z_k, \mathbf{pa}_k), \mathbf{pa}_k)|^{-1} \big|_{u_k = h_k^{-1}(x_k; g_k(z_k, \mathbf{pa}_k), \mathbf{pa}_k)} \,. \tag{6.4}$$

The approximate posterior distribution $Q(z_k \mid x_k, \mathbf{pa}_k)$ can for example be realised by an encoder function, $e_k(x_k; \mathbf{pa}_k)$, that outputs the parameters of a simple distribution over z_k (Fig. 6.1b), as in the auto-encoding variational Bayes (AEVB) framework (Kingma and Welling 2014).

Amortised, implicit: While the models above rely on (approximate) maximum-likelihood as training objective, it is admissible to train a non-invertible mechanism as a conditional implicit-likelihood model (Fig. 6.1c), optimising an adversarial objective (Donahue et al. 2017; Dumoulin et al. 2017; Mirza and Osindero 2014). Specifically, a deterministic encoder e_j would strive to fool a discriminator function attempting to tell apart tuples of encoded real data $(x_j, e_j(x_j; \mathbf{pa}_j), \mathbf{pa}_j)$ and generated samples $(f_j(\epsilon_j; \mathbf{pa}_j), \epsilon_j, \mathbf{pa}_j)$. This class of mechanism is proposed here for completeness, without empirical evaluation. However, following initial dissemination of our work, (Dash and Sharma 2020) reproduced our Morpho-MNIST experiments (Section 6.4) and demonstrated these amortised implicit-likelihood mechanisms can achieve comparable performance.

6.2.3 Deep counterfactual inference

Now equipped with effective deep models for representing mechanisms in DSCMs, we discuss the inference procedure allowing us to compute answers to counterfactual questions.

Abduction: As presented in Section 6.2.1, the first step in computing counterfactuals is abduction, i.e. to predict the exogenous noise, $\boldsymbol{\epsilon}$, based on the available evidence, $\mathbf{x}$. Because each noise variable is assumed to affect only the respective observed variable, $(\epsilon_k)_{k=1}^K$ are conditionally independent given $\mathbf{x}$, therefore this posterior distribution factorises as $P_{\mathfrak{G}}(\boldsymbol{\epsilon} \mid \mathbf{x}) = \prod_{k=1}^K P_{\mathfrak{G}}(\epsilon_k \mid x_k, \mathbf{pa}_k)$. In other words, it suffices to infer the noise independently for each mechanism, given the observed values of the variable and of its parents[3].

For invertible mechanisms, the noise variable can be obtained deterministically and exactly by just inverting the mechanism: $\epsilon_i = f_i^{-1}(x_i; \mathbf{pa}_i)$. Similarly, implicit-likelihood mechanisms can be approximately inverted by using the trained encoder function: $\epsilon_j \approx e_j(x_j; \mathbf{pa}_j)$.

[3]Note that here we assume full observability, i.e. no variables are missing when predicting counterfactuals. We discuss challenges of handling partial evidence in Section 6.7.

Some care must be taken in the case of amortised, explicit-likelihood mechanisms, as the 'high-level' noise z_k and 'low-level' noise u_k are not independent given x_k. Recalling that this mechanism is trained along with a conditional probabilistic encoder, $Q(z_k|e_k(x_k;\mathbf{pa}_k))$, the noise posterior can be approximated as follows, where $\delta_w(\,\cdot\,)$ denotes the Dirac delta distribution centred at w:

$$\begin{aligned} P_{\mathfrak{G}}(\epsilon_k|x_k,\mathbf{pa}_k) &= P_{\mathfrak{G}}(z_k|x_k,\mathbf{pa}_k)\,P_{\mathfrak{G}}(u_k|z_k,x_k,\mathbf{pa}_k) \\ &\approx Q(z_k|e_k(x_k;\mathbf{pa}_k))\,\delta_{h_k^{-1}(x_k;g_k(z_k;\mathbf{pa}_k),\mathbf{pa}_k)}(u_k)\,. \end{aligned} \tag{6.5}$$

Action: The causal graph is then modified according to the desired hypothetical intervention(s), as in the general case (Section 6.2.1). For each intervened variable x_k, its structural assignment is replaced either by a constant, $x_k := \widetilde{x}_k$—making it independent of its former parents (direct causes, $\mathbf{pa}_k$) and of its exogenous noise (ϵ_k)—or by a surrogate mechanism $x_k := \widetilde{f}_k(\epsilon_k;\widetilde{\mathbf{pa}}_k)$, forming a set of counterfactual assignments, $\widetilde{\mathbf{S}}$. This then defines a counterfactual SCM $\widetilde{\mathfrak{G}} = (\widetilde{\mathbf{S}}, P_{\mathfrak{G}}(\epsilon|\mathbf{x}))$.

Prediction: Finally, we can sample from $\widetilde{\mathfrak{G}}$. Noise variables that were deterministically inverted (either exactly or approximately) can simply be plugged back into the respective forward mechanism to determine the new output value. Notice that this step is redundant for observed variables that are not descendants of the ones being intervened upon, as they will be unaffected by the changes.

As mentioned above, the posterior distribution over (z_k, u_k) for an amortised, explicit-likelihood mechanism does not factorise (Eq. (6.5)), and the resulting distribution over the counterfactual x_k cannot be characterised explicitly. However, sampling from it is straightforward, such that we can approximate the counterfactual distribution via Monte Carlo as follows, for each sample s:

$$\begin{aligned} z_k^{(s)} &\sim Q(z_k|e_k(x_k;\mathbf{pa}_k)) \\ u_k^{(s)} &= h_k^{-1}(x_k;g_k(z_k^{(s)};\mathbf{pa}_k),\mathbf{pa}_k) \\ \widetilde{x}_k^{(s)} &= \widetilde{h}_k(u_k^{(s)};\widetilde{g}_k(z_k^{(s)};\widetilde{\mathbf{pa}}_k),\widetilde{\mathbf{pa}}_k)\,. \end{aligned} \tag{6.6}$$

Consider an uncorrelated Gaussian decoder for images as a concrete example, predicting vectors of means and variances for each pixel of x_k: $g_k(z_k;\mathbf{pa}_k) = (\mu(z_k;\mathbf{pa}_k), \sigma^2(z_k;\mathbf{pa}_k))$, with the low-level reparametrisation given by $h_k(u_k;(\mu,\sigma^2),\mathbf{pa}_k) = \mu + \sigma^2 \odot u_k$. Exploiting the reparametrisation trick, counterfactuals that preserve x_k's mechanism can be computed simply as

$$u_k^{(s)} = (x_k - \mu(z_k^{(s)};\mathbf{pa}_k)) \oslash \sigma(z_k^{(s)};\mathbf{pa}_k), \qquad \widetilde{x}_k^{(s)} = \mu(z_k^{(s)};\widetilde{\mathbf{pa}}_k) + \sigma(z_k^{(s)};\widetilde{\mathbf{pa}}_k) \odot u_k^{(s)}\,,$$

where $\oslash$ and $\odot$ denote element-wise division and multiplication, respectively. In particular, in the

constant-variance setting adopted for our experiments, counterfactuals further simplify to

$$\widetilde{x}_k^{(s)} = x_k + [\mu(z_k^{(s)}; \widetilde{\mathbf{pa}}_k) - \mu(z_k^{(s)}; \mathbf{pa}_k)] \,.$$

This showcases how true image counterfactuals are able to retain pixel-level details. Typical conditional generative models would output only $\mu(z_k; \widetilde{\mathbf{pa}}_k)$ (which is often blurry in vanilla variational auto-encoders (Larsen et al. 2016)), or would in addition have to sample $P(u_k)$ (resulting in noisy images).

6.2.4 Discrete counterfactuals

The Deep Structural Causal Model framework supports not only low- and high-dimensional continuous data, but also discrete variables. In particular, discrete mechanisms with a Gumbel–max parametrisation have been shown to lead to counterfactuals satisfying desirable properties (Oberst and Sontag 2019). For example, they are invariant to category permutations and are stable, such that increasing the odds only of the observed outcome cannot produce a different counterfactual outcome. More computational details and properties of the Gumbel distribution are found in (Maddison and Tarlow 2017).

Consider a discrete random variable over K categories, y, with a conditional likelihood described by logits $\boldsymbol{\lambda}$, assumed to be a function g_Y of its parents, $\mathbf{pa}_Y$:

$$P(y = k \,|\, \mathbf{pa}_Y) = \frac{e^{\lambda_k}}{\sum_{l=1}^{K} e^{\lambda_l}}, \qquad \boldsymbol{\lambda} = g_Y(\mathbf{pa}_Y) \,. \tag{6.7}$$

Under the Gumbel–max parametrisation, the mechanism generating y can be described as

$$y \coloneqq f_Y(\boldsymbol{\epsilon}_Y; \mathbf{pa}_Y) = \underset{1 \le l \le K}{\arg\max} \, (\epsilon_Y^l + \lambda_l), \qquad \epsilon_Y^l \sim \mathrm{Gumbel}(0, 1) \,. \tag{6.8}$$

Samples from the $\mathrm{Gumbel}(0, 1)$ distribution can be generated by computing $-\log(-\log U)$, where $U \sim \mathrm{Unif}(0, 1)$.

The Gumbel distribution has certain special properties (Maddison and Tarlow 2017) that enable tractable abduction. Given that we observed $y = k$, samples can be generated from the exact posterior $P(\boldsymbol{\epsilon}_Y \,|\, y = k, \mathbf{pa}_Y)$:

$$\begin{aligned} \epsilon_Y^k &= G_k + \log \textstyle\sum_l e^{\lambda_l} - \lambda_k, & G_k &\sim \mathrm{Gumbel}(0, 1), \\ \epsilon_Y^l &= -\log(e^{-G_l - \lambda_l} + e^{-\epsilon_Y^k - \lambda_k}) - \lambda_l, & G_l &\sim \mathrm{Gumbel}(0, 1), \quad \forall l \neq k \,. \end{aligned} \tag{6.9}$$

Finally, given an upstream counterfactual intervention such that $\widetilde{\boldsymbol{\lambda}} = \widetilde{g}_Y(\widetilde{\mathbf{pa}}_Y)$, the counterfactual

outcome for y can be determined simply as

$$y = f_Y(\boldsymbol{\epsilon}_Y; \widetilde{\mathbf{pa}}_Y) = \underset{1 \leq l \leq K}{\arg\max} \, (\epsilon_Y^l + \widetilde{\lambda}_l) \, . \tag{6.10}$$

Note that this entire derivation applies to a truly discrete variable, without the need for continuous relaxations as commonly used in deep generative models (Jang et al. 2017; Maddison et al. 2017), as the likelihood is given in closed form and no gradients of expectations are necessary.

6.2.5 Dealing with correlated parents

Deep learning models are known to pick up on spurious correlations and use shortcuts to fulfil the tasks they are trained to perform (Makar et al. 2021). Imagine a causal model with three variables: x, y and z that are factorised as $p(x, y, z) = p(z|x, y)\, p(y|x)\, p(x)$, such that x and y cause z and x causes y. Let x, y and z be Gaussian distributed according to the following SCM,

$$\begin{aligned}
x &\coloneqq f_x(\epsilon_x^*) = a_x \cdot \epsilon_x^* + b_x \, , & \epsilon_x^* &\sim \mathcal{N}(0, 1) \, , \\
y &\coloneqq f_y(\epsilon_y^*; x) = a_y(x) \cdot \epsilon_y^* + b_y(x) \, , & \epsilon_y^* &\sim \mathcal{N}(0, 1) \, , \\
z &\coloneqq f_z(\epsilon_z^*; x, y) = a_z(x, y) \cdot \epsilon_z^* + b_z(x, y) \, , & \epsilon_z^* &\sim \mathcal{N}(0, 1) \, ,
\end{aligned} \tag{6.11}$$

where $a(\cdot)$ and $b(\cdot)$ describe the effect of the parents on the respective variable. In the case of $a_y \to 0$, the variable y would collapse to be deterministic given x, $f_y(\epsilon_y^*; x) = b_y(x) = \bar{f}_y(x)$. Even though in practice, we will rarely encounter situations in which the mechanism collapses to a deterministic one, it offers a useful perspective of understanding the problem of estimating $f_z(\epsilon_z^*; x, y)$. The estimation of $f_z(\epsilon_z^*; x, y)$ from data alone would be susceptible to shortcuts because $f_z(\epsilon_z^*; x, y) \approx f_z(\epsilon_z^*; x, \bar{f}_y(x)) = \bar{f}_z(\epsilon_z^*; x)$. This scenario is equivalent to x and y having infinite correlation or zero conditional entropy, $H(y|x) = 0$.

However, ideally we want to learn the true $f_z(\epsilon_z^*; x, y)$ rather than the shortcut $\bar{f}_z(\epsilon_z^*; x)$ to enable interventional and causal queries to extrapolate to different $\tilde{p}(y\,|\,\mathbf{pa}_y)$. This problem touches upon topics of shortcut removal (Makar et al. 2021), extrapolation in generative models (Besserve et al. 2021) as well as identifiability of functions expressed by neural networks in general (Khemakhem et al. 2020; Mita et al. 2021; Roeder et al. 2021; Sorrenson et al. 2020; Zhou and Wei 2020). So far, the deep SCM framework learns the different mechanisms $f_{k,\theta}$ with parameters θ by optimising the explicit or implicit (conditional) likelihood of the observed data, $p_\theta(x_k\,|\,\mathbf{pa}_k)$. Without any further constraints to enforce identifiability, this approach can learn various solutions θ that optimise $p_\theta(x_k\,|\,\mathbf{pa}_k)$. Inspired by the notion of *genericity* from (Besserve et al. 2021), we use auxiliary distributions $q_\varphi(\mathbf{pa}_k\,|\,x_k)$ to constrain the space of equivalent solutions. Specifically, in addition to

the regular likelihood we optimise

$$\begin{aligned}&\arg\max_\theta \mathbb{E}_{p_\theta(\tilde{x}_k|\mathrm{do}(\mathbf{pa}_k=\widetilde{\mathbf{pa}}_k),x_k=x_k)}[q_\varphi(\mathbf{pa}_k=\widetilde{\mathbf{pa}}_k|\tilde{x}_k)]\\ &\quad\approx \arg\max_\theta \frac{1}{M}\sum_{m=1}^{M} q_\varphi(\mathbf{pa}_k=\widetilde{\mathbf{pa}}_k|\tilde{x}_k^{(m)}),\quad \tilde{x}_k^{(m)}\sim p_\theta(\tilde{x}_k|\mathrm{do}(\mathbf{pa}_k=\widetilde{\mathbf{pa}}_k),x_k=x_k)\,,\end{aligned} \tag{6.12}$$

where $\tilde{x}_k$ is a counterfactual of x_k with the intervention $\mathrm{do}(\mathbf{pa}_k=\widetilde{\mathbf{pa}}_k)$. Assuming that the auxiliary distributions $q_\varphi(\mathbf{pa}_k|x_k)$ are available or easier to learn than $p(x_k|\mathbf{pa}_k)$, this encourages f_k to account for all parents rather than only the ones with high conditional entropy. Specifically, the auxiliary distributions encourage disentanglement of the effect of the parents $\mathbf{pa}_k$ on the variable x_k by requiring that every parent can individually be recovered from counterfactuals $\tilde{x}_k$. If sufficient interventional or counterfactual data is available it would be possible to follow (Ilse et al. 2021) to directly optimise $p_\theta(\tilde{x}_k|\mathrm{do}(\mathbf{pa}_k=\widetilde{\mathbf{pa}}_k))$ as the intervention would expose the mechanism $f_k(\epsilon_k;\mathbf{pa}_k)$ independently of the observed distributions $p(\mathbf{pa}_k)$ over the parent variables $\mathbf{pa}_k$. Similarly, the application of adequately designed data augmentations can be used instead of interventional data (Ilse et al. 2020).

6.3 Related Work

Deep generative modelling has seen a wide range of contributions since the popularisation of variational auto-encoders (VAEs) (Kingma and Welling 2014), generative adversarial networks (GANs) (Goodfellow et al. 2014), and normalising flows (Rezende and Mohamed 2015). These models have since been employed to capture conditional distributions (Mirza and Osindero 2014; Sohn et al. 2015; Trippe and Turner 2017; Winkler et al. 2019), and VAEs and GANs were also extended to model structured data by incorporating probabilistic graphical models (Johnson et al. 2016; Li et al. 2018; Lin et al. 2018). In addition, deep generative models have been heavily used for (unsupervised) representation learning with an emphasis on disentanglement (Chen et al. 2016; Higgins et al. 2017; Kulkarni et al. 2015; Pati and Lerch 2020). However, even when these methods faithfully capture the distribution of observed data, they are capable of fulfilling only the association rung of the ladder of causation.

Interventions build on the associative capabilities of probabilistic models to enable queries related to changes in causal mechanisms. By integrating a causal graph into the connectivity of a deep model, it is possible to perform interventions with GANs (Kocaoglu et al. 2018) and causal generative NNs (Goudet et al. 2018). VAEs can also express causal links using specific covariance matrices between latent variables, which however restrict the dependences to be linear (Yang et al. 2020). Alternatively, assuming specific causal structures, (Tran and Blei 2018) and (Louizos et al. 2017) proposed different approaches for estimating causal effects in the presence of unobserved confounders. Despite reaching the second rung of the causal ladder, all of these methods

lack tractable abduction capabilities and therefore cannot generate counterfactuals.

Some machine-learning tasks such as explainability, image-to-image translation, or style transfer are closely related to counterfactual queries of the sort 'How would x (have to) change if we (wished to) modify y?'. Here, y could be the style of a picture for style transfer (Gatys et al. 2016), the image domain (e.g. drawing to photo) for image-to-image translation (Isola et al. 2017), the age of a person in natural images (Antipov et al. 2017) or medical scans (Xia et al. 2019), or a predicted output for explainability (Singla et al. 2020). However, these approaches do not explicitly model associations, interventions, nor causal structure. Potentially closest to our work is a method for counterfactual explainability of visual models, which extends CausalGANs (Kocaoglu et al. 2018) to predict reparametrised distributions over image attributes following an assumed causal graph (Martinez and Marca 2019). However, this approach performs no abduction step, instead resampling the noise of attributes downstream from the intervention(s), and does not include a generative model of imaging data. To the best of our knowledge, the proposed DSCM framework is the first flexible approach enabling end-to-end training and tractable inference on all three levels of the ladder of causation for high-dimensional data.

6.4 Case Study 1: Morpho-MNIST

We consider the problem of modelling the causal model of a synthetic dataset based on MNIST digits (LeCun et al. 1998b), where we defined stroke thickness to cause the brightness of each digit: thicker digits are brighter whereas thinner digits are dimmer. This simple dataset allows for examining the three levels of causation in a controlled and measurable environment.

6.4.1 Data Generation

We use morphological transformations on MNIST (Castro et al. 2019) to generate a dataset with known causal structure and access to the 'true' process of generating counterfactuals. The SCM

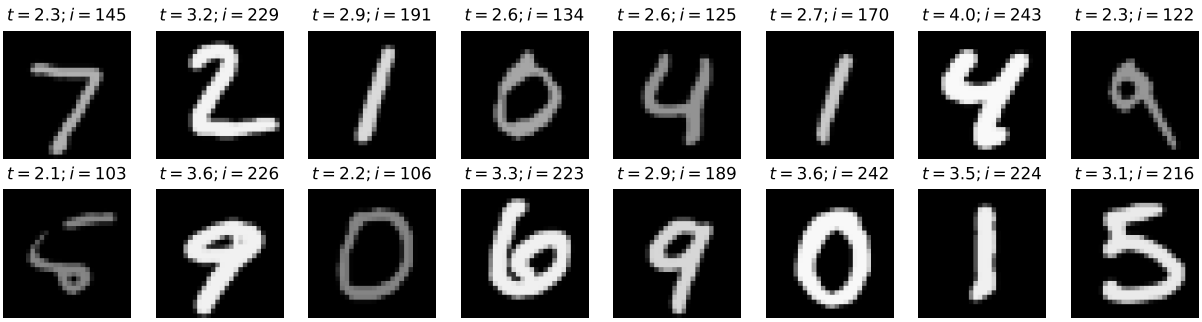
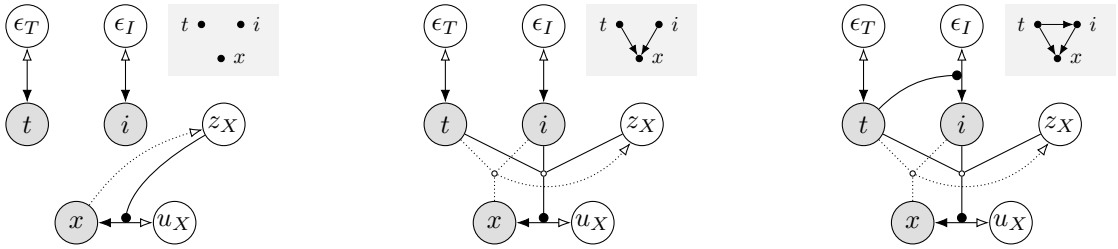


Figure 6.2: Random exemplars from the synthetically generated Morpho-MNIST test dataset.



(a) Independent **(b)** Conditional **(c)** Full

Figure 6.3: Computational graphs of the structural causal models for the Morpho-MNIST example. The image is denoted by x, stroke thickness by t, and image intensity by i. The corresponding causal diagrams are displayed in the top-right corners.

designed for this synthetic dataset is as follows:

$$\begin{aligned} t &:= f_T^*(\epsilon_T^*) = 0.5 + \epsilon_T^* \,, & \epsilon_T^* &\sim \Gamma(10, 5) \,, \\ i &:= f_I^*(\epsilon_I^*; t) = 191 \cdot \sigma(0.5 \cdot \epsilon_I^* + 2 \cdot t - 5) + 64 \,, & \epsilon_I^* &\sim \mathcal{N}(0, 1) \,, \\ x &:= f_X^*(\epsilon_X^*; i, t) = \mathrm{SetIntensity}(\mathrm{SetThickness}(\epsilon_X^*; t); i) \,, & \epsilon_X^* &\sim \mathrm{MNIST} \,, \end{aligned} \tag{6.13}$$

where $\mathrm{SetIntensity}(\,\cdot\,; i)$ and $\mathrm{SetThickness}(\,\cdot\,; t)$ refer to the operations that act on an image of a digit and set its intensity to i and thickness to t, x is the resulting image, ϵ^* is the exogenous noise for each variable, and $\sigma(\,\cdot\,)$ is the logistic sigmoid.

We use the original MNIST dataset (LeCun et al. 1998b) together with the morphometric measurements introduced with Morpho-MNIST (Castro et al. 2019) to add functionality to measure intensity as well as set the intensity and thickness to a given value.

We implement $\mathrm{MeasureIntensity}$ by following the processing steps proposed by (Castro et al. 2019), and measure the intensity i of an image as the median intensity of pixels within the extracted binary mask. Once the intensity is measured, the entire image is rescaled to match the target intensity, with values clamped between 0 and 255 (images are assumed to be in unsigned 8-bit format).

Originally, Morpho-MNIST only proposed relative thinning and thickening operations. We expand those operations to absolute values by calculating the amount of dilation or erosion based on the ratio between target thickness and measured thickness.

Finally, we follow Eq. (6.13) to modify each image within the MNIST dataset and randomly split the original training set into a training and validation set. We show random samples from the resulting test set in Fig. 6.2.

6.4.2 Experimental Setup

We use this setup to study the capabilities of our framework in comparison to models with less causal structure. We adapt the true causal graph from Eq. (6.13) and model thickness and inten-

sity using (conditional) normalising flows and employ a conditional VAE for modelling the image. In particular, we adopt the causal graphs shown in Fig. 6.3 and test a fully independent model (Fig. 6.3a), a conditional decoder model (Fig. 6.3b), as well as our full causal model (Fig. 6.3c). All our experiments were implemented within PyTorch (Paszke et al. 2019) using the Pyro probabilistic programming framework (Bingham et al. 2019).

We use (conditional) normalising flows for all variables apart from the images, which we model using (conditional) deep encoder-decoder architectures. The flows consist of components that constrain the support of the output distribution (where applicable) and components relevant for fitting the distribution. We use unit Gaussians as base distributions for all exogenous noise distributions $P(\boldsymbol{\epsilon})$ and, if available, we use the implementations in PyTorch (Paszke et al. 2019) or Pyro (Bingham et al. 2019) for all transformations. Otherwise, we adapt the available implementations, referring to (Durkan et al. 2019) for details. We indicate with θ the modules with learnable parameters.

We model the mechanisms of the thickness t and intensity i as

$$t \coloneqq f_T(\epsilon_T) = (\exp \circ \mathrm{AffineNormalisation} \circ \mathrm{Spline}_\theta)(\epsilon_T)\,, \tag{6.14}$$

$$i \coloneqq f_I(\epsilon_I; t) = \big(\mathrm{AffineNormalisation} \circ \mathrm{sigmoid} \circ \mathrm{ConditionalAffine}_\theta(\hat{t})\big)(\epsilon_I)\,. \tag{6.15}$$

In the independent model, where i is not conditioned on t, we use instead

$$i \coloneqq f_I(\epsilon_I) = (\mathrm{AffineNormalisation} \circ \mathrm{sigmoid} \circ \mathrm{Spline}_\theta \circ \mathrm{Affine}_\theta)(\epsilon_I)\,. \tag{6.16}$$

We found that including normalisation layers help learning dynamics[4] and therefore include flows to perform commonly used normalisation transformations. For a doubly bounded variable y we learn the flows in unconstrained space and then constrain them by a sigmoid transform and rescale to the original range using fixed affine transformations with bias $\min(Y)$ and scale $[\max(Y) - \min(Y)]$. We constrain singly bounded values by applying an exponential transform to the unbounded values and using an affine normalisation equivalent to a whitening operation in unbounded log-space. We denote those fixed normalisation transforms as $\mathrm{AffineNormalisation}$ and use a hat to refer to the unconstrained, normalised values (e.g. $\widehat{\mathbf{pa}}_k$). The Spline_θ transformation refers to first-order neural spline flows (Durkan et al. 2019), Affine_θ is an element-wise affine transformation, and $\mathrm{sigmoid}$ refers to the logistic function. $\mathrm{ConditionalAffine}_\theta(\cdot)$ is a regular affine transform whose transformation parameters are predicted by a context neural network taking $\cdot$ as input. In the case of $f_I(\epsilon_I; t)$, the context network is represented by a simple linear transform. Further, we

[4] We observed that not normalising the inputs can lead to the deep models prioritising learning the dependence on the variable with largest magnitude. We provide some first insights into a similar problem when studying correlated parents in Section 6.6. However, this phenomenon should be investigated further.

model x using a low-level flow:

$$h_X(u_X; \mathbf{pa}_X) = [\text{Preprocessing} \circ \text{ConditionalAffine}_\theta(\widehat{\mathbf{pa}}_X)](u_X)\,, \tag{6.17}$$

where the ConditionalAffine transform practically reparametrises the noise distribution into another Gaussian distribution and Preprocessing describes a fixed preprocessing transformation. We follow the same preprocessing as used with RealNVP (Dinh et al. 2017). The context network for the conditional affine transformation is the high-level mechanism $g_X(z_X; \mathbf{pa}_X)$ and is implemented as a decoder network that outputs the bias for of the affine transformation, while the log-variance is fixed to $\log \sigma^2 = -5$. We implement the decoder network as a CNN:

$$\begin{aligned} g_X(z_X; \mathbf{pa}_X) = (&\text{Conv}_\theta(1;1;1;0) \circ \text{ConvTranspose}_\theta(1;4;2;1) \circ \text{ReLU} \circ \text{BN}_\theta \\ &\circ \text{ConvTranspose}_\theta(64;4;2;1) \circ \text{Reshape}(64,7,7) \\ &\circ \text{ReLU} \circ \text{BN}_\theta \circ \text{Linear}_\theta(1024) \\ &\circ \text{ReLU} \circ \text{BN}_\theta \circ \text{Linear}_\theta(1024))([z_X, \widehat{\mathbf{pa}}_X])\,, \end{aligned} \tag{6.18}$$

where the operators describe neural network layers as follows: BN is batch normalisation; ReLU the ReLU activation function; $\text{Conv}(c;k;s;p)$ and $\text{ConvTranspose}(c;k;s;p)$ are a convolution or transposed convolution using a kernel with size k, a stride of s, a padding of p and outputting c channels; $\text{Linear}(h)$ is a linear layer with h output neurons; and $\text{Reshape}(\cdot)$ reshapes its inputs into the given shape $\cdot$. Lastly, $[z_X, \mathbf{pa}_X]$ denotes the concatenation of z_X and $\mathbf{pa}_X$, and $z_X \in \mathbb{R}^{16}$.

Equivalently, we implement the encoder function as a simple CNN that outputs mean and log-variance of an independent Gaussian:

$$\begin{aligned} e_X(x; \mathbf{pa}_X) = \big(&[\text{Linear}_\theta(16), \text{Linear}_\theta(16)] \circ [\text{LeakyReLU}(0.1), \widehat{\mathbf{pa}}_X] \\ &\circ \text{BN}_\theta \circ \text{Linear}_\theta(100) \circ \text{Reshape}(128 \cdot 7 \cdot 7) \\ &\circ \text{LeakyReLU}(0.1) \circ \text{BN}_\theta \circ \text{Conv}_\theta(128;4;2,1) \\ &\circ \text{LeakyReLU}(0.1) \circ \text{BN}_\theta \circ \text{Conv}_\theta(64;4;2,1)\big)(x)\,, \end{aligned} \tag{6.19}$$

where $\text{LeakyReLU}(\ell)$ is the leaky ReLU activation function with a leakiness of ℓ.

We use Adam (Kingma and Ba 2015) for optimisation with batch size of 256 and a learning rate of 10^{-4} for the encoder-decoder and 0.005 for the covariate flows. We set the number of particles (MC samples) for estimating the ELBO to 4. We use 32 MC samples for estimating reconstruction and counterfactuals. We train all models for 1000 epochs and report the results of the model with the best validation loss.

Table 6.1: Comparison of the associative abilities of the models shown in Fig. 6.3. The image is denoted by x, thickness by t, and intensity by i. Quantities with $\geq$ are lower bounds. MAE refers to the mean absolute error between pixels of the original image and of its reconstruction.

Model	$\log p(x,t,i) \geq$	$\log p(x\|t,i) \geq$	$\log p(t)$	$\log p(i\|t)$	$\mathrm{MAE}(x,x')$
Independent	-5925.26	-5919.14	-0.93	-5.19	4.50
Conditional	-5526.50	-5520.37	-0.93	-5.19	4.26
Full	-5692.94	-5687.71	-0.93	-4.30	4.43

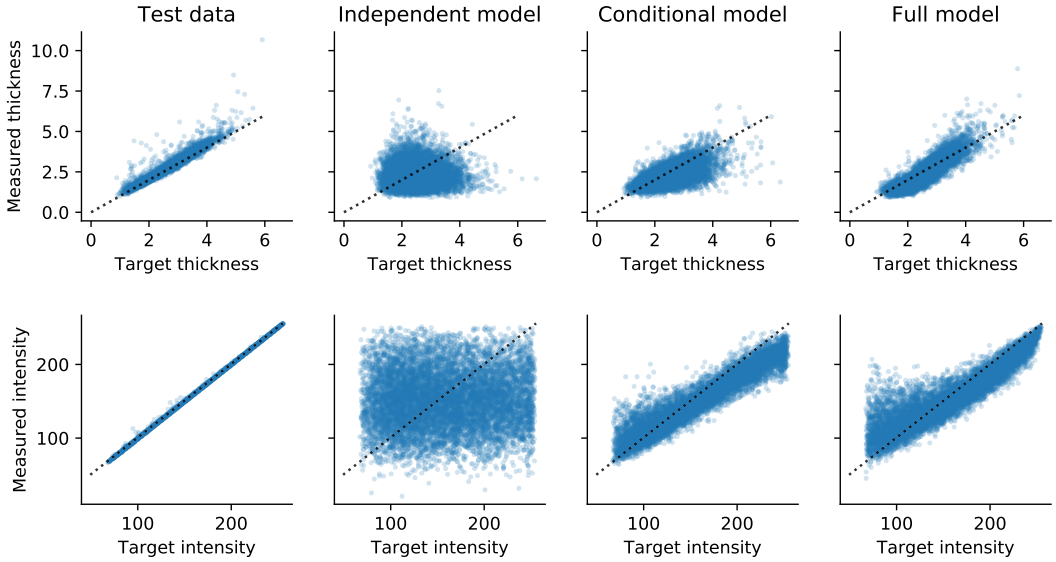
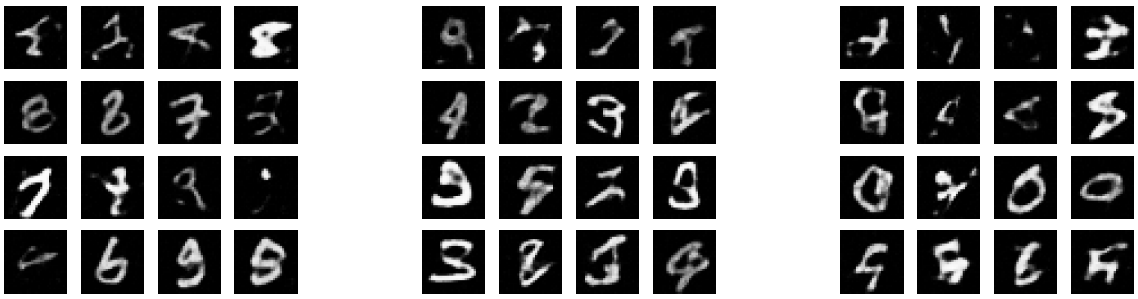


Figure 6.4: Comparison of the target covariates and the corresponding values measured from the generated images. The leftmost column refers to the accuracy of the $\mathrm{SetThickness}$ and $\mathrm{SetIntensity}$ transforms used in generating the synthetic dataset, and the remaining three columns describe the fidelity of samples generated by each of the learned models. While images sampled from the independent model are trivially inconsistent with the sampled covariates, the conditional and full models show comparable conditioning performance.

6.4.3 Results

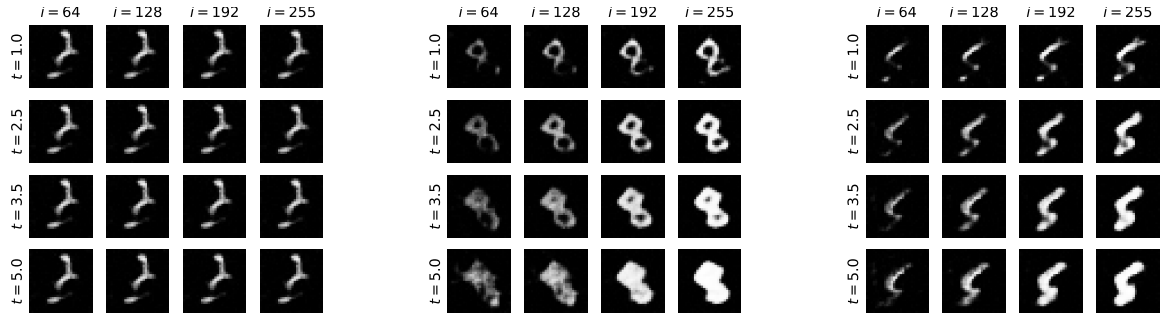
We quantitatively compare the associative capabilities of all models by evaluating their evidence lower bound (Eq. (6.3)), log-likelihoods and reconstruction errors as shown in Table 6.1. We find that performance improves consistently with the model's capabilities: enabling conditional image generation improves $p(x|t,i)$, and adding a causal dependency between t and i improves $p(i|t)$. Further, we examine samples of the conditional and unconditional distributions in Figs. 6.4 to 6.7.

The interventional distributions can be directly compared to the true generative process. Figure 6.8 shows that the densities predicted by our full model after intervening on t closely resemble the true behaviour. The conditional and independent models operate equivalently to each other and are incapable of modelling the relationship between t and i, capturing only their marginal distributions.



(a) Independent **(b)** Conditional **(c)** Full

Figure 6.5: Random samples generated by the independent, conditional and full model. Note how all models appear to have the same unconditional generation capacity.



(a) Independent **(b)** Conditional **(c)** Full

Figure 6.6: Conditional samples generated by the independent, conditional, and full model. The high-level noise, z_X, is shared for all samples from each model, ensuring the same 'style' of the generated digit. The independent model generates images independent of the thickness and intensity values, resulting in identical samples. For the conditional and full models, thickness and intensity change consistently along each column and row, respectively.

Additionally, we take an explicit look at the differences between intervening and conditioning in Fig. 6.9.

Lastly, we examine the full model's ability to generate counterfactuals. In this special case, where the true data generating process is known, it is possible to evaluate against reference counterfactuals that are impossible to obtain in most real-world scenarios. We compare all models on the task of generating counterfactuals with intervention $\mathrm{do}(t+2)$ and compute the mean absolute errors between the generated and the reference counterfactual image. For this task, the models perform in order of their complexity: the independent model achieved 41.6, the conditional 31.8, and the full model achieved a MAE of 17.6. This emphasises that, although wrongly specified models will give wrong answers to counterfactual queries (and interventions; see Fig. 6.8), the results are consistent with the assumptions of each model. The independent model lacks any relationship of the image on thickness and intensity and therefore does not change the image under the given intervention. The conditional model does not model any dependency of intensity on thickness, which leads to counterfactuals with varying thickness but constant intensity. Examples of previously unseen images and generated counterfactuals using the full model are shown in Fig. 6.10 for qualitative

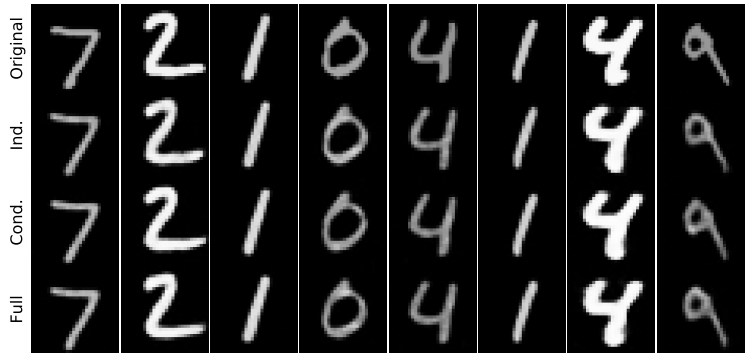


Figure 6.7: Reconstructions. These are computed as Monte Carlo averages approximating $\mathbb{E}_{Q(z_X|e_X(x;\mathbf{pa}_X))}[g_X(z_X;\mathbf{pa}_X)]$, where e_X and g_X are the image encoder and decoder networks. All models seem capable of producing faithful reconstructions.

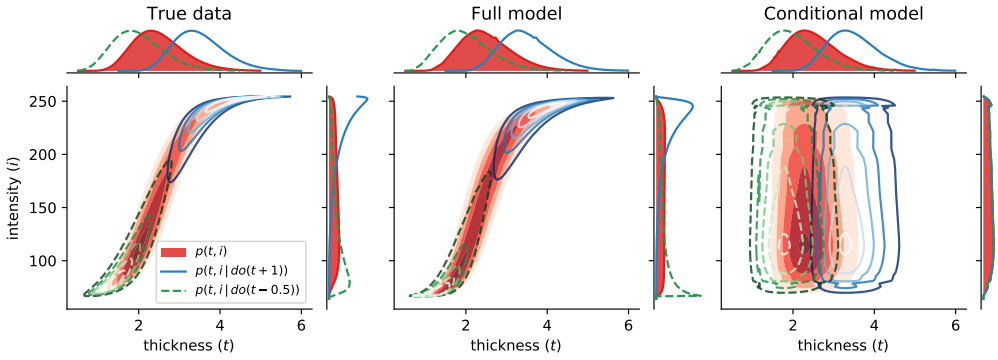


Figure 6.8: Distributions of thickness and intensity in the true data (left), and learned by the full (centre) and conditional (right) models. Contours depict the observational (red, shaded) and interventional joint densities for $\mathrm{do}(t := f_T(\epsilon_T) + 1)$ (blue, solid) and $\mathrm{do}(t := f_T(\epsilon_T) - 0.5)$ (green, dashed).

examination. We see that our model is capable of generating convincing counterfactuals that preserve the digit identity while changing thickness and intensity consistently with the underlying true causal model. A larger range of counterfactual samples can be seen in Fig. 6.11.

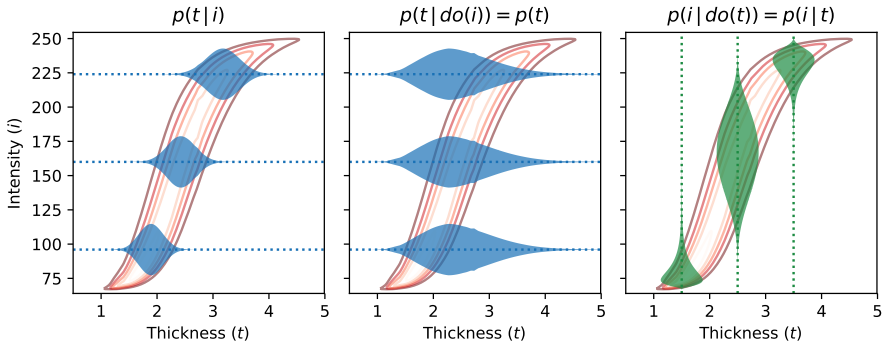


Figure 6.9: Difference between conditioning and intervening, based on the trained full model. The joint density $p(t, i)$ is shown as contours in the background, for reference, and the 'violin' shapes represent the density of one variable when conditioning or intervening on three different values of the other variable. Since t causes i, notice how $p(t \mid i)$ (left) is markedly different from $p(t \mid \mathrm{do}(i))$ (middle), which collapses to $p(t)$. On the other hand, $p(i \mid \mathrm{do}(t))$ and $p(i \mid t)$ (right) are identical.

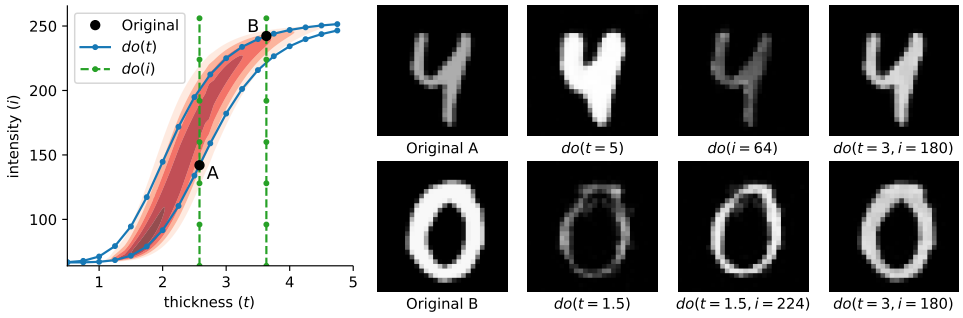


Figure 6.10: Counterfactuals generated by the full model. (left) Counterfactual 'trajectories' of two original samples, A and B, as their thickness and intensity are modified, overlaid on the learned joint density $p(t, i)$. (right) Original and counterfactual images corresponding to samples A and B.

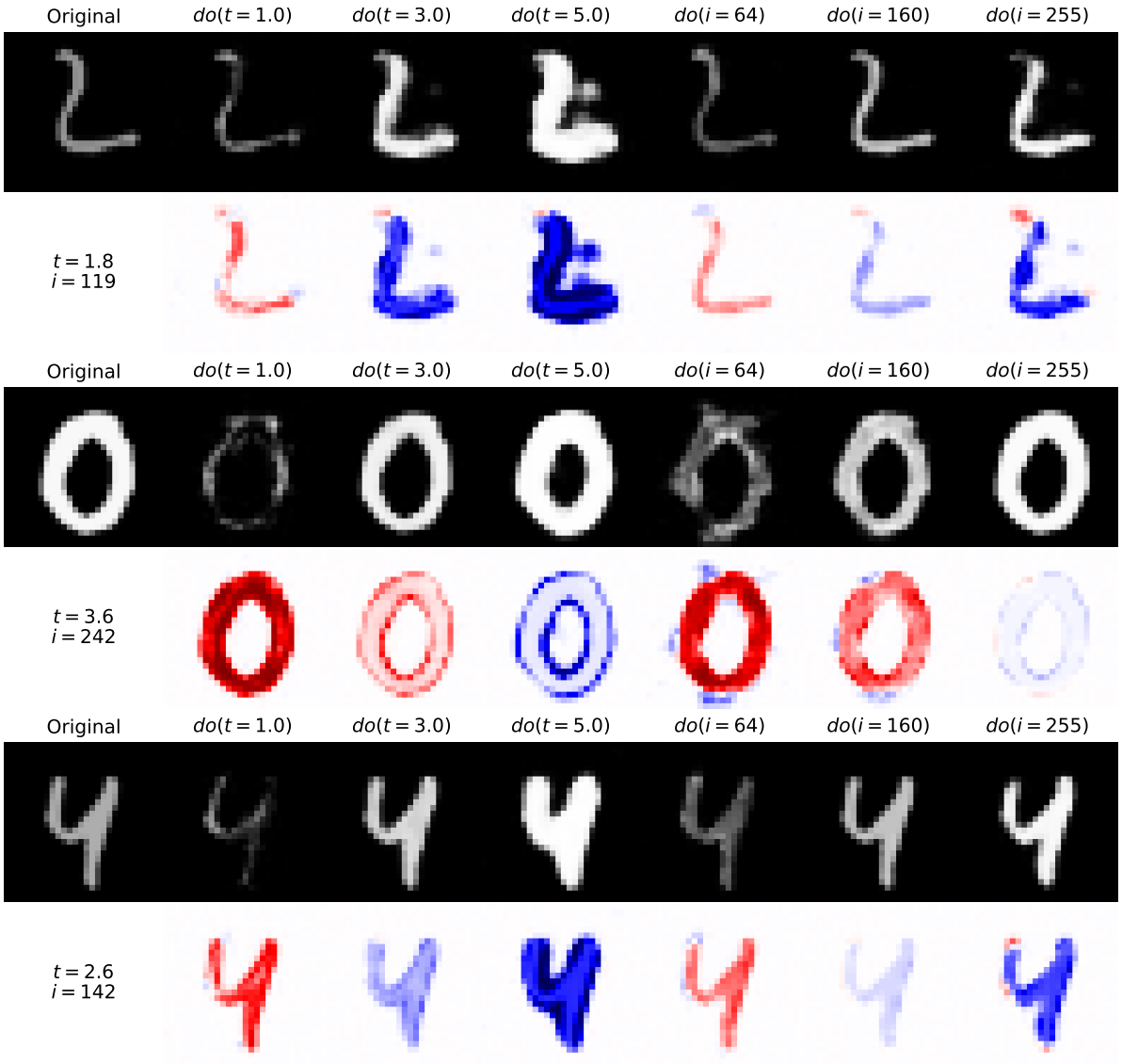


Figure 6.11: Original samples and counterfactuals from the full model. The first column shows the original image and true values of the non-imaging data. The even rows show the difference maps between the original image and the corresponding counterfactual image. We observe that all counterfactuals preserve the digits' identity and style. Our model even generates sensible counterfactual images (with some artefacts) in very low-density regions, e.g. '0' with $\mathrm{do}(i = 64)$ (thick but dim), and very far from the original, e.g. '2' with $\mathrm{do}(t = 5.0)$.

6.5 Case Study 2: Brain Imaging

Our real-world application touches upon fundamental scientific questions in the context of medical imaging: how would a person's anatomy change if particular traits were different? We illustrate with a simplified example that our DSCM framework may provide the means to answer such counterfactual queries, which may enable entirely new research into better understanding the physical manifestation of lifestyle, demographics, and disease. Note that any conclusions drawn from a model built in this framework are strictly contingent on the correctness of the assumed SCM. Here, we model the appearance of brain MRI scans given the person's age and biological sex, as well as brain and ventricle volumes[5], using population data from the UK Biobank (Sudlow et al. 2015). Ventricle and total brain volumes are two quantities that are closely related to brain age (Peters 2006) and can be observed relatively easily. We adopt the causal graph shown in Fig. 6.13 and otherwise follow the same training procedure as for the MNIST experiments.[6]

6.5.1 Data Generation

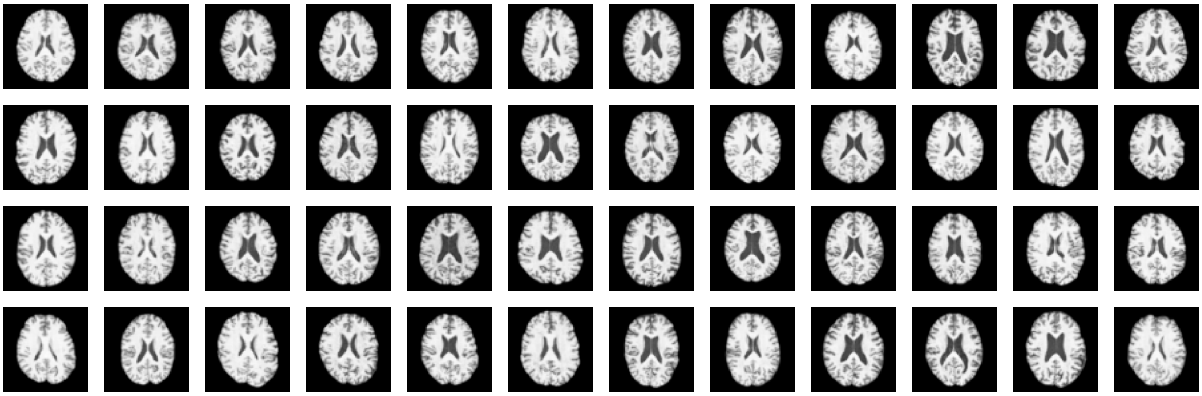


Figure 6.12: Random examplars from the test set of the adopted UK Biobank dataset (Sudlow et al. 2015).

The original three-dimensional (3D) T1-weighted brain MRI scans have been pre-processed by the data providers of the UK Biobank Imaging study using the FSL neuroimaging toolkit (Alfaro-Almagro et al. 2018). The pre-processing involves skull removal, bias field correction, and automatic segmentation of brain structures. In addition, we have rigidly registered all scans to the standard MNI atlas space using an in-house image registration tool, which enabled us to extract anatomically corresponding mid-axial 2D slices that were used for the experiments presented in this paper. The 2D slices were normalised in intensity by mapping the minimum and maximum values inside the brain mask to the range $[0, 255]$. Background pixels outside the brain were set to zero. Age and biological sex for each subject were retrieved from the UK Biobank database along with the

[5] Ventricles are fluid-filled cavities identified as the symmetric dark areas in the centre of the brain.

[6] Note that Fig. 6.13 shows s with a unidirectional arrow from ϵ_S: since s has no causal parents in this SCM, abduction of ϵ_S is not necessary. If it had parents and we wished to estimate discrete counterfactuals under upstream interventions, this could be done with a Gumbel–max parametrisation as described in Section 6.2.4.

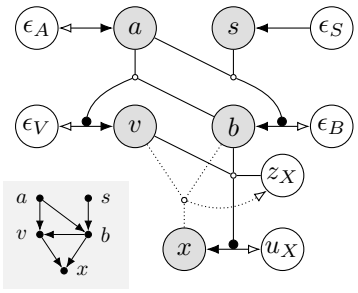


Figure 6.13: Computational graph for the brain imaging example: Variables are image (x), age (a), sex (s), and brain (b) and ventricle (v) volumes.

pre-computed brain and ventricle volumes. These volumes are derived from the 3D segmentation maps obtained with FSL, and although these are image-derived measurements, they may serve as reasonable proxies of the true measurements within our (simplified yet plausible) causal model of the physical manifestation of the brain anatomy.

6.5.2 Experimental Setup

The setup for the brain imaging experiment closely follows the MNIST example as described in Section 6.4.2. We randomly split the available $13,750$ brain images into train, validation and test sets with the respective ratios 70%, 15% and 15%. During training, we randomly crop the brain slices from their original size of $233\,\text{px} \times 197\,\text{px}$ to $192\,\text{px} \times 192\,\text{px}$ and use center crops during validation and testing. The cropped images are downsampled by a factor of 3 to a size of $64\,\text{px} \times 64\,\text{px}$.

We use the same low-level mechanism for the image x as with MNIST images but change the encoder and decoder functions to a deeper architecture with 5 scales consisting of 3 blocks of $(\text{LeakyReLU}(0.1) \circ \text{BN}_\theta \circ \text{Conv}_\theta)$ each as well as a linear layer that converts to and from the latent space with 100 dimensions. We directly learn the binary probability of the sex s and use the following invertible transforms to model the age a, brain volume b, and ventricle volume v as

$$a \coloneqq f_A(\epsilon_A) = \big(\exp \circ \text{AffineNormalisation} \circ \text{Spline}_\theta\big)(\epsilon_A)\,, \tag{6.20}$$

$$b \coloneqq f_B(\epsilon_B; s, a) = \big(\exp \circ \text{AffineNormalisation} \circ \text{ConditionalAffine}_\theta([s, \widehat{a}])\big)(\epsilon_B)\,, \tag{6.21}$$

$$v \coloneqq f_V(\epsilon_V; a, b) = \big(\exp \circ \text{AffineNormalisation} \circ \text{ConditionalAffine}_\theta([\hat{b}, \widehat{a}])\big)(\epsilon_V)\,, \tag{6.22}$$

where the context networks are implemented as a fully-connected network with 8 and 16 hidden units, and a $\text{LeakyReLU}(0.1)$ nonlinearity.

6.5.3 Results

The learned DSCM is capable of all three levels of the causal hierarchy. We present the analysis of lower levels in Figs. 6.14 to 6.17 and and focus here on counterfactuals, shown in Fig. 6.18.

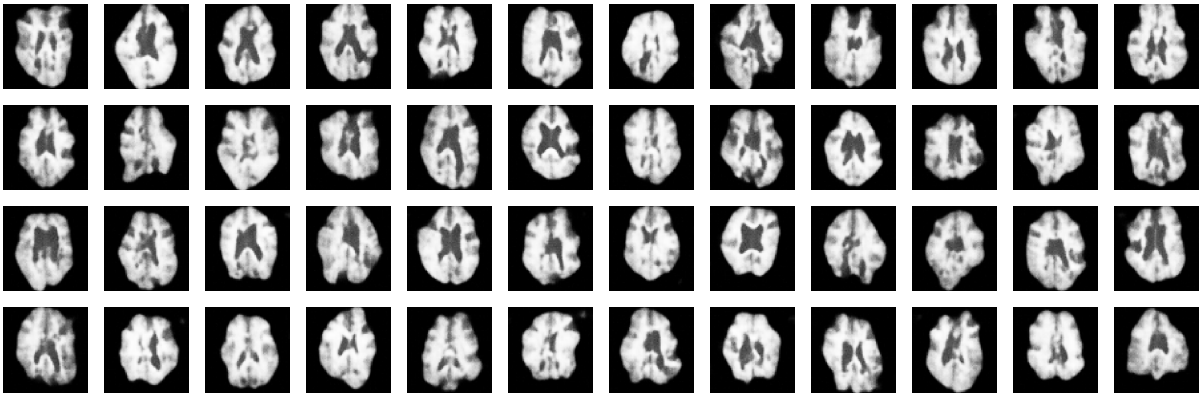


Figure 6.14: Random samples from the model trained on the UK Biobank dataset.

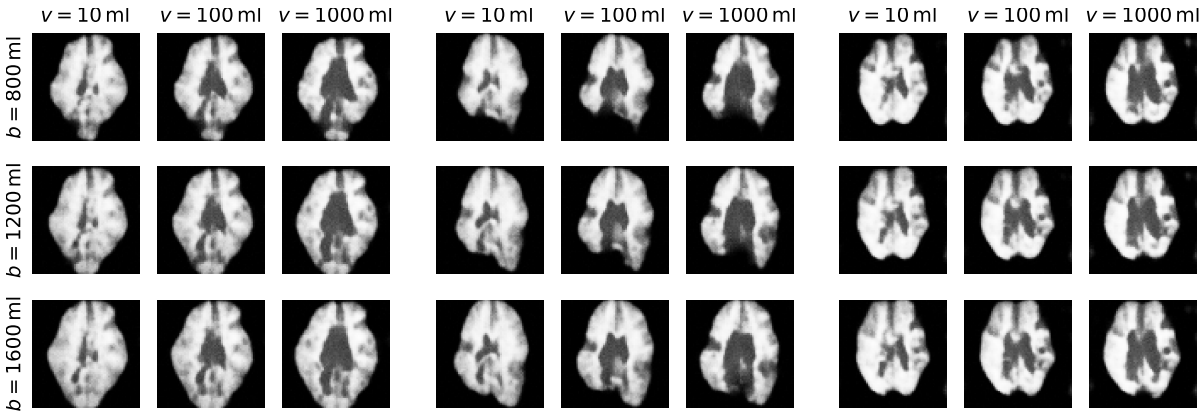


Figure 6.15: Conditional samples from the model trained on the UK Biobank dataset. Images in each 3×3 block share the same the high-level noise vector, z_X. Each row consistently changes the brain size, whereas each column changes the ventricle volume.

The difference maps show plausible counterfactual changes: increasing age causes slightly larger ventricles while decreasing the overall brain volume (fourth column). In contrast, directly increasing the brain volume has an opposite effect on the ventricles compared to changing age (sixth column). Intervening on ventricle volume has a much more localised effect (last column), while intervening on the categorical variable of biological sex has smaller yet more diffuse effects. Note how the anatomical 'identity' (such as the cortical folding) is well preserved after each intervention.

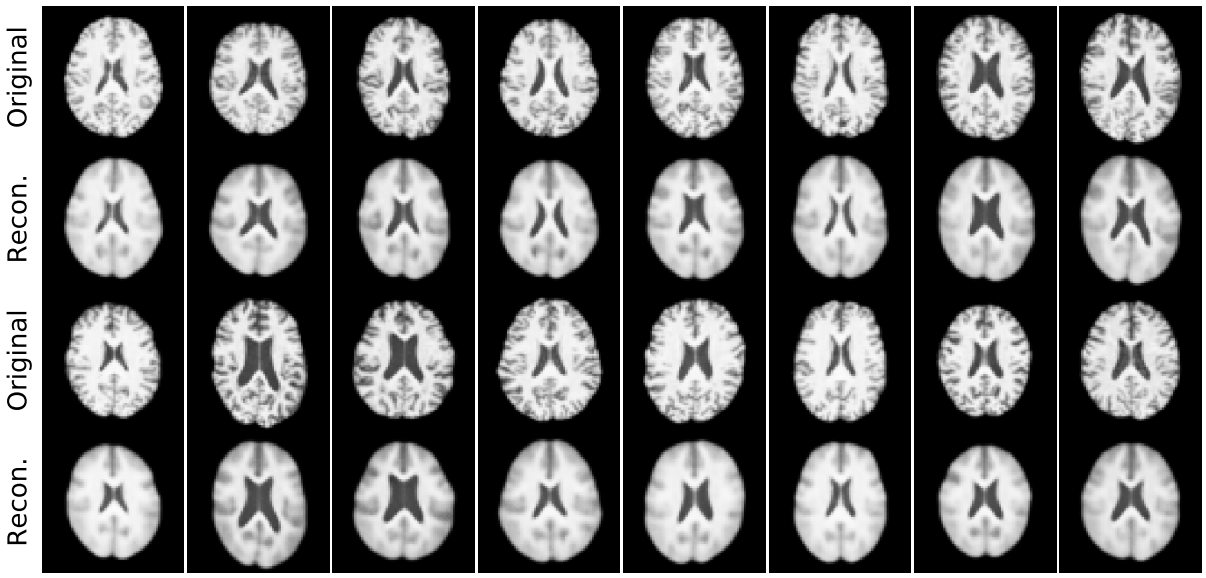
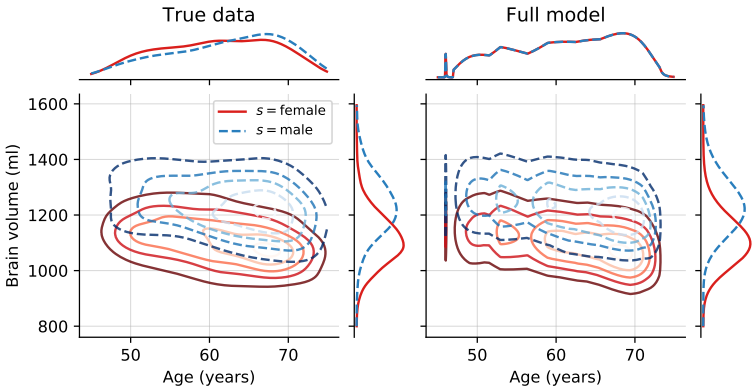
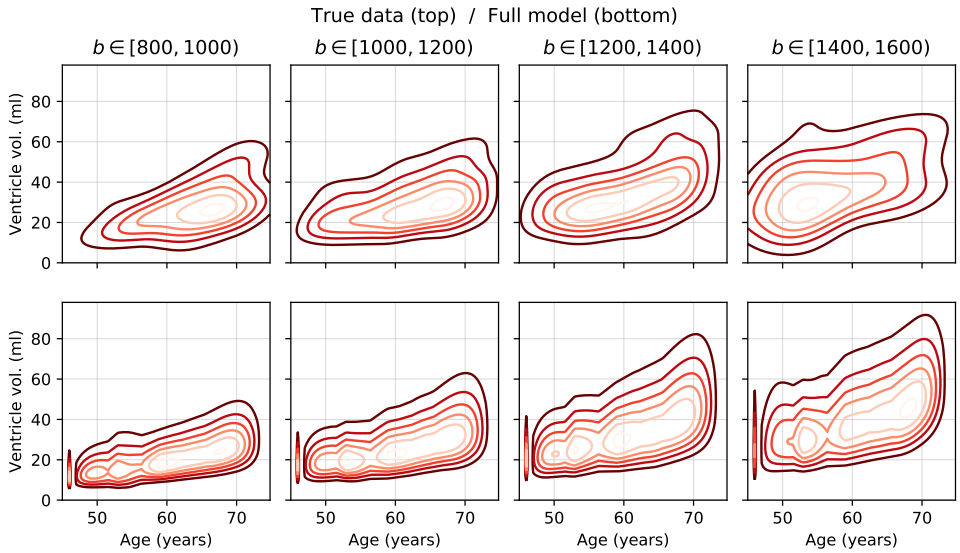


Figure 6.16: Original samples and reconstructions from the model trained on the UK Biobank dataset.



(a) Age vs. brain volume: $p(a, b \mid s)$. Here we see differences in head size across biological sexes (reflected in brain volume), as well as a downward trend in brain volume as age progresses.



(b) Age vs. ventricle volume: $p(a, v \mid b \in \cdot)$. As expected from the literature (Peters 2006), we observe a consistent increase in ventricle volume with age, in addition to a proportionality relationship with the overall brain volume.

Figure 6.17: Densities for the true data (KDE) and for the learned model. The overall trends and interactions present in the true data distribution seem faithfully captured by the model.

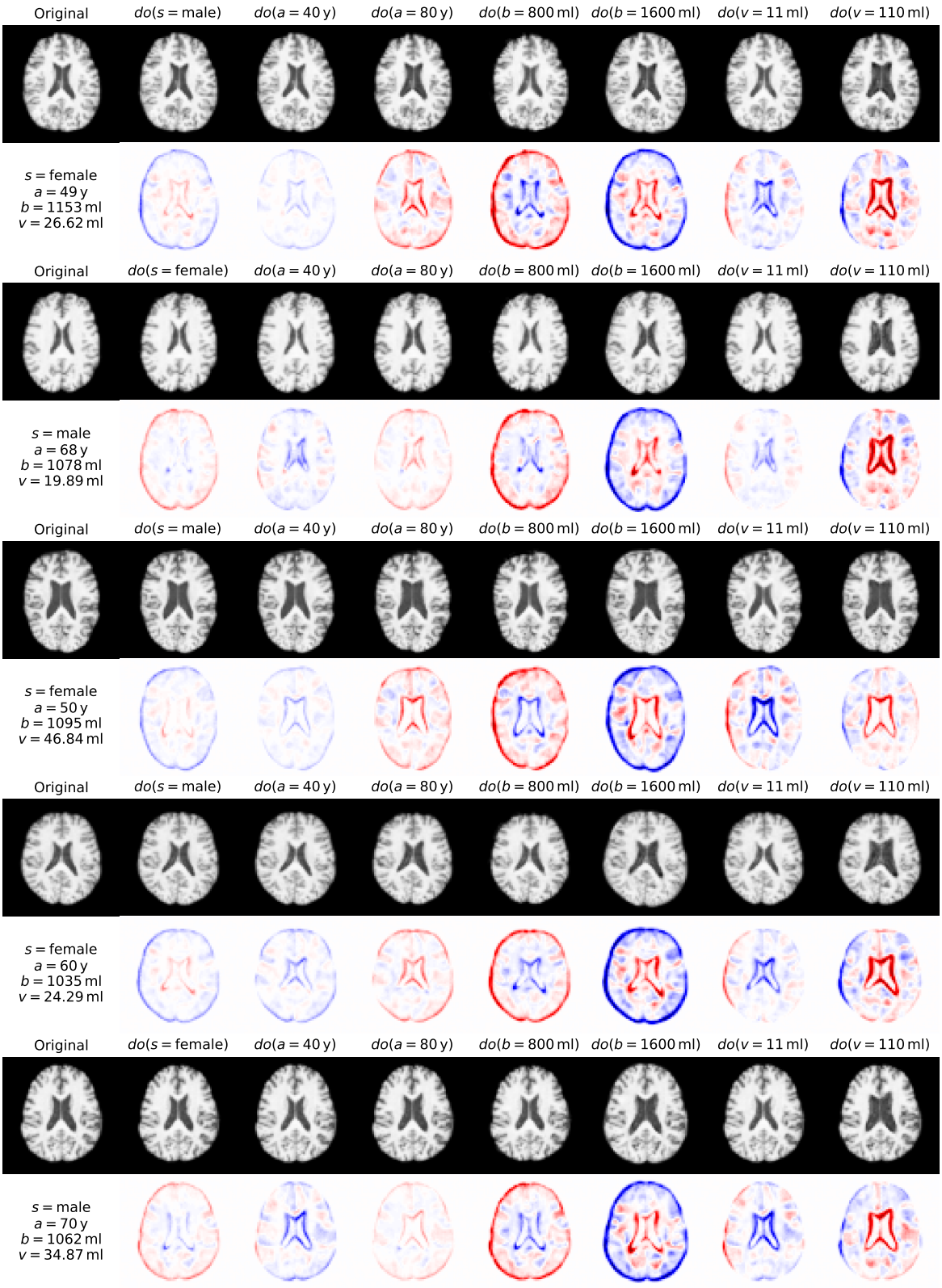


Figure 6.18: Original samples and counterfactuals from the model trained on the UK Biobank dataset. The first column shows the original image and true values of the non-imaging data. The even rows show the difference maps between the original image and the corresponding counterfactual image.

6.6 Case Study 3: Studying correlated parents on Morpho-MNIST

We follow the setup from Section 6.4 to investigate the behaviour of the deep SCM framework in the presence of correlated parents as described in Section 6.2.5. This experiment first shows that the naive training of a deep SCM-based model in this scenario is not capable of learning the true independent mechanisms. We then include the use of auxiliary distributions for constraining the learned model and show that this approach yields improved results.

6.6.1 Data Generation

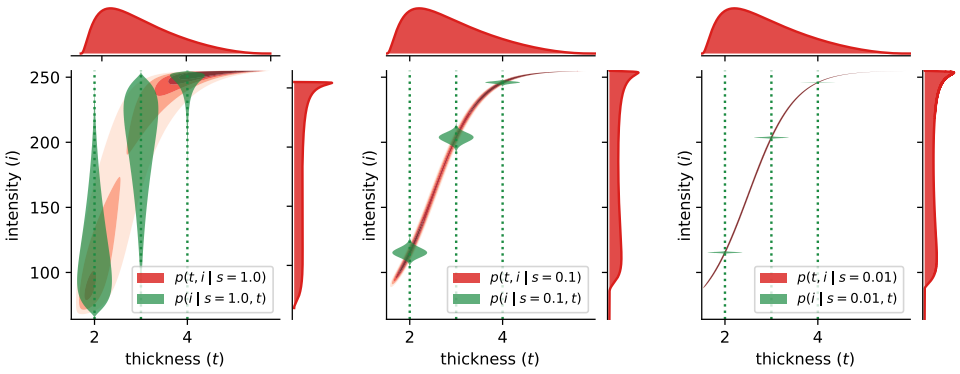


Figure 6.19: Impact of different correlation parameters s on the joint density. The joint density $p(t, i)$ is shown as contours in the background, for reference, and the 'violin' shapes represent the conditional density $p(i|t)$ of the intensity i when conditioning on three different values for thickness t. With decreasing correlation parameter s, we see that the intensity i becomes more deterministic as a function of the thickness t.

We use the same variables and morphological operations as before but change the mechanisms for the thickness variable t and the intensity variable i and regenerate the dataset. The SCM for the synthetic data[7] is as follows:

$$\begin{aligned} t &:= f_T^*(\epsilon_T^*) = 4.5 \cdot \sigma(\epsilon_T^* - 1) + 1.5\,, & \epsilon_T^* &\sim \mathcal{N}(0,1)\,, \\ i &:= f_I^*(\epsilon_I^*; t) = 191 \cdot \sigma(s \cdot \epsilon_I^* + 2 \cdot t - 5) + 64\,, & \epsilon_I^* &\sim \mathcal{N}(0,1)\,, \\ x &:= f_X^*(\epsilon_X^*; i, t) = \mathrm{SetIntensity}(\mathrm{SetThickness}(\epsilon_X^*; t); i)\,, & \epsilon_X^* &\sim \mathrm{MNIST}\,, \end{aligned} \tag{6.23}$$

where s is a parameter controlling the correlation between i and t – smaller s corresponding to higher correlation. We generate three variants of this dataset with s set to $s = 1$, $s = 0.1$, and $s = 0.01$. We visualise the impact of different correlation parameters s in Fig. 6.19, showing the joint distribution $p(t, i)$ of the thickness t and the intensity i together with violin plots for conditional probabilities of $p(i|t)$. We observe that with decreasing correlation parameter s, the conditional

[7] Note that this SCM differs to the on in Eq. (6.13) by changes in the mechanisms for i and t.

(a) Full **(b)** Full with auxiliary constraints

Figure 6.20: Computational graphs of the structural causal models for the Morpho-MNIST experiment studying the effect of correlation in parent variables. The image is denoted by x, stroke thickness by t, and image intensity by i. The corresponding causal diagrams are displayed in the top-right corners. The red dotted arrows from x to i and t in Fig. 6.20b refer to the auxiliary distributions described in Section 6.2.5.

probability $p(i|t)$ of intensity i given thickness t becomes more deterministic, while the marginal distribution $p(i)$ barely changes.

6.6.2 Experimental Setup

We use the same setup and the full model as in Section 6.4.2 to investigate the behaviour when naively training a deep SCM on the different synthetic datasets. We model the auxiliary distributions as $q_\varphi(i,t|x) = q_\varphi(i|x)\,q_\varphi(t|x)$ and approximate them with convolutional neuronal networks with four scales of $(\mathrm{ReLU} \circ \mathrm{Conv}_\theta)$ and a final linear layer. The training consists of a two step procedure. First, the auxiliary distributions are jointly trained with the rest of the model to optimise the log-likelihood of the observational data:

$$\begin{aligned}\arg\max_\theta \log p_\theta(i,t,x) &= \arg\max_\theta \log p_\theta(t) + \log p_\theta(i|t) + \log p_\theta(x|i,t)\\ \arg\max_\varphi \log q_\varphi(i,t|x) &= \arg\max_\varphi \log q_\varphi(i|x) + \log q_\varphi(t|x)\end{aligned} \tag{6.24}$$

Second, the image distribution $p_\theta(x|i,t)$ and its mechanism f_x are optimised to maximise the counterfactual log-likelihood of the auxiliary distribution $\log q(i=\tilde{i}, t=\tilde{t}|\tilde{x}) = \log q(t=\tilde{t}|\tilde{x}) + \log q(i=\tilde{i}|\tilde{x})$ while freezing φ:

$$\begin{aligned}&\arg\max_\theta \mathbb{E}_{p_\theta(\tilde{x}|\mathrm{do}(i=\tilde{i},t=\tilde{t}),x=x)}[q_\varphi(i=\tilde{i},t=\tilde{t}|\tilde{x})]\\ &\approx \arg\max_\theta \frac{1}{M}\sum_{m=1}^{M} q_\varphi(i=\tilde{i},t=\tilde{t}|\tilde{x}^{(m)}), \quad \tilde{x}^{(m)} \sim p_\theta(\tilde{x}|\mathrm{do}(i=\tilde{i},t=\tilde{t}),x=x)\,,\end{aligned} \tag{6.25}$$

This constrains the mechanism to rely both on intensity i and thickness t when generating the image x, because the auxiliary distributions encourage $\tilde{x}$ to contain information about values of its

Table 6.2: Comparison of the associative and counterfactual abilities of the full model on datasets generated with different variance parameter s. The image is denoted by x, thickness by t, and intensity by i. Quantities with $\geq$ are lower bounds. $\mathrm{MAE}(x, x')$ refers to the mean absolute error between pixels of the original image and of its reconstruction. $\mathrm{MAE}(x_{CF}, x'_{CF})$ refers to the mean absolute error between pixels of the true counterfactual image and the generated counterfactual image when performing the intervention $\mathrm{do}(i' = 255 - i + 64)$.

s	$\log p(x,t,i) \geq$	$\log p(x\|t,i) \geq$	$\log p(t)$	$\log p(i\|t)$	$\mathrm{MAE}(x,x')$	$\mathrm{MAE}(x_{CF}, x'_{CF})$
0.01	−6033.83	−6029.51	−1.12	−3.21	4.48	12.21
0.1	−6045.91	−6041.37	−1.12	−3.42	4.50	7.61
1	−6309.51	−6303.80	−1.13	−4.57	4.59	4.42

counterfactual parents, $\tilde{i}$ and $\tilde{t}$. In practice, we shuffle the intensity and thickness values in a batch to decide which interventions to perform. We use a single particle to estimate the counterfactual $\tilde{x}_k$ and use a factor of 0.001 to scale the negative log-likelihood of the auxiliary distributions for optimisation.

6.6.3 Results

We quantitatively compare the associative and counterfactual capabilities of the full model without auxiliary constraints trained on the datasets generated with different correlation parameters s in Table 6.2. By evaluating their evidence lower bound, log-likelihoods and reconstruction errors, we find that all models achieve comparable associative performance. Models trained on datasets with higher variance parameter s (lower correlation) exhibit slightly worse associative performance which we attribute to the higher complexity of the modelled data.

Previously, we hypothesised that with increased determinism of the intensity i given the thickness t, the learned model would increasingly rely on the thickness variable t and ignore the intensity i. We test this hypothesis by comparing the mean absolute error between true counterfactuals x_{CF} and the predicted ones $\tilde{x}_{CF}$ for the intervention $\mathrm{do}(i' = 255 - i + 64)$ as shown in Table 6.2. We experimentally verify this hypothesis by finding that the models trained on dataset generated with lower s yield worse MAEs. This can be visually examined in Fig. 6.22, which shows counterfactuals predicted from the model trained on the dataset generated with $s = 0.001$. The visual inspection indicates that the model is still capable of predicting counterfactuals for interventions on thickness, where it uses the thickness variable as an indicator for intensity as well.

We add the auxiliary constraints to the model and evaluate its associative and counterfactual abilities as shown in Table 6.3. Both models achieve comparable performance across the evidence lower bounds, log-likelihoods and reconstruction errors, with the constrained model showing improvements in modelling the image variable. This is aligned with the fact that the constraints are only applied to the learning of that mechanism. However, the constrained model achieves a sig-

Table 6.3: Comparison of the associative and counterfactual abilities of the full model with and without auxiliary constraints. The image is denoted by x, thickness by t, and intensity by i. Quantities with $\geq$ are lower bounds. $\mathrm{MAE}(x, x')$ refers to the mean absolute error between pixels of the original image and of its reconstruction. $\mathrm{MAE}(x_{CF}, x'_{CF})$ refers to the mean absolute error between pixels of the true counterfactual image and the generated counterfactual image when performing the intervention $\mathrm{do}(i' = 255 - i + 64)$.

Aux.	$\log p(x,t,i) \geq$	$\log p(x\|t,i) \geq$	$\log p(t)$	$\log p(i\|t)$	$\mathrm{MAE}(x,x')$	$\mathrm{MAE}(x_{CF}, x'_{CF})$
✗	-6033.83	-6029.51	-1.12	-3.21	4.48	12.21
✓	-5963.70	-5959.38	-1.11	-3.21	4.45	9.23

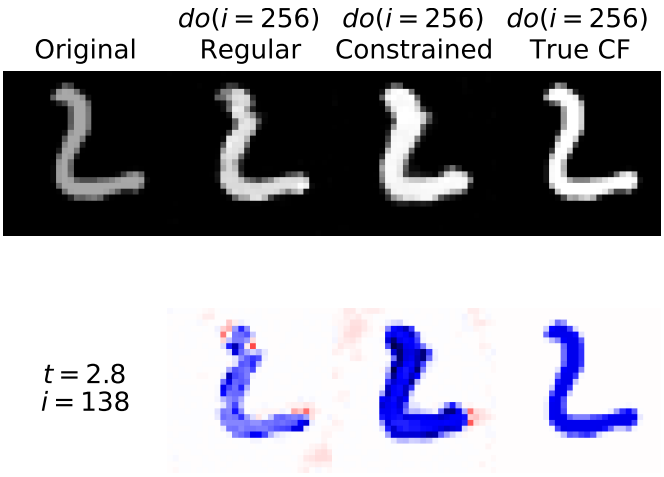


Figure 6.21: Comparison of counterfactuals from the full model with and without auxiliary constraints trained on the dataset generated with $s = 0.001$. The first column shows the original image and true values of the non-imaging data. The last column shows the true counterfactual image generated using the operations from the underlying true SCM. The second row shows the difference maps between the original image and the corresponding counterfactual image. The counterfactual generated by the constrained model resembles the true counterfactual more closely than the one from the regular model without constraints.

nificantly better mean absolute error on the counterfactual prediction task. This suggests that the addition of the counterfactual constraints helps the learning of the image mechanism f_x to rely less relying on spurious correlations. We also verify this comparison in Fig. 6.21. We find that the model with auxiliary constraints predicts counterfactuals that more closely resemble the true counterfactual. Visual inspection of multiple predicted counterfactuals in Fig. 6.23 shows that the learned model better disentangles the effect of thickness t and intensity i as interventions on intensity have a bigger visual effect than before and the counterfactuals look more similar to the ones predicted by the model trained on the original dataset as shown in Fig. 6.11. However, the counterfactuals in low density regions (high thickness and low intensity) contain image artefacts.

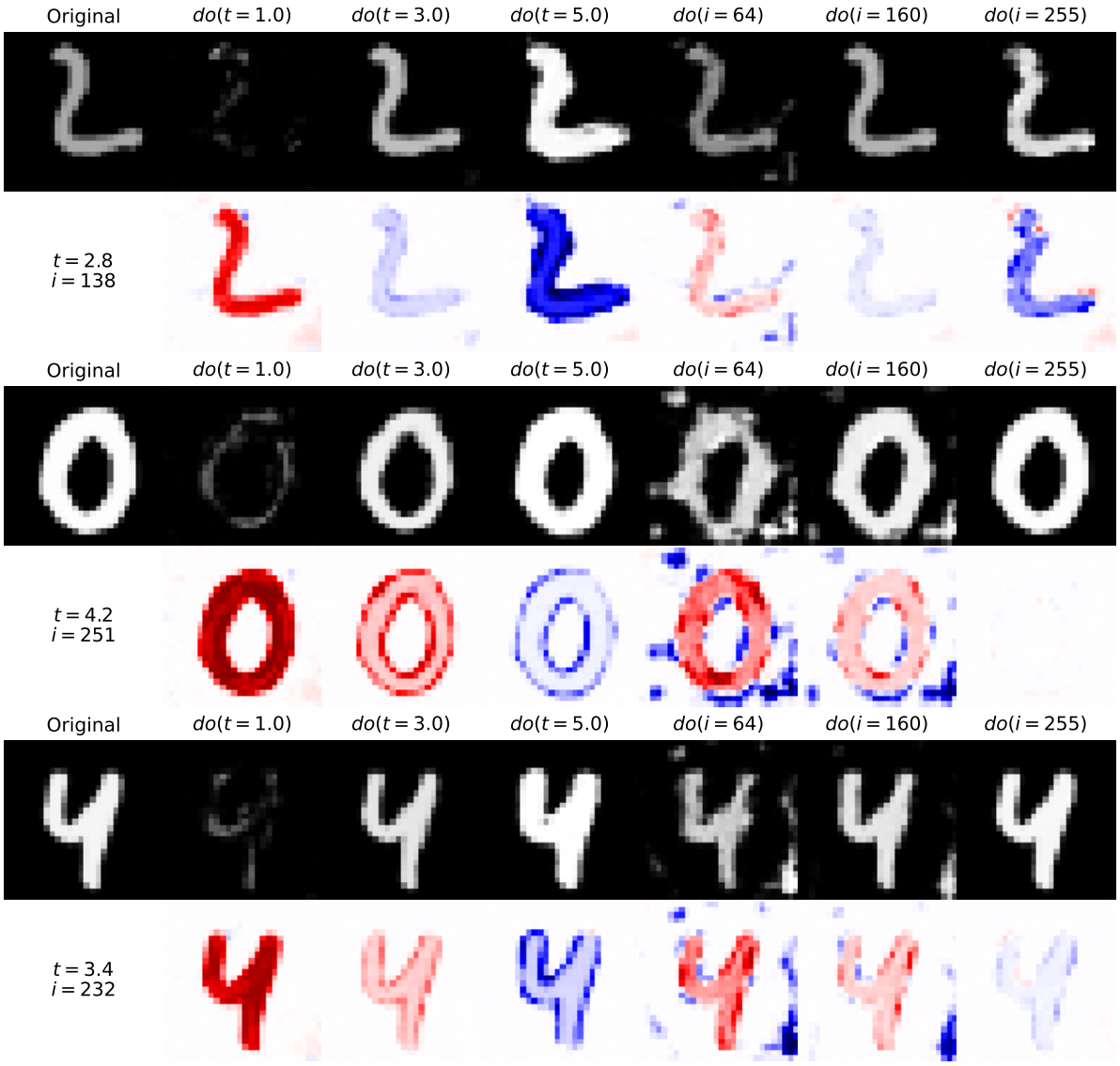


Figure 6.22: Original samples and counterfactuals from the full model trained on the dataset generated with $s = 0.001$. The first column shows the original image and true values of the non-imaging data. The even rows show the difference maps between the original image and the corresponding counterfactual image. We observe that all counterfactuals preserve the digits' identity and style. The model is still capable of generating sensible counterfactuals for interventions on thickness, where it changed both thickness and intensity. However, the model does not predict accurate counterfactuals for interventions on intensity, where it only predicts small visual changes compared to the performed interventions. The model trained on the original dataset was able to reliably predict all these counterfactuals as shown in Fig. 6.11.

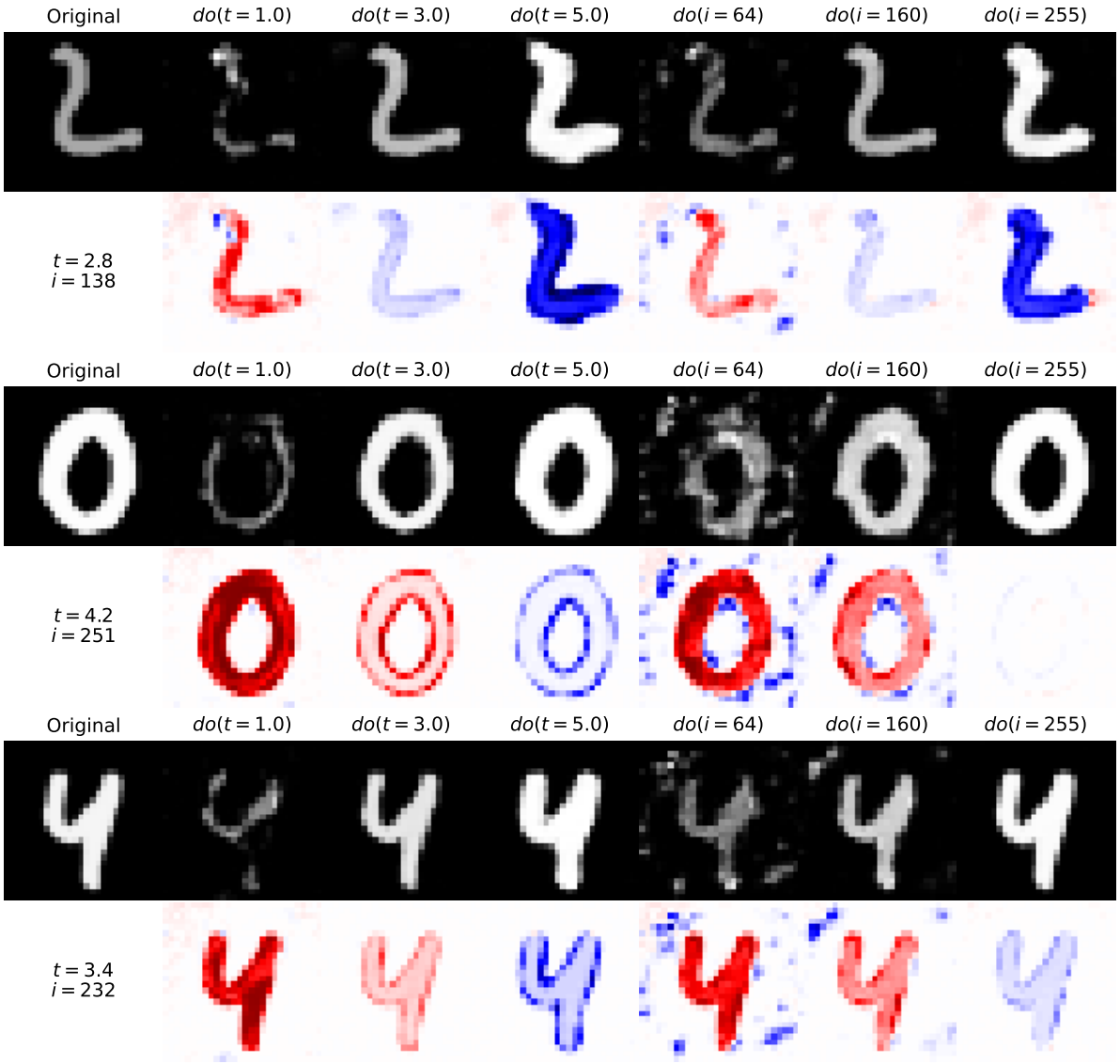


Figure 6.23: Original samples and counterfactuals from the full model with auxiliary constraints trained on the dataset generated with $s = 0.001$. The first column shows the original image and true values of the non-imaging data. The even rows show the difference maps between the original image and the corresponding counterfactual image. We observe that all counterfactuals preserve the digits' identity and style. The model is still capable of generating sensible counterfactuals for interventions on thickness, where it changed both thickness and intensity. The addition of the auxiliary constraints enables the model to predict more accurate counterfactuals for interventions on intensity compared to the model without (*c.f.* Fig. 6.22). However, the counterfactuals in low density regions (high thickness and low intensity) contain image artefacts not present in counterfactuals predicted by the model trained on the original dataset as shown in Fig. 6.11.

6.7 Conclusion

We introduce a novel general framework for fitting SCMs with deep mechanisms. Our deep SCM (DSCM) framework fulfils all three rungs of Pearl's causal hierarchy—in particular, it is the first to enable efficient abduction of exogenous noise, permitting principled counterfactual inference. We demonstrate the potential of DSCMs with three case studies: a synthetic task of modelling Morpho-MNIST digits with a known causal structure, an extension to a more complicated relation between the variables, and a real-world example with brain MRI.

The ability to correctly generate plausible counterfactuals could greatly benefit a wide variety of possible applications, e.g.: *explainability*, where differences between observed and counterfactual data can suggest causal explanations of outcomes; *data augmentation*, as counterfactuals can extrapolate beyond the range of observed data (e.g. novel combinations of attributes); and *domain adaptation*, since including the source of the data as an indicator variable in the causal model could enable generating counterfactual examples in a relevant target domain.

The proposed method does not come without limitations to be investigated in future work. Like the related approaches, the current setup precludes unobserved confounding and requires all variables to be observed when training and computing a counterfactual, which may limit its applicability in certain scenarios. This could be alleviated by imputing the missing data via MCMC or learning auxiliary distributions. Further work should study more closely the training dynamics of deep mechanisms in SCMs: We made some first observations of the neural networks not learning to cleanly disentangle the roles of its inputs on the output as expected. While we proposed a simple counterfactual regularisation technique by using auxiliary distributions, it might require further work. Potential directions could include approaches similar to losses used in image-to-image translation (Xia et al. 2019) and explainability (Singla et al. 2020).

The use of such flexible models also raises questions about the identifiability of the 'true' mechanism, as counterfactuals may not be uniquely defined. For real datasets, counterfactual evaluation is possible only in very constrained settings, as generally true counterfactuals can never be observed. For example, assuming our brain graph in Section 6.5 is correct, we may use a second MRI scan of a brain a few years later as an approximate counterfactual on age. Lastly, it would be interesting to examine whether this framework can be applied to causal discovery, attempting to uncover plausible causal structures from data.

Chapter 7

Conclusions

This chapter revisits the research goals set out in this thesis and takes a look at the contributions made in each chapter and their impact on the wider research community. I then discuss limitations and open questions of the presented work that could lead to future research.

7.1 Summary of Contributions

Let us recall the research goals set out for my PhD research from Section 1.2:

1. **To train neural networks that ignore spurious correlations**
2. **To teach machines to know when they do not know**
3. **To enable neural networks to leverage causal relations**

This section summarises the main contributions and conclusions of each chapter and frames them in the context of the set out research aims. If relevant, I also point out the impact some of these contributions have already had since their publication.

Chapter 3 explored the behaviour of neural networks with constrained receptive fields. The smaller receptive field means that the models are unable to model long range interactions within the input images and therefore less likely to pick up on long range spurious correlations. In a study on the proposed needle MNIST dataset we showed that the informed design of the neural network architecture can support the learning of the desired predictor and ignore present spurious correlations (**Goal 1**). This work further inspired later approaches of solving prediction tasks of small regions of interest in huge images, such as (Kong and Henao 2021). We then used the same approach to study whether long-range interactions are necessary for medical tasks such as biological sex and

age prediction from brain MRI. Our experiments showed that long-range informations are not necessary and texture information is enough for this task. Several works used this approach to further study texture information for segmentation (Fetit et al. 2020) or investigate the interpretability of these patch-based approaches (Bintsi et al. 2020).

Chapter 4 tackles the question of how to teach machines when they do not know something (**Goal 2**). To this end we proposed *Bayes by Hypernet* (BbH), a Bayesian deep learning technique to approximate the posterior distribution of neural network weights using another weight-generating network. The experiments on MNIST and CIFAR confirm that this approach is capable of producing competitive predictive performances while offering relevant uncertainty estimates. However, the complex engineering and complicated hyper-parameters required to successfully apply this method hindered its adoption. Nevertheless, it has been further studied in an extensive comparison of various uncertainty estimation methods (Yao et al. 2019). It also inspired work on using hypernetworks for continual learning (Oswald et al. 2019). Nevertheless, it is unclear whether highly complex variational distributions are even necessary for successful Bayesian deep learning (Farquhar et al. 2020).

We extend the work on "knowing what one doesn't know" in **Chapter 5**. Here, we have framed various medically relevant tasks in the light of outlier or novelty detection. First, we introduce the combination of generative modelling with Bayesian deep learning to capture the model uncertainty. We show in experiments on real world data that this model is capable of detecting lesions in CT images. However, non-deep learning baselines are still very competitive and hard to beat. Building on this introductory work, we applied the same VAE-based generative modelling framework to the task of lesion detection in brain MRI. Here, we found that the domain differences due to MR scanning physics prevent this method from outperforming classical baselines. Lastly, we turned to another type of generative model, normalising flows, and explored the task of classifying tumorous tissue from histopathology images. We confirmed that regular likelihood-based outlier detection metrics perform sub-optimally. Instead, we proposed the use of a Bayesian approach to estimate an outlier metric based on the variance of the predicted likelihood. This metric outperforms classical approaches and is competitive to a fully supervised model.

Even though our work on VAE-based outlier detection hit roadblocks in the form of domain shifts, it inspired a line of follow-up work using VAEs for outlier detection in medical images. We were involved in a range of collaborations that extended the basic VAE model using a locally Gaussian approximation (Chen et al. 2019b) and applying iterative optimisation to the reconstruction (Chen et al. 2021). VAE-based outlier detection has been applied to retinal images (Zhou et al. 2020), lung CT (Uzunova et al. 2019) and various brain MRI tasks (Baur et al. 2020; Zimmerer et al. 2019).

Finally, we take a stab at enabling neural networks to leverage causal relations (**Goal 3**) in Chapter 6. We propose deep structural causal models (DSCMs) as a framework combining deep generative models with concepts of causality to build deep learning models capable of all three rungs of Pearl's

ladder of causation (Pearl 2019). DSCMs naturally extend regular structural causal models from mostly low-dimensional data to high-dimensional data through the use of deep learning components. With sufficient prior knowledge or domain expertise, DSCM enable the design and learning of causal models of the data generating process capable of answering associative, interventional and counterfactual queries. The careful choice of deep learning components used allows users to follow the theoretically grounded three step procedure of "abduction, action and prediction" to predict counterfactuals. Through extensive experimentation, we validated that this framework is able to output plausible counterfactual images to arbitrary interventions while preserving details pertaining to the identity of the object being represented. The last case study also touches upon the aim of training neural networks that ignore spurious correlations by implementing auxiliary distributions to counterfactually constrain the learned functional mechanism.

Since its publication this work has raised awareness of causal methodology amongst medical imaging researchers and and was further extended. Even though our experiments did not explicitly validate the capabilities of the proposed adversarially learned mechanism, it has since been implemented by Dash and Sharma (2020). Work by Garrido et al. (2021) proposes a similar method to our DSCM framework and uses neural autoregressive models as generative models. Reinhold et al. (2021) adapted our published code to the task of modelling the MR images of multiple sclerosis patients. Similarly, Wang et al. (2021) applied the DSCM framework to the task of data harmonization of data derived from brain MRI.

7.2 Limitations and Future Research

This final part of this thesis discusses various limitations of the presented work and makes an attempt at suggesting future research directions.

7.2.1 Actionable Uncertainty Estimates

Various research works have proposed different ways of estimating the uncertainty of predictions from deep learning models. Bayesian approaches, including the proposed *Bayes by Hypernet*, claim to be theoretically grounded in Bayesian modelling. Other directions include solutions following frequentist beliefs (Lakshminarayanan et al. 2017) or aim to post-hoc correct the predicted uncertainty estimates (Guo et al. 2017). However, most of those works are positioned without human interaction in mind. What actions should be taken in cases of uncertain predictions? Are human experts actually more correct than the model in those cases?

Uncertainty estimates for deep learning methods are as hard to interpret as the model predictions themselves. Answering why is the model uncertain is as important as whether it is uncertain. Kendall and Gal (2017) divided uncertainty estimation into aleatoric and epistemic uncertainty – uncertainty relating to ambiguity or noise in the observations and model uncertainty due to limited

training data. Actions relating to an ambiguous observation might include acquiring another sample of that observation. However, if the model is uncertain because it has not seen enough similar training examples, this approach would be flawed. Malinin and Gales (2018) extended this separation and added the term of distributional uncertainty which is caused by a shift between training and test distributions. Even though each of these categories implies a different strategy of dealing with it, none take the human aspect into account.

Similarly, a growing area of research is developing models with reject options (Herbei and Wegkamp 2006). These models are designed and trained to have an additional option: to reject to make a prediction. This direction is orthogonal to uncertainty estimation as it makes a separate prediction as to whether the model is likely to be wrong or not. However, first methods again considered the model independently of potential deployment scenarios. Only recently, research has started to account for human expertise and shifted from a "reject option" to the study of a "deferral option" (Bansal et al. 2021; Mozannar and Sontag 2020).

Nevertheless, almost all of this research is confined to synthetic setups or cleanly curated research datasets. Human operators of machine learning systems are not stationary systems and their mental model of a machine's capabilities changes over time. It is imaginable that a human expert that is supported by an AI system might overwrite the prediction of the machine learning model because they have learned that the model has difficulties with certain properties of the current observation. As such it is necessary to design experiments and reader studies that investigate the effect of uncertainty estimation techniques and deferral options on human trust in AI systems. Additionally, real-world datasets should not only contain the target prediction but also a measure of human uncertainty to allow for the development of AI systems built for human-AI cooperation.

7.2.2 Likelihood as an Outlier Detection Measure

When we are learning a generative model of some data we hope to recover the data distribution as closely as possible. Intuitively, modelling "normal" data should lead to a normative distribution which allows us to detect abnormal samples as having low likelihood under that model. However, how do we distinguish between observations that have low likelihood because they are simply rare events and observations that are abnormal?

Various studies have shown that the likelihood is a flawed quantity to use when distinguishing between "normal" and "abnormal" data. First, it was observed that generative models can assign higher likelihoods to previously unseen samples than to training data (Nalisnick et al. 2019b). This has been related to the unintuitive nature of high-dimensional probability distributions and a solution has been introduced by relying on typicality rather than the likelihood (Nalisnick et al. 2019a). Later the issue of high likelihood for previously unseen data has been attributed to architectural choices in the design of the neural networks and it was shown that a change in inductive biases can alleviate the symptoms (Kirichenko et al. 2020).

However, even if generative models assign higher likelihood to the training distribution than to other regions in the high-dimensional space, they can still not distinguish between rare events and out-of-distribution data. The use of the likelihood alone is in itself a flawed concept (Le Lan and Dinh 2020). The use of derived quantities such as likelihood ratios might be better (Ren et al. 2019) and our experiments in Section 5.3 are in line with these findings. Future research should study the edge cases of distinguishing rare observations from samples of other distributions and derive outlier metrics adequate to this task.

7.2.3 Correctness of Causal Models and Unobserved Variables

Potentially the biggest weaknesses in our work on integrating causal reasoning into neural networks are the requirements that the causal relationships are assumed to be known and that all relevant variables are observed – ie there are no unobserved confounders and there is no missing data. In traditional works on causality, do-calculus identifies conditions under which causal effects are identifiable, and how they can be computed from the data or whether they might be impossible to compute without further assumptions (Pearl 2009). However, most of those assumptions or approaches fall short when having to deal with high-dimensional data.

Instead, our framework relies on the correctness of the causal model and the correct estimation of the parameters corresponding to the functional mechanisms. However, as seen in experiments in Section 6.6 there are no guarantees that the "black-box" deep learning components correctly learn the relationships between the variables even when the causal model is correct. It therefore becomes impossible to truly judge the correctness of the causal graph or the function estimation alone and one can only ever test both of them together. However, the nature of causal queries only ever allows the evaluation of associative and interventional queries – in real-world scenarios it is almost always impossible to evaluate predicted counterfactuals against their true value because we cannot go back in time and intervene on a previous action, or recreate the experiment with the exact same exogenous influences.

In certain settings it is possible to simulate specific interventions in controlled environments and compare the resulting counterfactuals with model predictions. One example of this is the counterfactual evaluation of our DSCM models in Section 6.4 and Section 6.6 where we use our synthetic data generating process to compute synthetic counterfactual for the interventions $\mathrm{do}(\mathrm{thickness} + 2)$ and $\mathrm{do}(\mathrm{intesity}' = 255 - \mathrm{intensity} + 64)$, respectively. Under specific assumptions one could use the same approach to counterfactuals in real world scenarios: Assuming that our causal graph in Section 6.5 is correct, one could collect one or more later brain MR scans of the same subject and interpret them as an approximate counterfactual on age. Often this will, however, not be possible and instead one might need to rely on derived metrics – similar to our use of auxiliary distributions in Section 6.2.5 and Section 6.6, one could imagine to use anti-causal predictive models to evaluate specific properties of the counterfactuals. The combination of this approach with the concept of

double machine learning (Chernozhukov et al. 2018; Jung et al. 2021) might allow the derivation of certain guarantees. However, the naive evaluation using auxiliary models failed due to the auxiliary models' reliance on spurious correlation.

Lastly, the training of the components within the DSCM framework as well as the computation of causal queries requires full observability of all variables. Various techniques have been proposed to deal with unobserved confounders such as the deconfounder (Wang and Blei 2019) or the integration of experimental data into the training process (Ilse et al. 2021) and could be adapted for the use with DSCMs. The problem of missing data during causal effect estimation could again be tackled with auxiliary models.

7.2.4 Identifiability and Spurious Correlations

We have seen in Section 3.1 and Section 6.6, spurious correlations are a constant problem when training neural networks as the training relies on shortcuts to learn the task at hand. The problem is not visible when investigating the observational distributions but only becomes apparent when evaluating interventional or counterfactual queries – that are inherently harder to evaluate. It therefore is an even more important problem to solve as it is easily hidden behind apparently good performances when modelling the observational distribution.

As there can be multiple – infinitely many – different models that entail the same observational distribution, this also becomes a question of the identifiability of neural networks: Which is the correct function that maps the inputs to the outputs? Various works have started addressing this question and defining conditions in which a neural network becomes identifiable (or identifiable up to a specific transformation). Khemakhem et al. (2020) study neural networks and their identifiability in the context of variational autoencoder and nonlinear ICA and introduce an adaptation of regular VAES, the identifiable VAE (or iVAE). Similarly, Khemakhem et al. (2021) shows that affine and additive autoregressive flows are identifiable. Besserve et al. (2021) studies the extrapolation of generative models and touches upon identifiability when talking about equivalent solution spaces. Further works on identifiability include (Mita et al. 2021; Roeder et al. 2021; Sorrenson et al. 2020; Zhou and Wei 2020). Another approach to tackling the issues in learning could follow recent works on domain generalisation and shortcut removal such as (Makar et al. 2021). Whatever form the solution might take, robust learning of the functional mechanisms in a causal model would improve the estimation of causal queries.

7.2.5 From modelling pixels to modelling objects

Traditional causality theory has been developed for the use in low-dimensional data settings where variables are scalars such as in epidimiology or economics. Many of the recent advances of the field still consider experiments with binary interventions – flipping a switch from on to off or administering a medical treatment or not. In the setting of high-dimensional variables, the concept of

causal modelling becomes more complicated. Imagine a natural image with multiple independent objects that are visible, e.g. a picture of a road with multiple cars and pedestrians. How would one formulate interventions of the form "Change the age of this single pedestrian from 20 to 30.", "Change the colour of the sunglasses on this pedestrian from red to blue", or "Change the time the picture was taken from daytime to nighttime."? All of those interventions interact with different levels of object hierarchies within the image and either require significantly different causal graphs for different interventions and images, or a very complex and flexible one that allows for different numbers of observed objects.

The question of how to model object hierarchies in images or other high-dimensional data is still an open one. Different approaches have been proposed to tackle questions such as perceptual grouping (Greff et al. 2016). Other approaches such as capsule networks (Hinton et al. 2018) are designed to automatically discover object-part hierarchies through a special neural network architecture. More and more potential solutions are proposed as idea papers (Hinton 2021) but none of the proposed methods have found widespread adoption and seem to solve the problem. Closely related are works on neuro-symbolic reasoning that aim to extract symbolic representations of objects contained within the image and perform reasoning on the relations between them (De Raedt et al. 2020). Neuro-symbolic approaches have been applied to different tasks such as visual question answering (Amizadeh et al. 2020) or object detection (Manigrasso et al. 2021).

I believe that the modelling of object-part hierarchies as well as the relation between objects can provide better interpretability of predictions made by those algorithms and enable more complex modelling of the world around us. Symbolic reasoning could provide safety guarantees while the neural components allow for modelling of high-dimensional data. Previous work such as (Garnelo et al. 2016) has shown improved interpretability on a toy task and (Sun et al. 2020) apply neuro-symbolic approaches to the problem of self-driving cars.

7.2.6 Towards fairer AI systems

Lastly, I want to touch upon the topic of fairness in machine learning. Fairness has been a growing area of interest within the wider machine learning and beyond. It has been shown that data-driven algorithms learn to reinforce biases present in the datasets they have been trained on (Zhao et al. 2017). Various research efforts are being devoted to developing fair machine learning algorithms to mitigate problems relating to fairness (Madras et al. 2018) but also analyse existing systems (Bird et al. 2020). Causality is a powerful tool in the analysis of the fairness of systems through the introduction of novel notions of fairness (Chiappa 2019; Kusner et al. 2017). Furthermore, robust learning of causal models means that they can transparently communicate which variables affect a certain prediction and allow for the building of fair systems by design. As such, it is clear that the application of causal principles to machine learning helps to build AI systems that are performant but also safe and reliable as well as fair and interpretable for everyone involved.

References

Abadi, M. et al. (2016). "TensorFlow: a system for large-scale machine learning". In: *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*. USENIX Association, pp. 265–283.

Alfaro-Almagro, F. et al. (2018). "Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank". In: *NeuroImage* 166, pp. 400–424. DOI: 10.1016/j.neuroimage.2017.10.034.

Almahairi, A., Ballas, N., Cooijmans, T., Zheng, Y., Larochelle, H., and Courville, A. C. (2016). "Dynamic Capacity Networks". In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by Balcan, M. and Weinberger, K. Q. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 2549–2558.

Amizadeh, S., Palangi, H., Polozov, A., Huang, Y., and Koishida, K. (2020). "Neuro-Symbolic Visual Reasoning: Disentangling "Visual" from "Reasoning"". In: *International Conference on Machine Learning*. PMLR, pp. 279–290.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). "Concrete problems in AI safety". In: *arXiv preprint arXiv:1606.06565*.

An, J. and Cho, S. (2015). "Variational autoencoder based anomaly detection using reconstruction probability". In: *SNU Data Mining Center, Tech. Rep.*

Antipov, G., Baccouche, M., and Dugelay, J.-L. (2017). "Face aging with conditional generative adversarial networks". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 2089–2093. DOI: 10.1109/ICIP.2017.8296650.

Antoniou, A., Pawlowski, N., Turner, J., Owers, J., Mellor, J., and Crowley, E. J. (2019). "Meta-meta-learning for Neural Architecture Search through arXiv Descent". In: *Proceedings of the 2019 ACH Special Interest Group on Harry Queue Bovik (SIGBOVIK)*. Association for Computational Heresy.

Aresta, G. et al. (2018). "BACH: Grand Challenge on Breast Cancer Histology Images". In: *CoRR* abs/1808.04277. arXiv: 1808.04277.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR, pp. 214–223.

Arpit, D. et al. (2017). "A Closer Look at Memorization in Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Precup, D. and Teh, Y. W. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 233–242.

Ayachi, R. and Amor, N. B. (2009). "Brain tumor segmentation using support vector machines". In: *European conference on symbolic and quantitative approaches to reasoning and uncertainty*. Springer, pp. 736–747.

Ba, J., Mnih, V., and Kavukcuoglu, K. (2015). "Multiple Object Recognition with Visual Attention". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Bengio, Y. and LeCun, Y.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., and Davatzikos, C. (2017). "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific data* 4, p. 170117.

Bakas, S. et al. (2018). "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge". In: *arXiv preprint arXiv:1811.02629*.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. (2021). "Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 13, pp. 11405–11414.

Bauer, S., Fejes, T., Slotboom, J., Wiest, R., Nolte, L.-P., and Reyes, M. (2012). "Segmentation of brain tumor images based on integrated hierarchical classification and regularization". In: *MICCAI-BraTS Workshop*.

Baur, C., Wiestler, B., Albarqouni, S., and Navab, N. (2018). "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images". In: *International MICCAI Brainlesion Workshop*. Springer, pp. 161–169.

– (2020). "Bayesian skip-autoencoders for unsupervised hyperintense anomaly detection in high resolution brain MRI". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1905–1909.

Becker, B. G., Klein, T., Wachinger, C., Initiative, A. D. N., et al. (2018). "Gaussian process uncertainty in age estimation as a measure of brain abnormality". In: *NeuroImage* 175, pp. 246–258.

Bejnordi, B. E. et al. (2017). "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer". In: *JAMA* 318.22, pp. 2199–2210.

Bengio, Y., Deleu, T., Rahaman, N., Ke, N. R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. J. (2020). "A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Besserve, M., Sun, R., Janzing, D., and Schölkopf, B. (2021). "A Theory of Independent Mechanisms for Extrapolation in Generative Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8, pp. 6741–6749.

Bingham, E. et al. (2019). "Pyro: Deep universal probabilistic programming". In: *Journal of Machine Learning Research* 20.28.

Bintsi, K.-M., Baltatzis, V., Kolbeinsson, A., Hammers, A., and Rueckert, D. (2020). "Patch-based brain age estimation from MR images". In: *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*. Springer, pp. 98–107.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). "Fairlearn: A toolkit for assessing and improving fairness in AI". In: *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.

Blum, A., Haghtalab, N., and Procaccia, A. D. (2015). "Variational Dropout and the Local Reparameterization Trick". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Ed. by Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., pp. 2575–2583.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). "Weight uncertainty in neural networks". In: *arXiv preprint arXiv:1505.05424*.

Bocklisch, T., Faulker, J., Pawlowski, N., and Nichol, A. (2017). "Rasa: Open source language understanding and dialogue management". In: *arXiv preprint arXiv:1712.05181*.

Borovec, J., Munoz-Barrutia, A., and Kybic, J. (2018). "Benchmarking of Image Registration Methods for Differently Stained Histological Slides". In: DOI: 10.1109/icip.2018.8451040.

Bousquet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C.-J., and Schoelkopf, B. (2017). "From optimal transport to generative modeling: the VEGAN cookbook". In: *arXiv preprint arXiv:1705.07642*.

Brendel, W. and Bethge, M. (2019). "Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Brock, A., Lim, T., Ritchie, J. M., and Weston, N. (2018). "SMASH: One-Shot Model Architecture Search through HyperNetworks". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). "Importance weighted autoencoders". In: *arXiv preprint arXiv:1509.00519*.

Bándi, P. et al. (2019). "From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge". In: *IEEE Transactions on Medical Imaging* 38.2, pp. 550–560. ISSN: 0278-0062. DOI: 10.1109/TMI.2018.2867350.

Carrera, D., Boracchi, G., Foi, A., and Wohlberg, B. (2015). "Detecting anomalous structures by convolutional sparse models". In: *IJCNN*.

Castro, D. C., Tan, J., Kainz, B., Konukoglu, E., and Glocker, B. (2019). "Morpho-MNIST: quantitative assessment and diagnostics for representation learning". In: *Journal of Machine Learning Research* 20.178, pp. 1–29.

Chalapathy, R., Menon, A. K., and Chawla, S. (2017). "Robust, deep and inductive anomaly detection". In: *ECML*.

Charakorn, R., Thawornwattana, Y., Itthipuripat, S., Pawlowski, N., Manoonpong, P., and Dilokthanakul, N. (2020). "An explicit local and global representation disentanglement framework with applications in deep clustering and unsupervised object detection". In: *arXiv preprint arXiv:2001.08957*.

Chen, R. T., Li, X., Grosse, R., and Duvenaud, D. (2018a). "Isolating sources of disentanglement in variational autoencoders". In: *arXiv preprint arXiv:1802.04942*.

Chen, T. Q., Behrmann, J., Duvenaud, D., and Jacobsen, J. (2019a). "Residual Flows for Invertible Generative Modeling". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., pp. 9913–9923.

Chen, T., Fox, E. B., and Guestrin, C. (2014). "Stochastic Gradient Hamiltonian Monte Carlo". In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1683–1691.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Ed. by Lee, D. D., Sugiyama, M., Luxburg, U. von, Guyon, I., and Garnett, R., pp. 2172–2180.

Chen, X. and Konukoglu, E. (2018). "Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders". In: *Medical Imaging with Deep Learning*.

Chen, X., Pawlowski, N., Glocker, B., and Konukoglu, E. (2019b). "Unsupervised Lesion Detection with Locally Gaussian Approximation". In: *International Workshop on Machine Learning in Medical Imaging*.

– (2021). "Normative ascent with local gaussians for unsupervised lesion detection". In: *Medical Image Analysis*.

Chen, X., Pawlowski, N., Rajchl, M., Glocker, B., and Konukoglu, E. (2018b). "Deep generative models in the real-world: An open challenge from medical imaging". In: *arXiv preprint arXiv:1806.05452*.

Chen, Y., Li, W., Sakaridis, C., Dai, D., and Van Gool, L. (2018c). "Domain adaptive faster r-cnn for object detection in the wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3339–3348.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters*.

Chiappa, S. (2019). "Path-specific counterfactual fairness". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 7801–7808.

Choi, H, Jang, E, and Alemi, A. (2018). "Waic, but why? generative ensembles for robust anomaly detection". In: *arXiv preprint arXiv:1810.01392*.

Cole, J. H., Marioni, R. E., Harris, S. E., and Deary, I. J. (2019). "Brain age and other bodily 'ages': implications for neuropsychiatry". In: *Molecular psychiatry* 24.2, pp. 266–281.

Dash, S. and Sharma, A. (2020). "Counterfactual Generation and Fairness Evaluation Using Adversarially Learned Inference". In: *arXiv preprint arXiv:2009.08270*. arXiv: 2009.08270.

De Raedt, L., Dumančić, S., Manhaeve, R., and Marra, G. (2020). "From statistical relational to neuro-symbolic artificial intelligence". In: *arXiv preprint arXiv:2003.08316*.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. (2009). "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

Dilokthanakul, N., Kaplanis, C., Pawlowski, N., and Shanahan, M. (2019). "Feature Control as Intrinsic Motivation for Hierarchical Reinforcement Learning". In: *IEEE Transactions on Neural Networks and Learning Systems*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). "Density estimation using Real NVP". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Doersch, C. (2016). "Tutorial on variational autoencoders". In: *arXiv preprint arXiv:1606.05908*.

Donahue, J., Krähenbühl, P., and Darrell, T. (2017). "Adversarial Feature Learning". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dosovitskiy, A. et al. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., and Courville, A. C. (2017). "Adversarially Learned Inference". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019). "Neural Spline Flows". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., pp. 7509–7520.

Ehteshami Bejnordi, B., Veta, M., Diest, P. Johannes van, Ginneken, B. van, Karssemeijer, N., Litjens, G., Laak, J. A. W. M. van der, and CAMELYON16 Consortium, the (2017). "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast CancerMachine Learning Detection of Breast Cancer Lymph Node MetastasesMachine Learning Detection of Breast Cancer Lymph Node Metastases". In: *JAMA* 318.22, pp. 2199–2210. ISSN: 0098-7484. DOI: 10.1001/jama.2017.14585.

Esteva, A. et al. (2021). "Deep learning-enabled medical computer vision". In: *NPJ digital medicine* 4.1, pp. 1–9.

Farquhar, S., Smith, L., and Gal, Y. (2020). "Liberty or Depth: Deep Bayesian Neural Nets Do Not Need Complex Weight Posterior Approximations". In: *Advances in Neural Information Processing*

Systems. Ed. by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. Vol. 33. Curran Associates, Inc., pp. 4346–4357.

Fetit, A. E., Cupitt, J., Kart, T., and Rueckert, D. (2020). "Training deep segmentation networks on texture-encoded input: application to neuroimaging of the developing neonatal brain". In: *Medical Imaging with Deep Learning*. PMLR, pp. 230–240.

Fidora, A. and Sierra, C. (2011). *Ramon Llull, from the Ars Magna to Artificial Intelligence*. Artificial Intelligence Research Institute Barcelona.

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., Group, B. D. C., et al. (2011). "Unbiased average age-appropriate atlases for pediatric studies". In: *Neuroimage* 54.1, pp. 313–327.

Gal, Y. and Ghahramani, Z. (2015). "Bayesian convolutional neural networks with Bernoulli approximate variational inference". In: *arXiv preprint arXiv:1506.02158*.

Garnelo, M., Arulkumaran, K., and Shanahan, M. (2016). "Towards deep symbolic reinforcement learning". In: *arXiv preprint arXiv:1609.05518*.

Garrido, S., Borysov, S., Rich, J., and Pereira, F. (2021). "Estimating causal effects with the neural autoregressive density estimator". In: *Journal of Causal Inference* 9.1, pp. 211–218. DOI: doi:10.1515/jci-2020-0007.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). "Image Style Transfer Using Convolutional Neural Networks". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp. 2414–2423. DOI: 10.1109/CVPR.2016.265.

Gibson, E. et al. (2018). "NiftyNet: a deep-learning platform for medical imaging". In: *Computer methods and programs in biomedicine* 158, pp. 113–122.

Goldsborough, P., Pawlowski, N., Caicedo, J. C., Singh, S., and Carpenter, A. E. (2017). "CytoGAN: Generative Modeling of Cell Images". In: *NIPS Workshop on Machine Learning for Computational Biology*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., pp. 2672–2680.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). "Explaining and Harnessing Adversarial Examples". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Bengio, Y. and LeCun, Y.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., and Sebag, M. (2018). "Learning Functional Causal Models with Generative Neural Networks". In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Ed. by Escalante, H. J., Escalera, S., Guyon, I.,

Baró, X., Güçlütürk, Y., Güçlü, U., and Gerven, M. van. Cham: Springer International Publishing, pp. 39–80. DOI: 10.1007/978-3-319-98131-4_3.

Graves, A. (2011). "Practical Variational Inference for Neural Networks". In: *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*. Ed. by Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q., pp. 2348–2356.

Greenland, S., Pearl, J., and Robins, J. M. (1999). "Causal Diagrams for Epidemiologic Research". In: *Epidemiology* 10.1, pp. 37–48.

Greff, K., Rasmus, A., Berglund, M., Hao, T., Valpola, H., and Schmidhuber, J. (2016). "Tagger: Deep unsupervised perceptual grouping". In: *Advances in Neural Information Processing Systems*, pp. 4484–4492.

Gregor, K., Danihelka, I., Graves, A., Rezende, D., and Wierstra, D. (2015). "Draw: A recurrent neural network for image generation". In: *International Conference on Machine Learning*. PMLR, pp. 1462–1471.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems*. Ed. by Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. Vol. 30. Curran Associates, Inc.

Gulshan, V. et al. (2016). "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". In: *Jama* 316.22, pp. 2402–2410.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Precup, D. and Teh, Y. W. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1321–1330.

Ha, D., Dai, A. M., and Le, Q. V. (2017). "HyperNetworks". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Haenlein, M. and Kaplan, A. (2019). "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence". In: *California management review* 61.4, pp. 5–14.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. (2018). "Co-teaching: Robust training of deep neural networks with extremely noisy labels". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Ed. by Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., pp. 8536–8546.

Handford, M. (1987). *Where's Wally?* Walker.

Hartford, J. S., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). "Deep IV: A Flexible Approach for Counterfactual Prediction". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Precup, D. and Teh, Y. W. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1414–1423.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

– (2016b). "Identity mappings in deep residual networks". In: *European conference on computer vision*. Springer, pp. 630–645.

Henning, C., Oswald, J. von, Sacramento, J., Surace, S. C., Pfister, J.-P., and Grewe, B. F. (2018). "Approximating the Predictive Distribution via Adversarially-Trained Hypernetworks". In: *Bayesian Deep Learning Workshop, NeurIPS (Spotlight) 2018*.

Herbei, R. and Wegkamp, M. H. (2006). "Classification with reject option". In: *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pp. 709–721.

Hernández-Lobato, J. M. and Adams, R. P. (2015). "Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Bach, F. R. and Blei, D. M. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1861–1869.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Hinton, G. (2021). "How to represent part-whole hierarchies in a neural network". In: *arXiv preprint arXiv:2102.12627*.

Hinton, G. E., Sabour, S., and Frosst, N. (2018). "Matrix capsules with EM routing". In: *International conference on learning representations*.

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-Excitation Networks". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. (2017a). "Snapshot Ensembles: Train 1, Get M for Free". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Huang, G., Liu, Z., Maaten, L. van der, and Weinberger, K. Q. (2017b). "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

Huszár, F. (2017). "Variational Inference using Implicit Distributions". In: *arXiv preprint arXiv:1702.08235*.

Ilse, M., Forré, P., Welling, M., and Mooij, J. M. (2021). "Efficient Causal Inference from Combined Observational and Interventional Data through Causal Reductions". In: *arXiv preprint arXiv:2103.04786*.

Ilse, M., Tomczak, J. M., and Forré, P. (2020). "Designing data augmentation for simulating interventions". In: *arXiv e-prints*, arXiv–2005.

Ilse, M., Tomczak, J. M., and Welling, M. (2018). "Attention-based Deep Multiple Instance Learning". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stock-*

holmsmässan, Stockholm, Sweden, July 10-15, 2018. Ed. by Dy, J. G. and Krause, A. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 2132–2141.

Ioffe, S. and Szegedy, C. (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Bach, F. R. and Blei, D. M. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 448–456.

Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2017). "Image-to-Image Translation with Conditional Adversarial Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp. 5967–5976. DOI: 10.1109/CVPR.2017.632.

Jang, E., Gu, S., and Poole, B. (2017). "Categorical Reparameterization with Gumbel-Softmax". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jastrzkebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. (2017). "Three factors influencing minima in sgd". In: *arXiv preprint arXiv:1711.04623*.

Jensen, J. L. W. V. (1906). "Sur les fonctions convexes et les inégalités entre les valeurs moyennes". In: *Acta Mathematica* 30.none, pp. 175 –193. DOI: 10.1007/BF02418571.

Jiang, B. (2018). "Approximate Bayesian Computation with Kullback-Leibler Divergence as Data Discrepancy". In: *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*. Ed. by Storkey, A. J. and Pérez-Cruz, F. Vol. 84. Proceedings of Machine Learning Research. PMLR, pp. 1711–1721.

Jiang, L., Zhou, Z., Leung, T., Li, L., and Fei-Fei, L. (2018). "MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Dy, J. G. and Krause, A. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 2309–2318.

Johnson, M. J., Duvenaud, D., Wiltschko, A. B., Adams, R. P., and Datta, S. R. (2016). "Composing graphical models with neural networks for structured representations and fast inference". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Ed. by Lee, D. D., Sugiyama, M., Luxburg, U. von, Guyon, I., and Garnett, R., pp. 2946–2954.

Jung, Y., Tian, J., and Bareinboim, E. (2021). "Estimating identifiable causal effects through double machine learning". In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017a). "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: *Medical image analysis* 36, pp. 61–78.

Kamnitsas, K. et al. (2017b). "Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation". In: *MICCAI Multimodal Brain Tumor Segmentation Challenge 2017*.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *International Conference on Learning Representations*.

Katharopoulos, A. and Fleuret, F. (2019). "Processing Megapixel Images with Deep Attention-Sampling Models". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Chaudhuri, K. and Salakhutdinov, R. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 3282–3291.

Kaushik, D., Hovy, E. H., and Lipton, Z. C. (2020). "Learning The Difference That Makes A Difference With Counterfactually-Augmented Data". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kendall, A. and Gal, Y. (2017). "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *Advances in Neural Information Processing Systems* 30, pp. 5574–5584.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). "Variational autoencoders and nonlinear ica: A unifying framework". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2207–2217.

Khemakhem, I., Monti, R., Leech, R., and Hyvarinen, A. (2021). "Causal autoregressive flows". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3520–3528.

Kingma, D. P. and Ba, J. (2015). "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Bengio, Y. and LeCun, Y.

Kingma, D. P. and Welling, M. (2014). "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Bengio, Y. and LeCun, Y.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). "Improved variational inference with inverse autoregressive flow". In: *Advances in neural information processing systems* 29, pp. 4743–4751.

Kirichenko, P., Izmailov, P., and Wilson, A. G. (2020). "Why Normalizing Flows Fail to Detect Out-of-Distribution Data". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. (2018). "CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Komura, D and Ishikawa, S (2018). "Machine Learning Methods for Histopathological Image Analysis". In: *Computational and Structural Biotechnology Journal* 16, pp. 34–42.

Kong, F. and Henao, R. (2021). "Efficient Classification of Very Large Images with Tiny Objects". In: *arXiv preprint arXiv:2106.02694*.

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images*. Tech. rep.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. Ed. by Bartlett, P. L., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., pp. 1106–1114.

Krueger, D., Huang, C.-W., Islam, R., Turner, R., Lacoste, A., and Courville, A. (2017). "Bayesian Hypernetworks". In: *arXiv preprint arXiv:1710.04759*.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. B. (2015). "Deep Convolutional Inverse Graphics Network". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Ed. by Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., pp. 2539–2547.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2017). "Counterfactual Fairness". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Guyon, I., Luxburg, U. von, Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., pp. 4066–4076.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Guyon, I., Luxburg, U. von, Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., pp. 6402–6413.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). "Unsupervised Machine Translation Using Monolingual Corpora Only". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). "Autoencoding beyond pixels using a learned similarity metric". In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by Balcan, M. and Weinberger, K. Q. Vol. 48. JMLR Workshop and Conference Proceedings. PMLR. JMLR.org, pp. 1558–1566.

Le Lan, C and Dinh, L (2020). "Perfect density models cannot guarantee anomaly detection". In: *"I Can't Believe It's Not Better!"NeurIPS 2020 workshop*.

LeCun, Y. and Bengio, Y. (1998). "The Handbook of Brain Theory and Neural Networks". In: ed. by Arbib, M. A. Cambridge, MA, USA: MIT Press. Chap. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258. ISBN: 0-262-51102-9.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998a). "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791.

LeCun, Y., Cortes, C., and Burges, C. J. (1998b). *The MNIST database of handwritten digits*.

Lee, B. and Paeng, K. (2018). "A Robust and Effective Approach Towards Accurate Metastasis Detection and pN-stage Classification in Breast Cancer". In: *CoRR* abs/1805.12067. arXiv: 1805.12067.

Lee, M. C., Petersen, K., Pawlowski, N., Glocker, B., and Schaap, M. (2019). "Template Transformer Networks for Image Segmentation". In: *Medical Imaging with Deep Learning Abstract track*.

Lee, M. C., Petersen, K., Pawlowski, N., Glocker, B., and Schaap, M. (2019). "TETRIS: Template Transformer Networks for Image Segmentation with Shape Priors". In: *IEEE Transactions on Medical Imaging*.

Li, C., Welling, M., Zhu, J., and Zhang, B. (2018). "Graphical Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Ed. by Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., pp. 6072–6083.

Liew, S.-L. et al. (2018). "A large, open source dataset of stroke anatomical brain images and manual lesion segmentations". In: *Scientific data* 5.1, pp. 1–11.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). "Microsoft COCO: Common objects in context". In: *European conference on computer vision*. Springer, pp. 740–755.

Lin, W., Hubacher, N., and Khan, M. E. (2018). "Variational Message Passing with Structured Inference Networks". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Laak, J. A. van der, Ginneken, B. van, and Sánchez, C. I. (2017). "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42, pp. 60–88.

Liu, Y. et al. (2020). "A deep learning system for differential diagnosis of skin diseases". In: *Nature medicine* 26.6, pp. 900–908.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D. A., Zemel, R. S., and Welling, M. (2017). "Causal Effect Inference with Deep Latent-Variable Models". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Guyon, I., Luxburg, U. von, Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., pp. 6446–6456.

Louizos, C. and Welling, M. (2016). "Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors". In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by Balcan, M. and Weinberger, K. Q. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1708–1716.

– (2017). "Multiplicative Normalizing Flows for Variational Bayesian Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia,*

6-11 August 2017. Ed. by Precup, D. and Teh, Y. W. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 2218–2227.

Lu, X., Perrone, V., Hasenclever, L., Teh, Y. W., and Vollmer, S. J. (2017). "Relativistic Monte Carlo". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*. Ed. by Singh, A. and Zhu, X. J. Vol. 54. Proceedings of Machine Learning Research. PMLR, pp. 1236–1245.

Lu, Y. and Huang, B. (2020). "Structured Output Learning with Conditional Generative Flows". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 5005–5012.

Ma, Y., Chen, T., and Fox, E. B. (2015). "A Complete Recipe for Stochastic Gradient MCMC". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Ed. by Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., pp. 2917–2925.

MacKay, D. J. (1992). "A practical Bayesian framework for backpropagation networks". In: *Neural computation* 4.3, pp. 448–472.

– (1995). "Bayesian neural networks and density networks". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 354.1, pp. 73–80.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Maddison, C. J. and Tarlow, D. (2017). *Gumbel Machinery*.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2019). "A Simple Baseline for Bayesian Uncertainty in Deep Learning". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., pp. 13132–13143.

Madras, D., Pitassi, T., and Zemel, R. (2018). "Predict responsibly: improving fairness and accuracy by learning to defer". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 6150–6160.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and Maaten, L. van der (2018). "Exploring the limits of weakly supervised pretraining". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 181–196.

Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., and D'Amour, A. (2021). "Causally-motivated Shortcut Removal Using Auxiliary Labels". In: *arXiv preprint arXiv:2105.06422*.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). "Adversarial autoencoders". In: *arXiv preprint arXiv:1511.05644*.

Malinin, A. and Gales, M. (2018). "Predictive uncertainty estimation via prior networks". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7047–7058.

Manigrasso, F., Miro, F. D., Morra, L., and Lamberti, F. (2021). "Faster-LTN: a neuro-symbolic, end-to-end object detection architecture". In: *International Conference on Artificial Neural Networks*. Springer, pp. 40–52.

Martinez, Á. P. and Marca, J. V. (2019). "Explaining Visual Models by Causal Attribution". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, pp. 4167–4175.

Mayor, A. (2020). *Gods and robots: myths, machines, and ancient dreams of technology*. Princeton University Press.

McKinney, S. M. et al. (2020). "International evaluation of an AI system for breast cancer screening". In: *Nature* 577.7788, pp. 89–94.

Meng, Q., Pawlowski, N., Rueckert, D., and Kainz, B. (2019). "Representation Disentanglement for Multi-task Learning with application to Fetal Ultrasound". In: *arXiv preprint arXiv:1908.07885*.

Menze, B. H. et al. (2015). "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10, pp. 1993–2024.

Mescheder, L. M., Nowozin, S., and Geiger, A. (2017). "Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Precup, D. and Teh, Y. W. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 2391–2400.

Minka, T. P. (2001). "Expectation propagation for approximate Bayesian inference". In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 362–369.

Mirza, M. and Osindero, S. (2014). "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784*.

Mita, G., Filippone, M., and Michiardi, P. (2021). "An Identifiable Double VAE For Disentangled Representations". In: *International Conference on Machine Learning*. PMLR, pp. 7769–7779.

Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). "Recurrent Models of Visual Attention". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., pp. 2204–2212.

Mnih, V. et al. (2015). "Human-level control through deep reinforcement learning". In: *nature* 518.7540, pp. 529–533.

MONAI Consortium (Mar. 2020). *MONAI: Medical Open Network for AI*. DOI: 10.5281/zenodo.4323058.

Monteiro, M., Folgoc, L. L., Castro, D. C. de, Pawlowski, N., Marques, B., Kamnitsas, K., Wilk, M. van der, and Glocker, B. (2020). "Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncertainty". In: *Advances in Neural Information Processing*.

Mozannar, H. and Sontag, D. (2020). "Consistent estimators for learning to defer to an expert". In: *International Conference on Machine Learning*. PMLR, pp. 7076–7087.

Nalisnick, E., Matsukawa, A., Teh, Y. W., and Lakshminarayanan, B. (2019a). "Detecting out-of-distribution inputs to deep generative models using a test for typicality". In: *arXiv preprint arXiv:1906.02994* 5, p. 5.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. (2019b). "Do Deep Generative Models Know What They Don't Know?" In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Neal, R. M. (1995). "Bayesian Learning for Neural Networks". AAINN02676. PhD thesis. Toronto, Canada. ISBN: 0-612-02676-0.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. (2018). "Sensitivity and Generalization in Neural Networks: an Empirical Study". In: *International Conference on Learning Representations*. OpenReview.net.

Oberst, M. and Sontag, D. A. (2019). "Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Chaudhuri, K. and Salakhutdinov, R. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 4881–4890.

OpenAI et al. (2019). "Dota 2 with Large Scale Deep Reinforcement Learning". In: *arXiv preprint arXiv:1912.06680*.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2015). "Is object localization for free? - Weakly-supervised learning with convolutional neural networks". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, pp. 685–694. DOI: 10.1109/CVPR.2015.7298668.

Osband, I., Blundell, C., Pritzel, A., and Roy, B. V. (2016). "Deep Exploration via Bootstrapped DQN". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Ed. by Lee, D. D., Sugiyama, M., Luxburg, U. von, Guyon, I., and Garnett, R., pp. 4026–4034.

Oswald, J. von, Henning, C., Sacramento, J., and Grewe, B. F. (2019). "Continual learning with hypernetworks". In: *arXiv e-prints*, arXiv–1906.

Paisley, J., Blei, D. M., and Jordan, M. I. (2012). "Variational Bayesian inference with stochastic search". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1363–1370.

Pang, J., Li, C., Shi, J., Xu, Z., and Feng, H. (2019). "$\mathcal{R}^2$-CNN: Fast Tiny Object Detection in Large-scale Remote Sensing Images". In: *arXiv preprint arXiv:1902.06042*.

Papamakarios, G., Murray, I., and Pavlakou, T. (2017). "Masked Autoregressive Flow for Density Estimation". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Guyon, I., Luxburg, U. von, Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., pp. 2338–2347.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2019). "Normalizing Flows for Probabilistic Modeling and Inference". In: *arXiv preprint arXiv:1912.02762*.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., and Schölkopf, B. (2018). "Learning Independent Causal Mechanisms". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Dy, J. G. and Krause, A. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 4033–4041.

Paszke, A. et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., pp. 8024–8035.

Pati, A. and Lerch, A. (2020). "Attribute-based Regularization of VAE Latent Spaces". In: *arXiv preprint arXiv:2004.05485*.

Pawlowski, N., Bhooshan, S., Ballas, N., Ciompi, F., Glocker, B., and Drozdzal, M. (2019). "Needles in Haystacks: On Classifying Tiny Objects in Large Images". In: *arXiv preprint arXiv:1908.06037*.

Pawlowski, N., Brock, A., Lee, M. C., Rajchl, M., and Glocker, B. (2017a). "Implicit Weight Uncertainty in Neural Networks". In: *NeurIPS Workshop on Bayesian Deep Learning*.

Pawlowski, N., Castro, D. C., and Glocker, B. (2020). "Deep Structural Causal Models for Tractable Counterfactual Inference". In: *Advances in Neural Information Processing Systems*.

Pawlowski, N. and Glocker, B. (2019). "Is Texture Predictive for Age and Sex in Brain MRI?" In: *Medical Imaging with Deep Learning Abstract track*.

– (2021). "Abnormality Detection in Histopathology via Density Estimation with Normalising Flows". In: *Medical Imaging with Deep Learning Short Paper Track*.

Pawlowski, N., Jaques, M., and Glocker, B. (2017b). "Efficient variational Bayesian neural network ensembles for outlier detection". In: *ICLR Workshop Track*.

Pawlowski, N., Ktena, S. I., Lee, M. C., Kainz, B., Rueckert, D., Glocker, B., and Rajchl, M. (2017c). "DLTK: State of the Art Reference Implementations for Deep Learning on Medical Images". In: *NIPS Workshop on Medical Imaging meets NIPS*.

Pawlowski, N. et al. (2018). "Unsupervised Lesion Detection in Brain CT using Bayesian Convolutional Autoencoders". In: *Medical Imaging with Deep Learning Abstract track*.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd. Cambridge University Press.

– (2019). "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3, pp. 54–60. DOI: 10.1145/3241036.

Pérez-Cruz, F. (2008). "Kullback-Leibler divergence estimation of continuous distributions". In: *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*. IEEE, pp. 1666–1670.

Pérez-García, F., Sparks, R., and Ourselin, S. (2021). "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning". In: *Computer Methods and Programs in Biomedicine*, p. 106236. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2021.106236.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press.

Peters, R. (2006). "Ageing and the brain". In: *Postgraduate Medical Journal* 82.964, pp. 84–88.

Prastawa, M., Bullitt, E., Ho, S., and Gerig, G. (2004). "A brain tumor segmentation framework based on outlier detection". In: *Medical image analysis* 8.3, pp. 275–283.

Press, G. (2020). *12 Artificial Intelligence (AI) Milestones: 2. Ramon Llull And His 'Thinking Machine'*.

Radford, A., Metz, L., and Chintala, S. (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXiv preprint arXiv:1511.06434*.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. (2018). "Tighter Variational Bounds are Not Necessarily Better". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Dy, J. G. and Krause, A. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 4274–4282.

Rajchl, M., Pawlowski, N., Rueckert, D., Matthews, P. M., and Glocker, B. (2018). "NeuroNet: Fast and Robust Reproduction of Multiple Brain Image Segmentation Pipelines". In: *Medical Imaging with Deep Learning*.

Ratzlaff, N. and Li, F. (2019). "HyperGAN: A Generative Model for Diverse, Performant Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Chaudhuri, K. and Salakhutdinov, R. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 5361–5369.

Reinhold, J. C., Carass, A., and Prince, J. L. (2021). "A Structural Causal Model for MR Images of Multiple Sclerosis". In: *arXiv preprint arXiv:2103.03158*.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., and Lakshminarayanan, B. (2019). "Likelihood Ratios for Out-of-Distribution Detection". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., pp. 14680–14691.

Resta, S., Acito, N., Diani, M., Corsini, G., Opsahl, T., and Haavardsholm, T. V. (2011). "Detection of small changes in airborne hyperspectral imagery: Experimental results over urban areas". In: *2011 6th International Workshop on the Analysis of Multi-temporal Remote Sensing Images (Multi-Temp)*, pp. 5–8. DOI: 10.1109/Multi-Temp.2011.6005033.

Rezende, D. J. and Mohamed, S. (2015). "Variational Inference with Normalizing Flows". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Bach, F. R. and Blei, D. M. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1530–1538.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1278–1286.

Robinson, R. D. (2020). "Reliable Machine Learning for Medical Imaging Data through Automated Quality Control and Data Harmonization". In:

Roeder, G., Metz, L., and Kingma, D. (2021). "On linear identifiability of learned representations". In: *International Conference on Machine Learning*. PMLR, pp. 9030–9039.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.

Rosca, M., Lakshminarayanan, B., Warde-Farley, D., and Mohamed, S. (2017). "Variational approaches for auto-encoding generative adversarial networks". In: *arXiv preprint arXiv:1706.04987*.

Roy, A. G. et al. (2021). "Does Your Dermatology Classifier Know What It Doesn't Know? Detecting the Long-Tail of Unseen Conditions". In: *arXiv preprint arXiv:2104.03829*.

Sato, D., Hanaoka, S., Nomura, Y., Takenaga, T., Miki, S., Yoshikawa, T., Hayashi, N., and Abe, O. (2018). "A primitive study on unsupervised anomaly detection with an autoencoder in emergency head ct volumes". In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol. 10575. International Society for Optics and Photonics, 105751P.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery". In: *IPMI*. Springer, pp. 146–157.

Schmidhuber, J. (2021). *Turing Oversold*. URL: https://people.idsia.ch//~juergen/turing-oversold.html (visited on 09/16/2021).

Schölkopf, B. (2019). "Causality for Machine Learning". In: *arXiv preprint arXiv:1911.10500*.

Setio, A. A. A. et al. (2017). "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge". In: *Medical Image Analysis* 42, pp. 1–13. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.06.015.

Shi, J., Sun, S., and Zhu, J. (2018). "Kernel Implicit Variational Inference". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Silver, D. et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587, pp. 484–489.

Silver, D. et al. (2017). "Mastering the game of go without human knowledge". In: *Nature* 550.7676, p. 354.

Simonyan, K. and Zisserman, A. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Bengio, Y. and LeCun, Y.

Singla, S., Pollack, B., Chen, J., and Batmanghelich, K. (2020). "Explanation by Progressive Exaggeration". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Sohn, K., Lee, H., and Yan, X. (2015). "Learning Structured Output Representation using Deep Conditional Generative Models". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Ed. by Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., pp. 3483–3491.

Sorrenson, P., Rother, C., and Köthe, U. (2020). "Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN)". In: *International Conference on Learning Representations*.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). "Dropout: a simple way to prevent neural networks from overfitting." In: *JMLR* 15.1, pp. 1929–1958.

Subbaswamy, A., Schulam, P., and Saria, S. (2019). "Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport". In: *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*. Ed. by Chaudhuri, K. and Sugiyama, M. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 3118–3127.

Suckling, J. (1994). "The Mammographic Image Analysis Society Digital Mammogram Database"Exerpta Medica". In: *Exerpta Medica International Congress Series* 1069.

Sudlow, C. et al. (2015). "UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS Medicine* 12.3.

Sudre, C. H. et al. (2018). "3D multirater RCNN for multimodal multiclass detection and characterisation of extremely small objects". In: *arXiv preprint arXiv:1812.09046*.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era". In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp. 843–852. DOI: 10.1109/ICCV.2017.97.

Sun, J., Sun, H., Han, T., and Zhou, B. (2020). "Neuro-Symbolic Program Search for Autonomous Driving Decision Module Design". In: *Conference on Robot Learning 2020*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. (2014). "Intriguing properties of neural networks". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Bengio, Y. and LeCun, Y.

Tabak, E. G. and Turner, C. V. (2013). "A Family of Nonparametric Density Estimation Algorithms". In: *Communications on Pure and Applied Mathematics* 66.2, pp. 145–164. DOI: 10.1002/cpa.21423.

Taylor, J. R., Williams, N., Cusack, R., Auer, T., Shafto, M. A., Dixon, M., Tyler, L. K., Henson, R. N., et al. (2017). "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample". In: *Neuroimage* 144, pp. 262–269.

Tieleman, T. and Hinton, G. (2012). *Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude*. COURSERA: Neural Networks for Machine Learning.

Tomas-Fernandez, X. and Warfield, S. K. (2015). "A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation". In: *IEEE transactions on medical imaging* 34.6, pp. 1349–1361.

Tomita, N., Abdollahi, B., Wei, J., Ren, B., Suriawinata, A., and Hassanpour, S. (2018). "Finding a Needle in the Haystack: Attention-Based Classification of High Resolution Microscopy Images". In: *arXiv preprint arXiv:1811.08513*.

Tran, D. and Blei, D. M. (2018). "Implicit Causal Models for Genome-wide Association Studies". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Tran, D., Ranganath, R., and Blei, D. M. (2017). "Deep and Hierarchical Implicit Models". In: *arXiv preprint arXiv:1702.08896*.

Trippe, B. L. and Turner, R. E. (2017). "Conditional density estimation with Bayesian normalising flows". In: *NIPS 2017 Workshop on Bayesian Deep Learning*.

Tunyasuvunakool, K. et al. (2021). "Highly accurate protein structure prediction for the human proteome". In: *Nature* 596.7873, pp. 590–596.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). "Adversarial discriminative domain adaptation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176.

Ukai, K., Matsubara, T., and Uehara, K. (2018). "Hypernetwork-based Implicit Posterior Estimation and Model Averaging of CNN". In: *Proceedings of The 10th Asian Conference on Machine Learning*. Ed. by Zhu, J. and Takeuchi, I. Vol. 95. Proceedings of Machine Learning Research. PMLR, pp. 176–191.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. (2016). "Instance Normalization: The Missing Ingredient for Fast Stylization". In: arXiv: 1607.08022 [cs.CV].

Uzunova, H., Schultz, S., Handels, H., and Ehrhardt, J. (2019). "Unsupervised pathology detection in medical images using conditional variational autoencoders". In: *International journal of computer assisted radiology and surgery* 14.3, pp. 451–461.

Valindria, V. V., Pawlowski, N., Rajchl, M., Lavdas, I., Aboagye, E. O., Rockall, A. G., Rueckert, D., and Glocker, B. (2018). "Multi-Modal Learning from Unpaired Images: Application to Multi-Organ Segmentation in CT and MRI". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., and Suetens, P. (2001). "Automated segmentation of multiple sclerosis lesions by model outlier detection". In: *IEEE transactions on medical imaging* 20.8, pp. 677–688.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. (2018). "Rotation Equivariant CNNs for Digital Pathology". In: arXiv: 1806.03962 [cs.CV].

Wang, R., Chaudhari, P., and Davatzikos, C. (2021). "Harmonization with Flow-based Causal Inference". In: *arXiv preprint arXiv:2106.06845*.

Wang, Y. and Blei, D. M. (2019). "The blessings of multiple causes". In: *Journal of the American Statistical Association* 114.528, pp. 1574–1596.

Wei, J. W., Tafe, L. J., Linnik, Y. A., Vaickus, L. J., Tomita, N., and Hassanpour, S. (2019). "Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks". In: *CoRR* abs/1901.11489. arXiv: 1901.11489.

Welling, M. and Teh, Y. W. (2011). "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In: *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*. Ed. by Getoor, L. and Scheffer, T. Omnipress, pp. 681–688.

Winkler, C., Worrall, D., Hoogeboom, E., and Welling, M. (2019). "Learning Likelihoods with Conditional Normalizing Flows". In: *arXiv preprint arXiv:1912.00042*.

Wold, H. O. A. (1954). "Causality and Econometrics". In: *Econometrica* 22.2, pp. 162–177. DOI: 10.2307/1907540.

World Health Organization (2018). *Cancer*. URL: https://www.who.int/news-room/fact-sheets/detail/cancer (visited on 02/12/2020).

Xia, G., Bai, X., Ding, J., Zhu, Z., Belongie, S. J., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. (2018). "DOTA: A Large-Scale Dataset for Object Detection in Aerial Images". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, pp. 3974–3983. DOI: 10.1109/CVPR.2018.00418.

Xia, T., Chartsias, A., Wang, C., and Tsaftaris, S. A. (2019). "Learning to synthesise the ageing brain without longitudinal data". In: *arXiv preprint arXiv:1912.02620*.

Yan, X., Yang, J., Sohn, K., and Lee, H. (2016). "Attribute2image: Conditional image generation from visual attributes". In: *European Conference on Computer Vision*. Springer, pp. 776–791.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. (2020). "CausalVAE: Structured Causal Disentanglement in Variational Autoencoder". In: *arXiv preprint arXiv:2004.08697*.

Yao, J., Pan, W., Ghosh, S., and Doshi-Velez, F. (2019). "Quality of uncertainty quantification for Bayesian neural network inference". In: *arXiv preprint arXiv:1906.09686*.

Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. (2017). "Semantic image inpainting with deep generative models". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5485–5493.

Zeng, K., Erus, G., Sotiras, A., Shinohara, R. T., and Davatzikos, C. (2016). "Abnormality detection via iterative deformable registration and basis-pursuit decomposition". In: *IEEE transactions on medical imaging* 35.8, pp. 1937–1951.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). "Understanding deep learning requires rethinking generalization". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Zhang, Q.-s. and Zhu, S.-c. (2018). "Visual interpretability for deep learning: a survey". In: *Frontiers of Information Technology & Electronic Engineering* 19.1, pp. 27–39.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2979–2989. DOI: 10.18653/v1/D17-1323.

Zhou, D. and Wei, X.-X. (2020). "Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE". In: *Advances in Neural Information Processing Systems*. Ed. by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. Vol. 33. Curran Associates, Inc., pp. 7234–7247.

Zhou, K., Xiao, Y., Yang, J., Cheng, J., Liu, W., Luo, W., Gu, Z., Liu, J., and Gao, S. (2020). "Encoding structure-texture relation with P-Net for anomaly detection in retinal images". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, pp. 360–377.

Zikic, D et al. (2012). "Context-sensitive Classification Forests for Segmentation of Brain Tumor Tissues". In: *MICCAI-BraTS Workshop*.

Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., and Maier-Hein, K. (2019). "Unsupervised anomaly localization using variational auto-encoders". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 289–297.

Appendices

A Posterior Distributions for Bayes by Hypernet

A.1 Toy Example

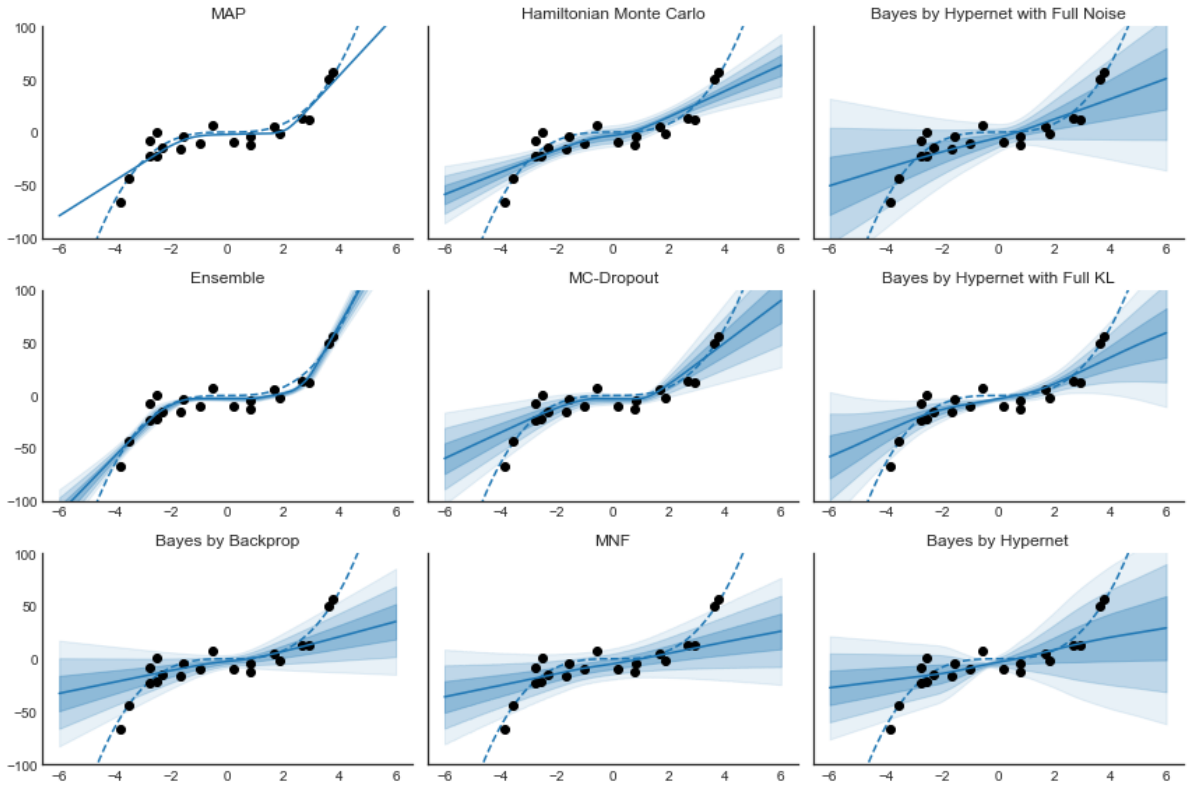
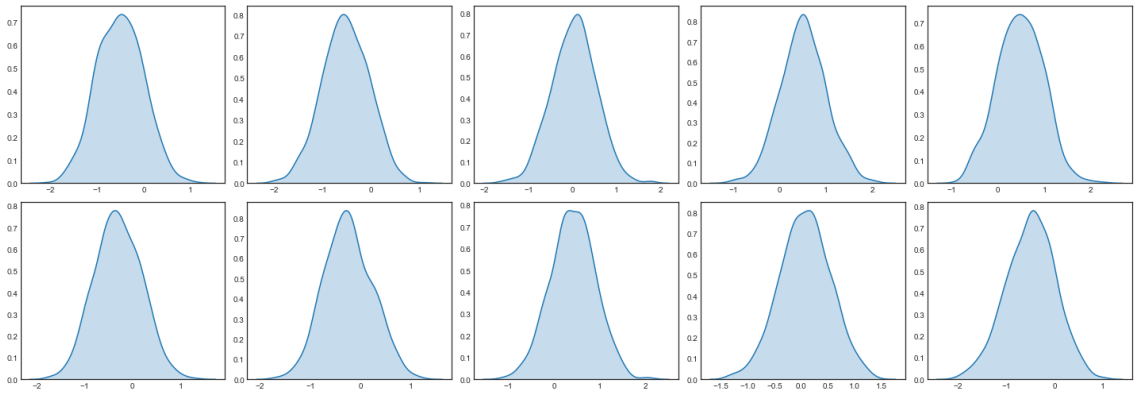
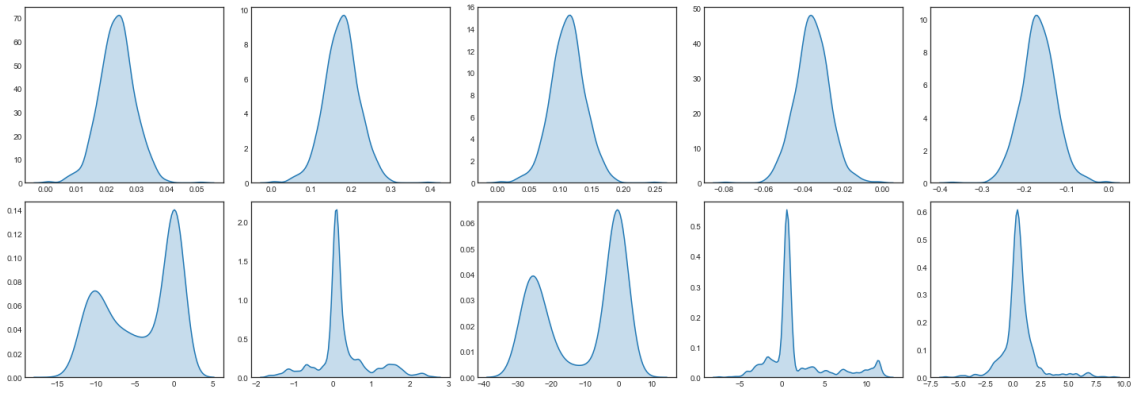


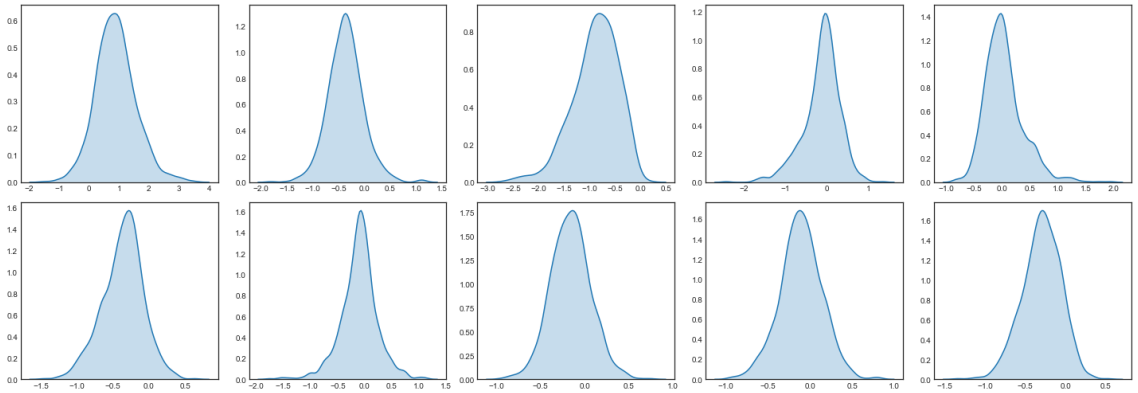
Figure A.1: Fit of a toy cubic function by various Bayesian deep learning methods, regular deep learning methods and HMC. The *BbH* variants refer to the use of an auxiliary noise vector of equivalent dimensionality as the generated weights (BbH with full noise), calculating the KL divergence using the kernel method with the assumption of independent weights (BbH), and calculating the full KL divergence using the kernel method (BbH with full KL).



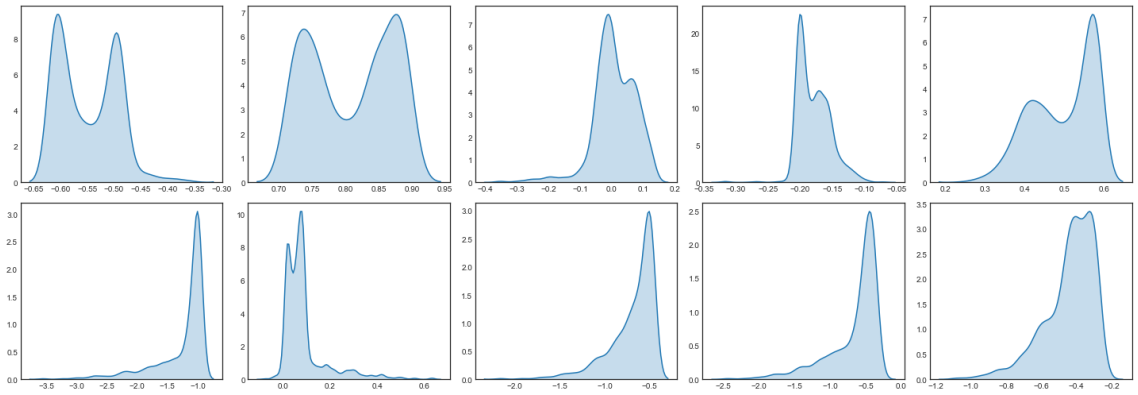
(a) Examples of the posterior weight distributions using BBB.



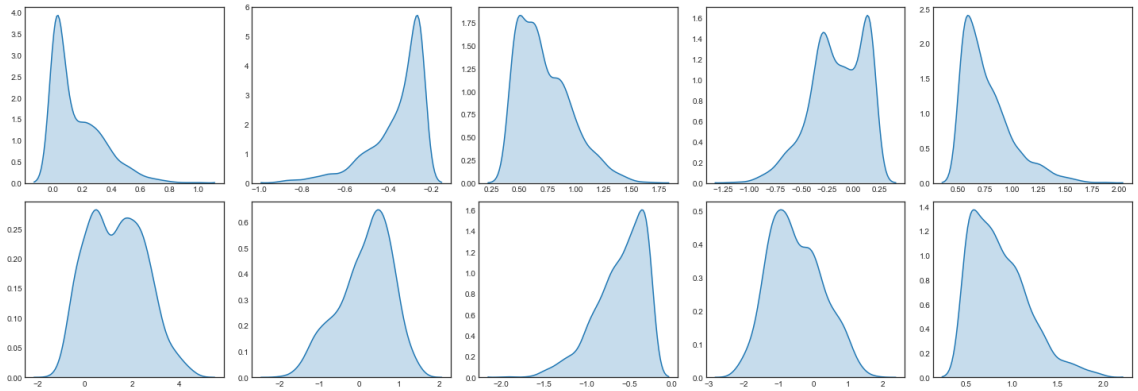
(b) Examples of the posterior weight distributions using MNF.



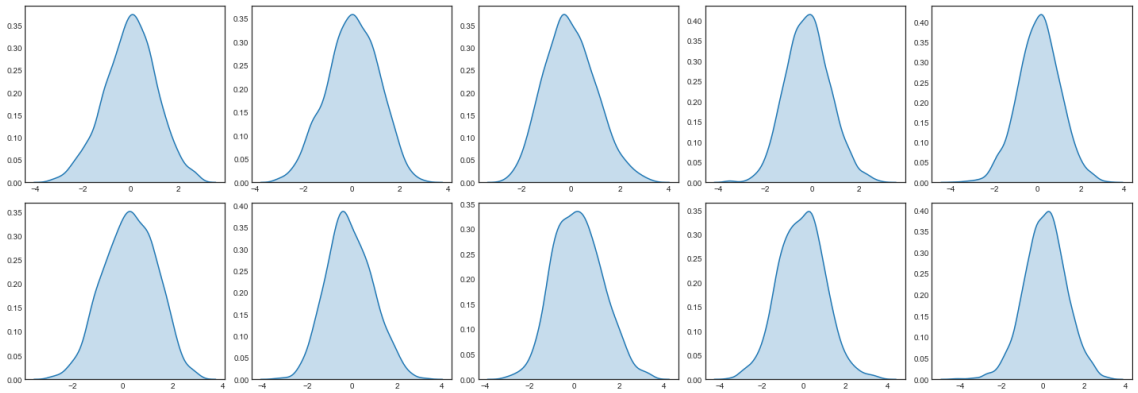
(c) Examples of the posterior weight distributions using *BbH* with an auxiliary noise vector of equivalent dimensionality as the generated weights.



(d) Examples of the posterior weight distributions using *BbH* calculating the full KL divergence using the kernel method.

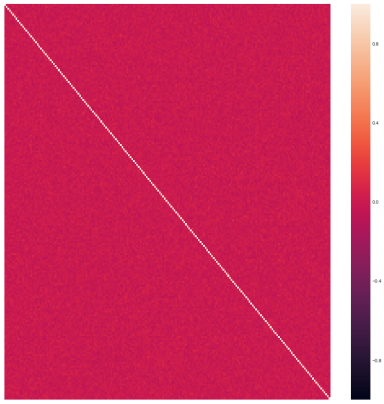


(e) Examples of the posterior weight distributions using *BbH* calculating the KL divergence using the kernel method and the assumption of independent weights.

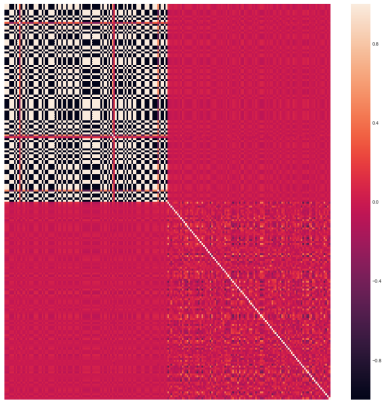


(f) Examples of the posterior weight distributions using HMC.

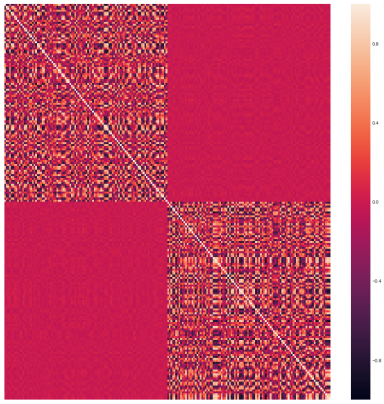
Figure A.2: Illustration of the posterior distributions of the first weights of a fully connected network trained on a toy regression task approximated by BBB (a), MNF (b), various settings of BbH (c-e) and HMC (f).



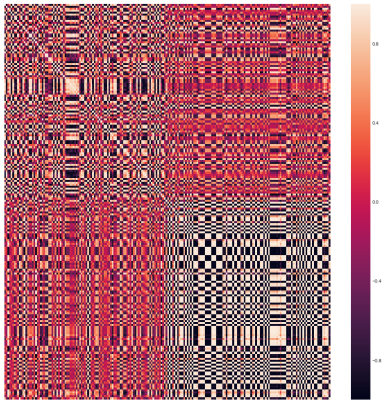
(a) Examples of the correlations within the posterior weight distributions using BBB.



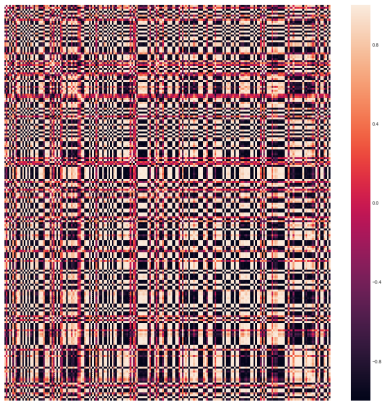
(b) Examples of the correlations within the posterior weight distributions using MNF.



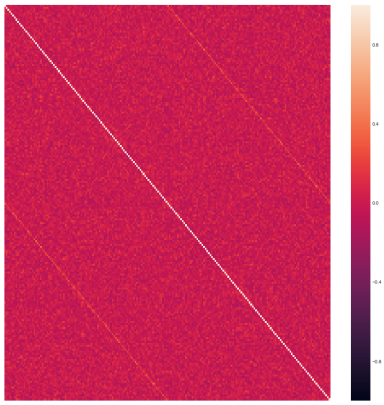
(c) Examples of the posterior correlations within the weight distributions using *BbH* with an auxiliary noise vector of equivalent dimensionality as the generated weights.



(d) Examples of the correlations within the posterior weight distributions using *BbH* calculating the full KL divergence using the kernel method.



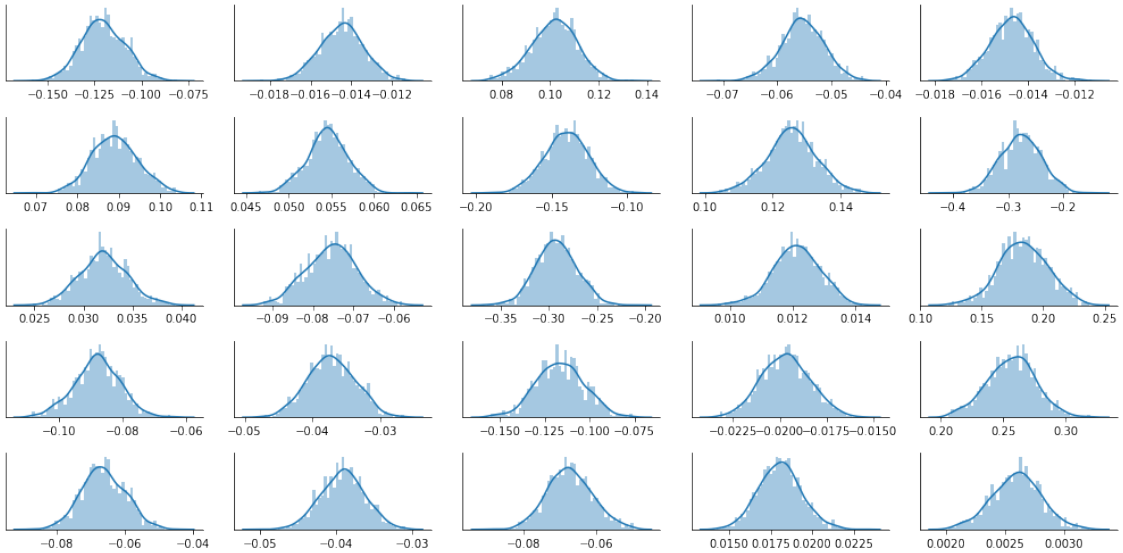
(e) Examples of the correlations within the posterior weight distributions using *BbH* calculating the KL divergence using the kernel method and the assumption of independent weights.



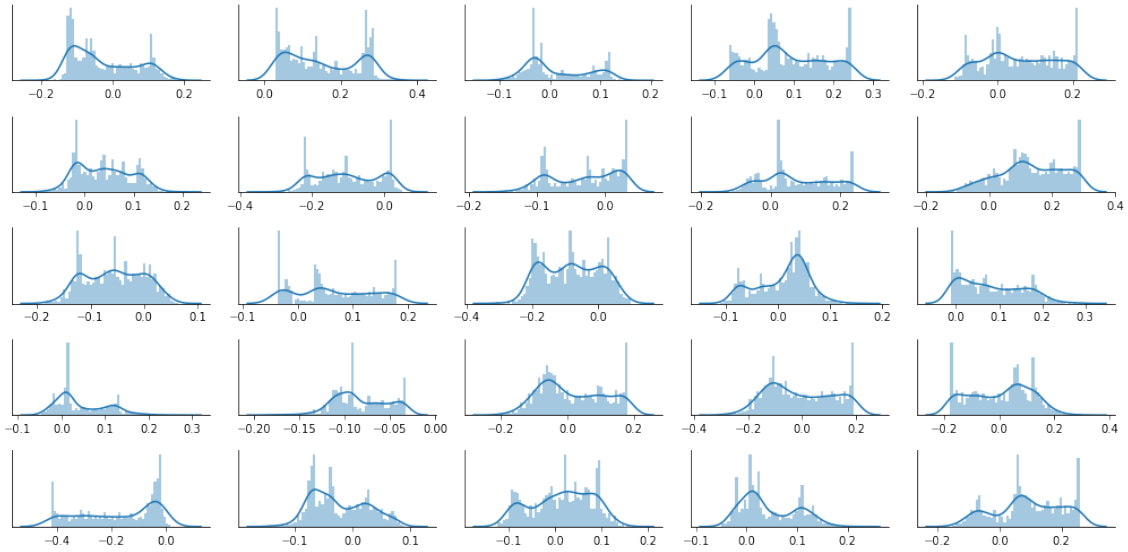
(f) Examples of the correlations within the posterior weight distributions using HMC.

Figure A.3: Illustration of the correlations within the posterior distributions of the first weights of a fully connected network trained on a toy regression task approximated by BBB (a), MNF (b), various settings of BbH (c-e) and HMC (f).

A.2 LeNet on MNIST



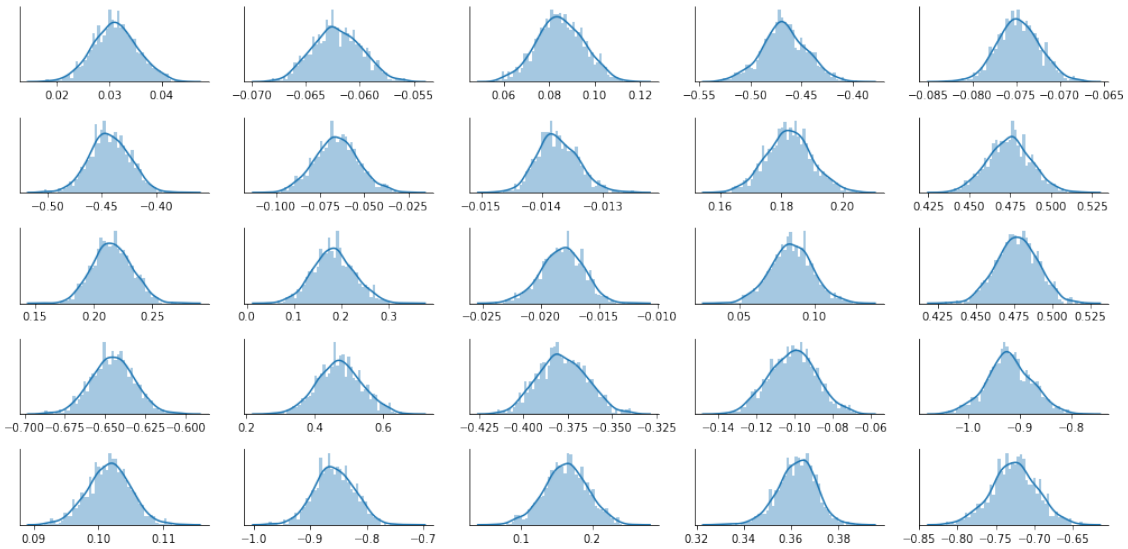
(a) Examples of the posterior weight distributions of the LeNet using MNF.



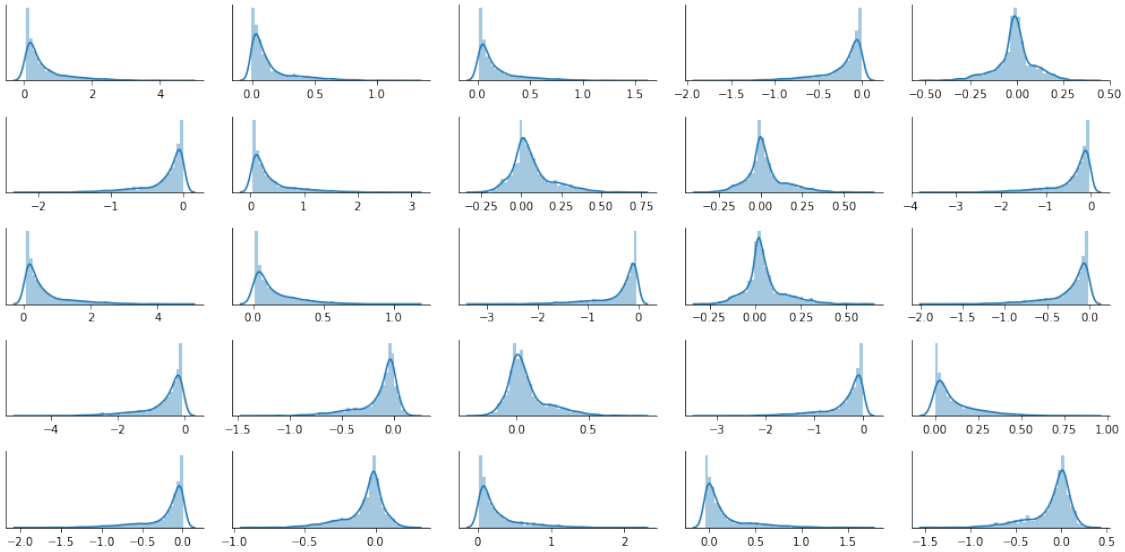
(b) Examples of the posterior weight distributions of the LeNet using BbH.

Figure A.4: Illustration of the posterior distributions of the 25 first weights of a LeNet trained on the MNIST digit classification task approximated by MNF (a) and BbH (b). BbH clearly generates more complex approximations whereas MNF's resemble Gaussians.

A.3 ResNet-32 on CIFAR-5

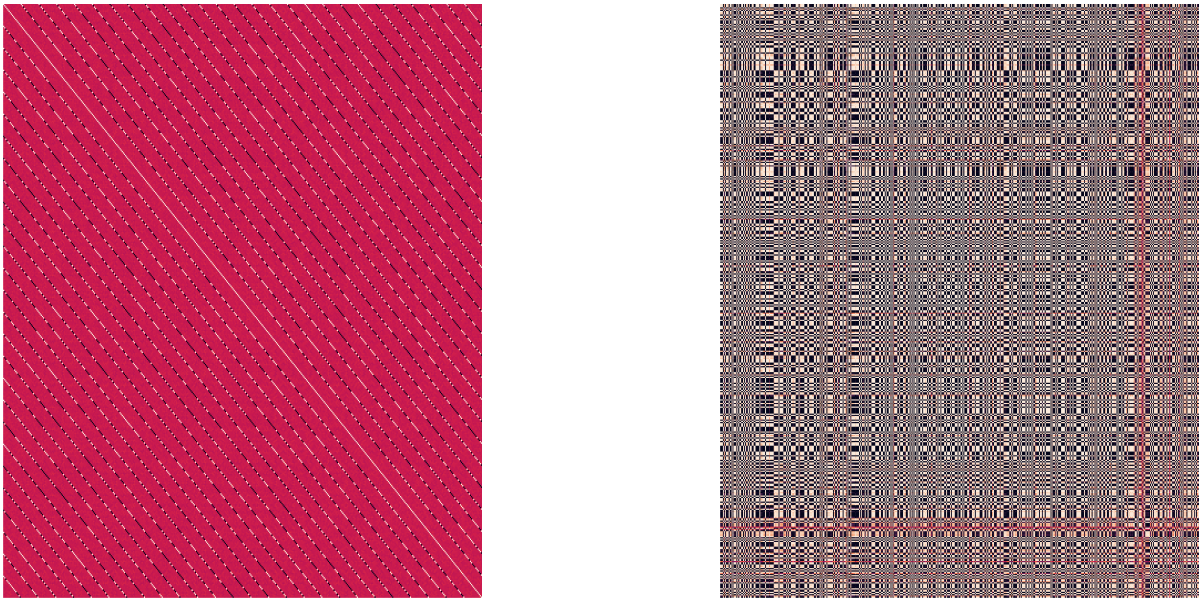


(a) Examples of the posterior weight distributions of the LeNet using MNF.



(b) Examples of the posterior weight distributions of the LeNet using BbH.

Figure A.5: Illustration of the posterior distributions of the 25 first weights of a ResNet-32 trained on the CIFAR-5 classification task approximated by MNF (a) and BbH (b). MNF models posterior distributions that resemble Gaussians. *BbH* models distributions that are more complex than those of MNF but less than the ones it modelled for the MNIST task.

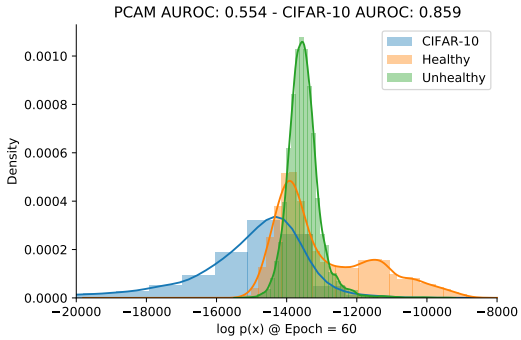


(a) Correlations modelled by MNF.

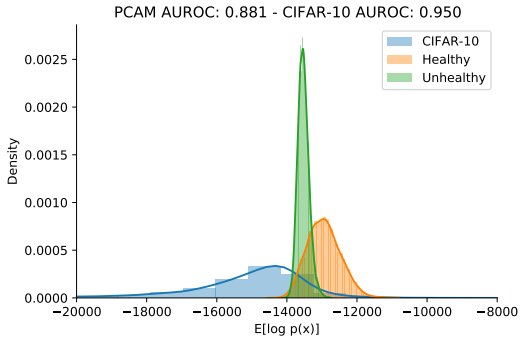
(b) Correlations modelled by *BbH*.

Figure A.6: Illustration of the correlations between the weights in the first convolutional layer of a ResNet-32 trained on the CIFAR-5 classification task modelled by the posterior distributions approximated by MNF (a) and *BbH* (b). Dark spots indicate negative and bright spots positive correlation. *BbH* models complex dependencies between the weights whereas MNF is only capable of modelling dependencies along the dimension of the multiplicative factor.

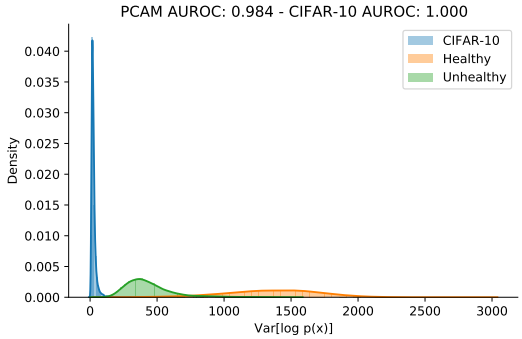
B Density plots for histopathology OOD detection



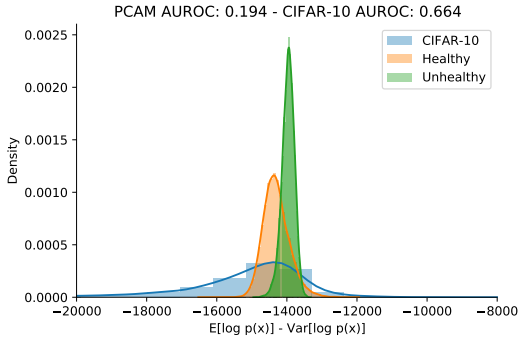
(a) Using the regular log likelihood, $\log p(x)$, for OOD detection.



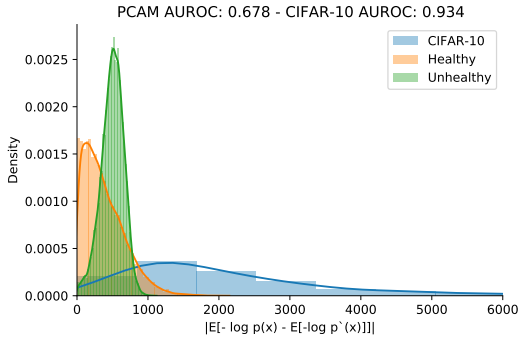
(b) Using the expected log likelihood, $\mathbb{E}[\log p(x)]$, for OOD detection.



(c) Using the variance of log likelihood, $\mathrm{Var}[\log p(x)]$, for OOD detection.



(d) Using the WAIC, $\mathbb{E}[\log p(x)] - \mathrm{Var}[\log p(x)]$, for OOD detection.



(e) Using the expected typicality, $|\mathbb{E}_{epochs}[-\log p(x)] - \mathbb{E}_x[-\log p_{train}(x)]|$, for OOD detection.

Figure B.7: Comparison of the distribution of the different outlier metrics on the validation set of healthy and unhealthy PatchCamelyon images as well as on CIFAR10.