

University of London  
Imperial College of Science, Technology and Medicine  
Department of Computing

# **Generating Semantically Enriched Diagnostics for Radiological Images using Machine Learning**

Aydan Gasimova

Submitted in part fulfilment of the requirements for the degree of  
Doctor of Philosophy in Computing of the University of London and  
the Diploma of Imperial College, September 2021



## 0.1 Statement of Originality

This work is my own, aside from where appropriate references are stated. The final chapter is a joint work completed by myself and Gavin Seegoolam, and is indicated as such.

## 0.2 Copyright Statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

## Abstract

Development of Computer Aided Diagnostic (CAD) tools to aid radiologists in pathology detection and decision making relies considerably on manually annotated images. With the advancement of deep learning techniques for CAD development, these expert annotations no longer need to be hand-crafted, however, deep learning algorithms require large amounts of data in order to generalise well. One way in which to access large volumes of expert-annotated data is through radiological exams consisting of images and reports. Using past radiological exams obtained from hospital archiving systems has many advantages: they are expert annotations available in large quantities, covering a population-representative variety of pathologies, and they provide additional context to pathology diagnoses, such as anatomical location and severity. Learning to auto-generate such reports from images presents many challenges such as

the difficulty in representing and generating long, unstructured textual information, accounting for spelling errors and repetition or redundancy, and the inconsistency across different annotators. In this thesis, the problem of learning to automate disease detection from radiological exams is approached from three directions. Firstly, a report generation model is developed such that it is conditioned on radiological image features. Secondly, a number of approaches are explored aimed at extracting diagnostic information from free-text reports. Finally, an alternative approach to image latent space learning from current state-of-the-art is developed that can be applied to accelerated image acquisition.



## Acknowledgements

I would like to express my gratitude to my supervisor, Daniel Rueckert, who has always made me feel like I can do this. I would like to thank him for always being understanding and supportive, and for always having time for catch-ups despite his busy schedule.

I would like to thank Amani for always being there as a source of support and helping me with finding the extra time and funding to finish this PhD. I'd also like to thank her for supporting Women in Computing in the department, and hope that the community we all helped to create continues to inspire and support more women.

I would like to thank my friends and fellow researchers at BioMedIA for fostering such an open, welcoming, and stimulating environment. Special mentions to Gavin and Harvey for being in the office at odd hours and weekends and keeping me company.

I would like to thank my mother, Nigar, for letting me live at home during the writing up of this thesis and making what was a stressful time more bearable through home-cooked meals and full use of the sun-room.s



# Contents

0.1	Statement of Originality . . . . .	i
0.2	Copyright Statement . . . . .	i
	<b>Abstract</b>	<b>i</b>
	<b>Acknowledgements</b>	<b>iii</b>
<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Thesis overview . . . . .	4
1.3	Contributions . . . . .	6
1.4	Publications . . . . .	7
1.5	Datasets . . . . .	8
1.5.1	Indiana University Chest X-ray Dataset . . . . .	9
1.5.2	Imperial College Local Hospital Brain DWI . . . . .	12
1.6	Evaluation Metrics . . . . .	15
1.6.1	BLEU . . . . .	15
1.6.2	ROUGE . . . . .	16

1.6.3	DAPS . . . . .	16
<b>2</b>	<b>Background</b>	<b>18</b>
2.1	Theoretical Background . . . . .	18
2.1.1	Building Blocks of Encoder-Decoder Frameworks for Image Captioning .	19
2.1.2	Medical Concept Extraction and Word Representation Learning . . . . .	24
2.1.3	Image Latent Space Learning through Autoencoders . . . . .	29
2.2	State-of-the-art in Image Caption Generation . . . . .	29
2.2.1	Image Caption Generation in Computer Vision . . . . .	29
2.2.2	Radiology Reports As Image Annotations . . . . .	36
2.3	Summary . . . . .	39
<b>3</b>	<b>Report Generation for Single and Multi-View Radiological Images</b>	<b>41</b>
3.1	Introduction . . . . .	41
3.2	Datasets . . . . .	43
3.2.1	Preprocessing - Chest X-rays . . . . .	44
3.2.2	Preprocessing - Brain DWI . . . . .	45
3.3	Static Image Embedding Models . . . . .	45
3.3.1	Related Work . . . . .	45
3.3.2	Model Architectures . . . . .	46
3.3.3	Experiments . . . . .	49
3.3.4	Results . . . . .	51

3.3.5	Summary . . . . .	58
3.4	Dynamic Image Embedding . . . . .	59
3.4.1	Dataset . . . . .	59
3.4.2	Related Work . . . . .	59
3.4.3	Model Architectures . . . . .	60
3.4.4	Experiments . . . . .	63
3.4.5	Results . . . . .	63
3.5	Conclusion . . . . .	65
3.6	Related Publications . . . . .	66
<b>4</b>	<b>Medical Concept Extraction from Free-Text Diagnostic Reports</b>	<b>70</b>
4.1	Introduction . . . . .	70
4.2	Datasets . . . . .	73
4.3	Statistical and ontology-based concept extraction of chest X-ray reports . . . . .	74
4.3.1	Related work . . . . .	74
4.3.2	Methods . . . . .	75
4.4	Anatomical brain region mapping using manual extraction and a hierarchical ontology . . . . .	83
4.4.1	Method . . . . .	84
4.4.2	Results . . . . .	85
4.5	Conclusion . . . . .	87

<b>5</b>	<b>Abstractive Concept Extraction from Free-Text Diagnostic Reports</b>	<b>91</b>
5.1	Introduction . . . . .	91
5.1.1	Related work . . . . .	92
5.1.2	Method . . . . .	93
5.1.3	Experimental Settings . . . . .	95
5.1.4	Results . . . . .	97
5.2	Conclusion . . . . .	98
5.3	Related Publications . . . . .	99
<b>6</b>	<b>Image Latent Space Learning for Diagnostic Report Generation</b>	<b>100</b>
6.1	Introduction . . . . .	100
6.2	Related Work . . . . .	102
6.2.1	Latent space learning of accelerated MRI . . . . .	102
6.3	Methods . . . . .	103
6.3.1	Latent space learning . . . . .	104
6.3.2	Report generation model . . . . .	106
6.4	Dataset . . . . .	106
6.5	Experiments . . . . .	107
6.6	Results . . . . .	108
6.7	3D-DWI Extension . . . . .	109
6.8	Conclusion . . . . .	111
6.9	Related Publications . . . . .	112

**7 Conclusion 115**

7.1 Summary of Thesis Achievements . . . . . 115

7.2 Limitations and Future Work . . . . . 118

**Bibliography 119**

.1 Patterns and diagnoses groupings of MeSH terms . . . . . 139

.2 MetaMap extracted terms . . . . . 139





# List of Tables

1.1	Top 20 most commonly occurring MeSH findings, taken from MeSH annotations of the IU-CX images [1]. . . . .	12
1.2	IU-CX [1] report vocabulary and length statistics. IU-CX MeSH-sp-subset subset of exams with one a single reported abnormal pattern. Avg. s/e refers to the average number of sentences per exam, avg. t/e is average tokens per exam, and STD t/e is standard deviation of tokens per exam. . . . .	12
1.3	MeSH terms categorised into diagnosis, patterns, anatomy, position and severity labels. . . . .	12
3.1	IU-CX [2] statistics of free text reports (IU-CX free-text and ICH-Brain-DWI), MeSH (IU-CX MeSH) and MeSH single-pattern subset (IU-CX MeSH-sp-subset) annotations after processing. Avg. s/e refers to the average number of sentences per exam, avg. t/e is average tokens per exam, and STD t/e is standard deviation of tokens per exam. . . . .	45
3.2	Image and text augmentation parameters. . . . .	49
3.3	BLEU n-gram scores of static image embedding models train (tr) an test (test) metrics, evaluated on the IU-CX MeSH-sp PA-view subset, compared with re-current cascade model of [3]. . . . .	53

3.4	Disease, anatomy, position and severity (DAPS) averaged F1 scores of static image embedding models train (tr) an test (test) metrics, evaluated on the IU-CX MeSH-sp PA-view subset. . . . .	53
3.5	BLEU n-gram scores of SERepGen-merge trained and evaluated on the IU-CX with PA- and L-view images and ICH-Brain-DWI dataset with axial slice images. Reported performance on test set. Results are compared against initialising the trained report generation model on random noise (in the shape of the image, with matched mean and variance to the image space). . . . .	57
3.6	Disease, anatomy, position and severity (DAPS) averaged F1 scores of SERepGen-merge trained and evaluated on the IU-CX MeSH-single-pattern subset and full IU-CX MeSH dataset with PA- and L-view images. Reported performance on test set. . . . .	57
3.7	BLEU n-gram and ROUGE-1 F1 scores of DAREpGen trained and evaluated on the IU-CX free-text single (PA) and multi-view (PA+L) datasets, and ICH-Brain-DWI single axial slice and 20 axial slice datasets. SERepGen-merge results displayed for comparison. Reported performance on test set. . . . .	65
4.1	Top 20 Tags Assigned to the X-Ray Reports by MetaMap . . . . .	76
4.2	Total vocabulary and percentage of ‘normal’ cases when using multiple MetaMap tags to extract disease terms and phrases. . . . .	77
4.3	Top 10 Metamap extracted ‘Disease or Syndrome’, ‘Finding’ and ‘Pathologic Function’ terms and their percentage overlap with other phrases. Percentage is calculated as total appearance as a single phrase divided by appearances with and within other phrases. For instance, the word ‘normal’ appears as single Metamap term 2512 times, but also appears within other Metamap terms such as ‘normal heart size’. . . . .	78

4.4	Sample of disease phrases extracted from largest cluster of rare words from tf-idf and word2vec k-means. . . . .	81
4.5	Multi-class experimental set-up for extractive and clustering labelling methods. .	82
4.6	Classification comparison of labelling methods using micro-precision, -recall, -F1 and macro-precision, -recall and -F1. Results reported on test set. Micro-P/R/F1 refers to the average over classes, and macro-P/R/F1 are averaged over samples. . . . .	82
4.7	Top 20 classes of brain regions after re-assignment based on a hierarchical ontology.	84
4.8	ICH Brain DWI multi-label classification results using extractive labelling technique. The accuracy of the ‘infarct’ class is reported separately as well as part of the mean average accuracy, precision and recall (mAA, mAP, mAR) and Hamming Loss (HL) of all classes. . . . .	86
5.1	TextCNN2Seq performance comparison with seq2seq models. Reported are Rouge unigram bigram, and longest sequence F1 scores, and BLEU 1-4-gram scores. All metrics are reported on the test set. . . . .	97
6.1	F1, MSE and PSNR metrics of ground truth and accelerated MRI. F1 score is taken at the output of the classifier module and MSE and PSNR metrics from the output of the reconstruction module of the autoencoder. All metrics are reported as average over samples. . . . .	109
6.2	BLEU1,2,3,4-gram and ROUGE1 F1, precision (P) and recall (R) metric comparisons on increasingly accelerated image embeddings. . . . .	109
6.3	Results of ablation study . . . . .	110
6.4	Sample ground truth and generated reports from fully sampled and undersampled 3D brain DWI. Correctly identified concepts are highlighted. . . . .	111

1	Chest X-ray patterns and diagnoses manually extracted from MeSH annotations of the IU-CX images [1]. Definition of patterns and diagnosis taken from Smithuis and van Delden [4]. . . . .	140
2	Metamap extracted ‘Disease or Syndrome’, ‘Finding’ and ‘Pathalogic Function’ terms that appear in at least 30 reports, and their degree of overlap with other terms. . . . .	141

# List of Figures

1.1	Sample chest X-ray exam from IU-CX [2] . . . . .	10
1.2	Patterns of lung abnormalities resulting in increased density on chest X-rays [4]	11
1.3	Samples of three unique brain DWI exams, shown as slices with respective filtered reports. Slices were selected based on maximum ischaemic infarct segmentation area using the segmentation network of Chen et al. [5]. . . . .	13
1.4	2D Brain DWI subset procedure: 3D images are passed through a brain ischemia segmentation network of Chen et al. [5]. Slices fitting a heuristic thresholding criteria are selected as having a visible acute ischemia, and non-acute slices are sampled from normal brain images according to the same axial plane distribution as the acute slices. . . . .	14
2.1	Single-layer perceptron . . . . .	20
2.2	Convolutional neural network with 2 hidden layers. . . . .	21
2.3	Recurrent neural network, rolled and unrolled. . . . .	22
2.4	Encoder-decoder configuration for image captioning. . . . .	24
2.5	MetaMap System Diagram . . . . .	26
2.6	Autoencoder network for dense image representation through reconstruction. . .	29

2.7	Illustration of ‘Meaning’ space triplet of object/action/scene, reproduced from Farhadi et al. [6]. Framed as a Markov Random Field, where each object, action, scene is a node and edges are relationships between nodes. . . . .	31
2.8	Textual descriptions to inform semantic segmentations through a conditional random field, reproduced from Fidler et al. [7]. Object relations are extracted from text and used to re-rank candidate bounding boxes for object detection. . .	32
2.9	Neural image caption model, reproduced from Vinyals et al. [8]. CNN encodes image into a dense representation and input into the LSTM at time step -1. Images and words are mapped into the same embedding space. . . . .	33
2.10	Mind’s eye: image captioning through visual feature reconstruction, reproduced from Chen et al. [9]. The green modules represent the RNN language model, $\mathbf{v}$ is a vector of observed visual features, and $\tilde{v}$ is the reconstruction of the visual features. . . . .	35
2.11	High-level illustration of recurrent attention model of Xu et al., reproduced from [10]. Instead of a single vector representation as the input to the language model RNN, they propose the use of the lower-level features maps. ‘Attention’ is characterised as the learned weights over these features, as well as the previously generated words. . . . .	36
2.12	Recurrent neural cascade training sequence, reproduced from Shin et al. [3]. . .	38
3.1	Image-report learning architectures using a single static image embedding as an aggregate representation of all input image views. . . . .	48
3.2	Multi-view static embedding. All multi-view images from a single exam are passed through a pre-trained classification CNN where the classification layer has been removed. The outputs are then the considered the image representations and are combined into a single static vector through a combination operation, such as max, concat, sum. . . . .	49

- 3.3 5-Fold averaged cross validation hyperparameter studies of static image embedding models SERepGen-init, SERepGen-inject and SERepGen-merge. All models were trained and evaluated on the IU-CX MeSH-sp PA view subset. Reported metrics are BLEU-1, BLEU-2, BLEU-3, BLEU-4, Rouge-F1, and the F1 scores of groupings of disease, anatomy, position and severity terms. Results on training set are at half-transparency and results on validation set are opaque. The first four experiments in the figure are the hyperparameter studies of SERepGen-init, the next five are SERepGen-inject, and the last three are SERepGen-merge. . . . 52
- 3.4 Sample MeSH annotation prediction generated by SERepGen-merge on test set of IU-CX-sp PA view subset. Per-sample BLEU n-gram scores are reported beneath each prediction based on the original MeSH annotation. The disease, anatomy, position and severity recall and precision are reported on the batch at the bottom of the figure. . . . . 56
- 3.5 An illustration of soft-attention image captioning, reproduced from model description in [10] . . . . . 61
- 3.6 Dynamic attention report generation (DAREpGen) module where attention is computed using the previous hidden state of the LSTM. Image convolutional features are a concatenation of conv features from either a single or multiple views. 62
- 3.7 IU-CX free-text sample test report predictions and attention maps of DAREpGen. 67
- 3.8 ICH-Brain-DWI sample test report predictions and attention maps of DAREpGen. 68

3.9	20-slice brain DWI exams with predicted reports. Slices highlighted in blue on the left are the slices selected by the 2D brain-DWI subset selection procedure described in Section 1.5.2. They are displayed for comparison in order to highlight which slice (if any) contains the ischemia according to the segmentation network of Chen et al. [5]. Slices on the right, highlighted in orange, are selected based on max attention weights at each time-step. The generated word and selected slice number for that word are displayed in the top left corner of each image. The original reports are displayed underneath each sample. . . . .	69
4.1	Sample report and Medical Subject Heading (MeSH) annotations. Highlighted are phrases in the report that contribute the most to the MeSH annotations. . .	73
4.2	MetaMap sample output: identified phrase ‘interval development of bandlike opacity’ and two out of 18 candidate mappings, with corresponding MetaMap score. . . . .	75
4.3	K-means clustering performed on tf-idf representations of disease phrases extracted by MetaMap. . . . .	88
4.4	K-means clustering performed on averaged word2vec representations of disease phrases extracted by MetaMap. . . . .	89
4.5	Log of disease phrase frequency distributions over k-means clusters performed on tf-idf and word2vec representations. . . . .	90
4.6	Central slice of sample DWI exam with corresponding clinical report, clinical diagnosis, manually extracted brain regions and region mappings. . . . .	90
4.7	Attention-guided clinical report generation model. . . . .	90
5.1	TextCNN2Seq-Att model schematic. Best viewed in colour. . . . .	96
5.2	Sample output MeSH summary from the TextCNN2Seq+Att with largest and second-largest attention weights highlighted in colour. . . . .	98



6.1	An autoencoder is trained to reconstruct the fully-sampled image through an L2 loss. The latent space is conditioned to encode pathological information by performing a classification of ischaemia, trained with a binary cross-entropy loss. The latent space encoding learned at the bottleneck is used as a training target for the encoding branch which only sees the accelerated image. . . . .	103
6.2	Left to right: (1) An example of a brain with ischaemia (2) The corresponding x16 accelerated image is zero-fill reconstructed from k-space using a 2D Fourier Transform. Note that this image suffers from heavy aliasing artefacts. (3) A projection of the first two principle components in a PCA analysis of the latent space. Some clustering can be seen (4) a t-SNE projection of the latent space showing clear clustering. . . . .	105
6.3	Clinical report generation model from accelerated image latent space embeddings.	105
6.4	Sample brain slices and associated reports generated from non-accelerated and increasingly accelerated image embeddings. Correctly identified pathology (acute/non-acute) and spatial contexts are highlighted in blue. . . . .	113
6.5	Average BLEU-n scores of accelerated brain volumes. . . . .	114



# Chapter 1

## Introduction

### 1.1 Motivation

Computer aided detection (CAdE) and diagnosis (CAdx) systems are tools designed to assist radiologists in the interpretation of medical images such as x-rays, MRIs and ultrasound. Early CAD systems were developed in order to combat fatigue that radiologists experience from interpreting large volumes of images, as well as human error that may occur in the interpretation of these images [11]. Traditional CAdE tools relied on hand-crafted image features tailored to specific tasks, and their purpose was to highlight these features, or abnormalities, to the radiologist, who ultimately made the diagnostic judgement [12, 13, 14, 15]. CAdx tools go one step further and attempt to classify these image features, for instance, classifying nodule features into benign or malignant [16, 17, 18]. Modern-day CAD systems can incorporate both detection and diagnosis in order to optimise the work-flow of radiologists and thereby reduce reading time, as well as improve the accuracy and consistency of diagnoses [19, 20]. A typical CAD scheme will incorporate one or more of the following: 1) image preprocessing, such as noise reduction, contrast and exposure levelling 2) image segmentation into anatomical regions such as organs, tissue types or possible lesions; 3) quantitative analysis of regions of interest or abnormalities, such as form, size, location; 4) evaluation/classification of image features [21].

Techniques used in developing CAD diagnosis tools began with rule-based approaches [22, 23,

24] that relied on computing filters and using anatomical knowledge to select for filter responses. These filters and if-then reasoning relied on a number of assumptions about the data, which may not be valid in every situation to which they are applied. Statistical and machine learning approaches, such as feature selection methods and classifiers, can be given a range of features, with fewer assumptions on which features will result in a ‘better’ prediction [25, 26, 18, 27]. The choice of features is still an important part of machine learning approaches, and hand-crafting them will still introduce bias. Additionally, even with the most comprehensive feature extraction techniques, there is an unavoidable loss of underlying information from the images. With the increase in computational power and the availability of larger volumes of imaging data from picture archiving and communication systems (PACS) of increasingly higher resolution quality, recent approaches have begun to adopt deep learning (DL) techniques [28, 29, 30]. Deep learning refers to a subset of machine learning algorithms that use multiple layers of artificial neural networks to extract features of increasing levels of abstraction by iteratively updating the weights of the network with respect to a pre-defined task. In this way, DL does away with the need for domain knowledge and hand-crafted features and instead, learns to extract features directly from the images.

Observer performance studies are generally in agreement that CAD systems improve the sensitivity, specificity and AUC of a radiologist’s diagnosis [31, 32, 33, 34]. They do, however, have their limitations. Hand-crafted image features and radiologists’ expertise as ground truth introduce human bias. DL algorithms reduce some of this bias by removing the need for manual feature engineering and can instead learn an intermediate image space, potentially even learning features not previously considered by radiologists. However, DL algorithms are still trained on human-labeled ground-truths, and therefore require much larger databases in order to generalise well. Without large and diverse training data, DL algorithms are liable to suffer from overfitting, whereby the networks learn features unique to the data they are trained on, and perform poorly on unseen data. Additionally, supervised training algorithms require task-specific expert annotations, sometimes more than is required as part of the diagnostic process, as in the case of segmentation. It is therefore an additional time-consuming task for radiologists to create annotations of the quality required for supervised DL. The annotations may

also require updating when applying the algorithms to different populations, or once imaging techniques advance.

This thesis demonstrates how data gathered as part of hospital data management, specifically radiological images and their corresponding free-text reports, can be used as part of a supervised learning framework in order to automate the generation of diagnostics from unseen radiological images. This has potential uses as part of a CAD tool for assisting radiologists in making diagnoses, and additionally framing the diagnoses in a textual report that summarises the visual features of the pathologies in a similar style to that of a radiologist. The motivation for using PACS databases directly is two-fold: firstly, given the necessary permissions (e.g. ethical approvals since patient data is identifiable in clinical PACS), they are available in large quantities and can be gathered from many hospitals, and secondly, ‘ground-truth’ diagnostics are gathered as part of radiological exams in the form of diagnostic reports and so dispenses the need for manual image annotation.

There are, of course, still practical challenges to consider in gathering data from PACS for the purpose of automated diagnostics generation: a large, standardised, representative and fully characterised set of examples is required. Collecting standardised patient data is difficult as data gathering protocols vary across institutions, contain manually recorded ‘free-text’ information, and reports are typically unstructured and possibly incomplete and/or contain errors. There has been a push to introduce structured reporting systems into the workflow of radiologists to improve consistency and reproducibility, which in turn would benefit data mining and machine learning [35, 36, 37]. However, these systems are difficult for radiologists to adopt into their workflow as they tend to require more time, as well as limit the level of expression and detail a radiologist can provide [38]. Acquiring fully representative data is challenging due to the low prevalence of certain diseases, but also due to biases introduced when data is acquired from one hospital, or specific geographical regions. Fully-characterised patient data may consist of multiple modalities: radiological images (multi-modality and multi-view), radiologist reports, lab reports, patient history, interviews and physician’s notes. These may be incomplete for many patients, and conversely, many of these features may have no diagnostic value and introduce unnecessary noise.

Even with structured and standardised diagnostic reports, there are challenges in designing a report generation model based on supervised learning. Firstly, in order to learn to generate reports from a radiological image, the image must be encoded in such a way that the encoding retains semantic information - for instance, the pathology present in the image and its anatomical location. This is a challenge as most work on generating image embeddings is done for the purpose of classification and segmentation, meaning we cannot directly take advantage of pre-trained models for transfer learning. Secondly, given a semantically-rich image space, the report generation model must balance learning a syntactic language model (a grammatically coherent structure) as well as a semantic model where the generated words are conditioned on image features.

It is also important to consider whether a syntactic language model needs to be learned at all, and whether it is more prudent to approach this as a classification task whereby disease labels are extracted from the reports as an intermediate step, and assigned as class labels to the images. The main challenges in this approach are that extracting disease labels from free-text reports made by humans is not a simple case of using regular expressions: radiologists have many ways of describing the same disease, or a disease may be referred to in multiple ways. Additionally, one disadvantage of extracting only disease labels is that we lose all the contextual and visual information about the disease, for instance its severity, visual characteristics, and anatomical location, all of which is useful knowledge when making a differential diagnosis.

## 1.2 Thesis overview

The goal of achieving an effective supervised framework from radiological exams is approached from three main directions:

- The challenge of learning a text generation model that is conditioned on radiological image features
- The challenge of distilling diagnostic information from free-text reports to be assigned as image label

- The challenge of encoding radiological images such that the embedding space retains semantic information.

To begin with, in Chapter 3 the focus is on investigating encoder-decoder image captioning models that are trained to generate diagnostic reports from radiological image representations. The ‘encoder’ portion of the model can either be an encoder of images, (for instance, using a CNN to encode images into dense representations), or a multi-modal encoder of image-text (using a CNN and RNN). The decoder portion then takes this representation as input and generates an output, in this case, the diagnostic report. Different encoder-decoder configurations are considered in this chapter, beginning with encoding the images into a single image feature vector using a CNN pre-trained on natural images (due to the low availability of labeled radiological images). These image embeddings are then incorporated into an RNN sequence learning model either by initialisation, concatenation at the input, or concatenation at the output. Each regime was trained and evaluated on single and multi-view radiological image datasets, with varying degrees of report complexity. In the same chapter, this work was extended by exploring image captioning frameworks that aim to capture dynamic (attention-based) and location-specific image representations. This is motivated by the fact that using lower-level image features and recurrent attention over these features provides a richer input to the sequence learning model.

The next challenge addressed by this thesis in Chapter 4 and Chapter 5 is the extraction of diagnostic information from free-text radiological reports. This is useful for many further applications, including image retrieval, image classification and image captioning tasks. Two concept extraction approaches are explored and compared: a combination of ontological and statistical tools for classification label extraction, and abstractive text summarisation using machine learning. Chapter 4 describes the statistical and ontological approach: a series of off-the-shelf tools were used to extract findings from images, grouped using clustering of word representations, and then assigned as image labels for image classification. Chapter 5 then explores mapping the free-text reports into a summary of findings using abstractive text summarisation techniques. In this thesis, a summary report is defined as one that is vocabulary controlled and contains pathologies present in the image(s), their anatomical location, and visually-descriptive

features. The reasoning behind this structure is that these concepts directly correspond to visual features in the image and are more suitable for image-captioning tasks than free-text reports. In addition, they are consistent with the natural reporting methods of radiologists, and so can be used as part of a CAD tool in daily reporting.

The last challenge addressed in this thesis in Chapter 6 is the challenge of capturing semantic information in the image space. The assumption that pre-trained CNN classification networks extract task-agnostic features is certainly compelling, however, it can be improved upon. The goal of this chapter is to explore the use of autoencoders for encoding images into a latent space to be used for report generation. The use of an autoencoder for this task is two-fold: autoencoders can be used in an unsupervised way to learn a dense representation of the image for report generation, and the same network can be used in conjunction with an auxiliary task of denoising such that report generation can be performed on noisy or aliased images. This is particularly useful in magnetic resonance imaging (MRI), where scans can take hours, but can be sped up through sampling in the frequency space. This approach of report generation from latent space is evaluated on brain diffusion-weighted MRI of stroke patients, where diagnosis of an ischaemic stroke is time-sensitive. The method is first evaluated on pre-processed 2D axial slices, and then on the full 3D brain images. The results showed that image embeddings taken from the latent space of an autoencoder trained on the dual task of classification and reconstruction out-performed the attention network on report generation, even at high image acceleration rates. The results on 3D images were poorer than for 2D slices, however, it demonstrated that reports can be generated from highly-accelerated 3D images without the need for pre-trained networks.

## 1.3 Contributions

The main contributions of this work are as follows:

- Investigation into optimal multi-view encoder-decoder image captioning configurations for radiological image report generation.



- Evaluation of report generation methods when trained on vocabulary-controlled, structured versus varied and unstructured raw reports.
- Investigation into combined techniques of statistical ontological tools for label extraction from radiological reports.
- Evaluation of report label extraction methods on radiological image classification.
- Application of abstractive text summarisation to radiology reports in order to map raw reports into vocabulary-controlled summaries of key pathologies and their visual descriptions.
- Investigation into an alternative approach to image latent space learning for the task of accelerated image report generation through the use of autoencoders.

## 1.4 Publications

One of the techniques described in Chapter 3 was first published in a study done on knee X-ray report generation: Gasimova, A. (2017). Automated Knee X-ray Report Generation, In NeurIPS Workshop on Machine Learning for Health, 2017.

The deep learning-based model for concept extraction described in Chapter 5 and consequent report generation described in Chapter 3 was published in a study that first used annotated reports to train a model for concept extraction, and then used the extracted concepts for chest X-ray report generation: Gasimova, A. (2019). Automated enriched medical concept generation for chest X-ray images. In Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support (pp. 83-92). Springer, Cham.

The latent space learning for accelerated report generation described in Chapter 6 was published in a study on report generation for brain diffusion weighted imaging: Gasimova, A., Seegoolam, G., Chen, L., Bentley, P., Rueckert, D. (2020, October). Spatial Semantic-Preserving Latent Space Learning for Accelerated DWI Diagnostic Report Generation. In International

Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 333-342). Springer, Cham.

## 1.5 Datasets

The datasets necessary to evaluate automated radiology report generation from images must contain expert-generated radiology reports that can be used as a ground-truth. The Indiana University Chest X-ray collection (IU-CX) [2] is suitable for this as it is a collection of outpatient examinations with associated free-text radiology reports. Additionally, the text reports were manually annotated by two experts with MeSH terms [39], supplemented by Radiology Lexicon, RadLex [40], codes. These annotations were made according to the principles outlined in NLM Indexing Manual and Technical Memoranda [41]: using MeSH terms with qualifiers (such as MeSH term ‘lung’ with qualifier ‘upper lobe’), where available in the free-text reports. These vocabulary-controlled, semi-structured annotations are suitable for the second task in this thesis of evaluating automatic extraction of diagnostic information from free-text reports, explored in Chapter 5.

A second dataset of brain-DWI examinations and radiology reports was available from Imperial College local hospitals. This dataset was also suitable for the evaluation of automated radiology report generation as the DWI reports were manually parsed to extract 1-2 sentences summarising the findings in the image, which, for stroke patients, is limited to determining whether there has been an ischaemic stroke or haemorrhage, and if so, where and to what extent. The scope of the pathology, and therefore the diversity of the language, is therefore less diverse than for the IU-CX (which report on all possible chest abnormalities). Therefore, even though they do not have expert annotations, they do have a potential to be parsed by a non-expert in a semi-automated way to extract and categorise brain regions, explored in Chapter 4.

### 1.5.1 Indiana University Chest X-ray Dataset

The Indiana University Chest X-ray collection (IU-CX) [2] is available publicly from the National Library of Medicine [1]. This dataset consists of 3,955 exams and 7,470 images from the hospital’s picture archiving systems. Each exam contains a posterir-anterior (PA) chest X-ray view, an associated radiological report, and a series of Medical Subject Heading annotations (MeSH<sup>®</sup>) all made by qualified radiologists. Some exams have additional images such as a second PA and/or a lateral chest X-ray view. Exams have all been fully anonymised to remove patient names and any identifiable information. Reports are made up of *Indication*: symptoms, *Findings*: visual features noted by the radiologist in the X-ray scan, and *Impression*: pathology diagnosis. MeSH annotations are (with some exceptions) formatted as  $[finding_0/description_0, \dots, finding_n/description_n]$  where *description* is a combination of *anatomy/position/severity*. An example of a full exam consisting of two PA view and one lateral view chest X-ray images, together with the full radiological report and MeSH annotations is illustrated in Figure 1.1. The ‘XXXX’ characters represent redacted information.

The Basic Interpretation guide by Smithuis and van Delden from the Radiology Department of the Rijnland Hospital [4] proposes a systematic inside-out approach to interpreting chest X-rays: examining first the heart, then mediastinum and hili, followed by lungs, lungborders and finally the chest wall and abdomen. If abnormalities are identified, a pattern approach is used to come up with the most likely differential diagnosis. Lung abnormalities typically present as areas of increased density and can be split into the following patterns: consolidation (small airways fill with dense material), atelectasis (lung collapse), nodule or mass, and interstitial (scarring), see Figure 1.2. The pleura (membrane covering the lungs) are examined for opacification, which can indicate pleural effusion or masses, and shape/displacement, which can indicate a pneumothorax. In the case of the heart, only the outer contours are visible on an X-ray, hence the main findings are either that the heart is normal or enlarged. The mediastinum and hili are also examined for displacement. Chest wall abnormalities are identified by rib deformation (typically due to old rib fractures) or metastases in vertebral bodies and ribs. Abdominal abnormalities can be identified by examining the diaphragm, where an

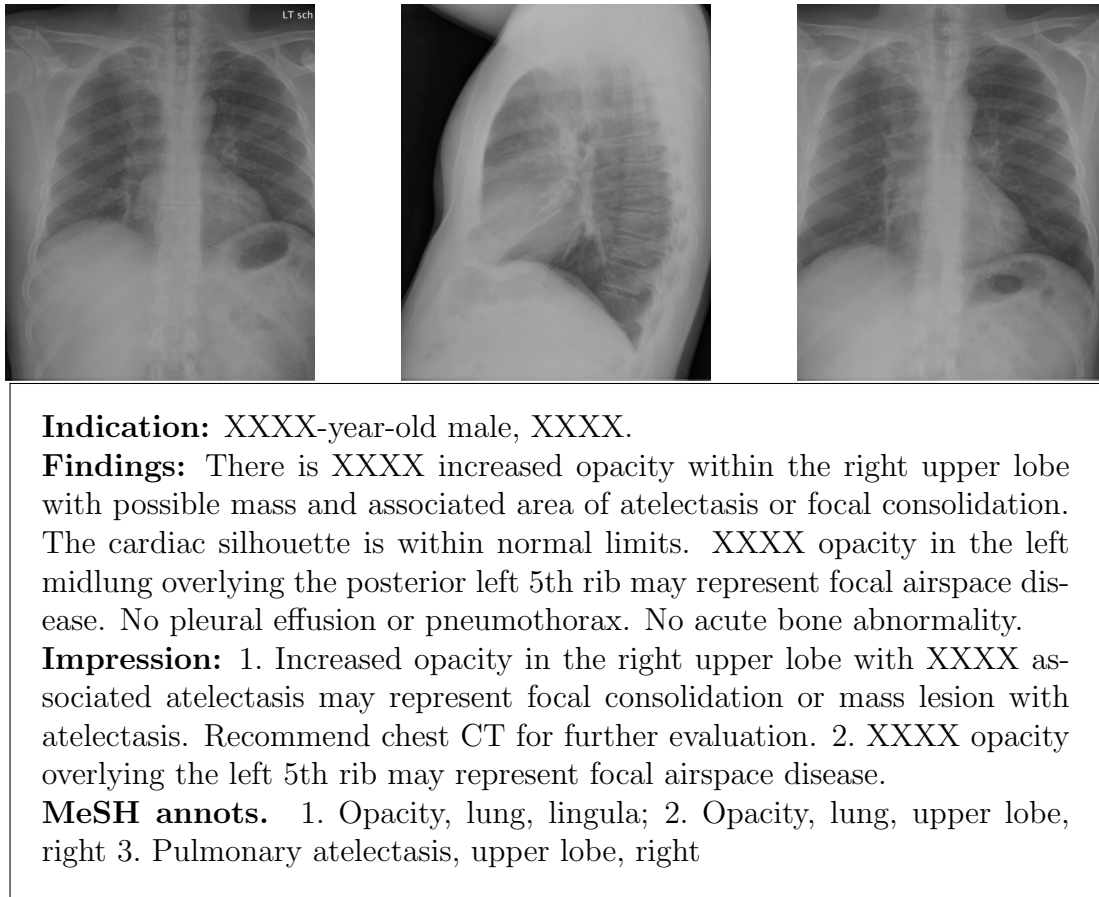


Figure 1.1: Sample chest X-ray exam from IU-CX [2]

unusual positioning of the diaphragm can suggest free abdominal air.

With this prior information, the findings in the reports can be grouped according to the patterns they present, as opposed to the diagnosis. The motivation for this is that a pattern may be caused by any number of different diseases, and diagnosing the disease is typically based on a number of other observations. In addition, the disease diagnostic made by a radiologist is not confirmed in the report itself, and hence cannot be treated as a ground-truth annotation for the purpose of supervised learning from reports.

A list of finding labels was created by compiling the *finding* terms of the MeSH annotations, i.e. any word(s) that were not anatomical location. Where a finding was made up of two or more words, they were combined into one label, such as ‘pulmonary atelectasis’. There were 94 finding labels identified, the top 20 of which are listed in Table 1.1 with respective occurrence frequency. The finding terms include a mixture of abnormal patterns and possible diagnoses. In the example in Figure 1.1, opacity and pulmonary atelectasis are general abnormal patterns. An

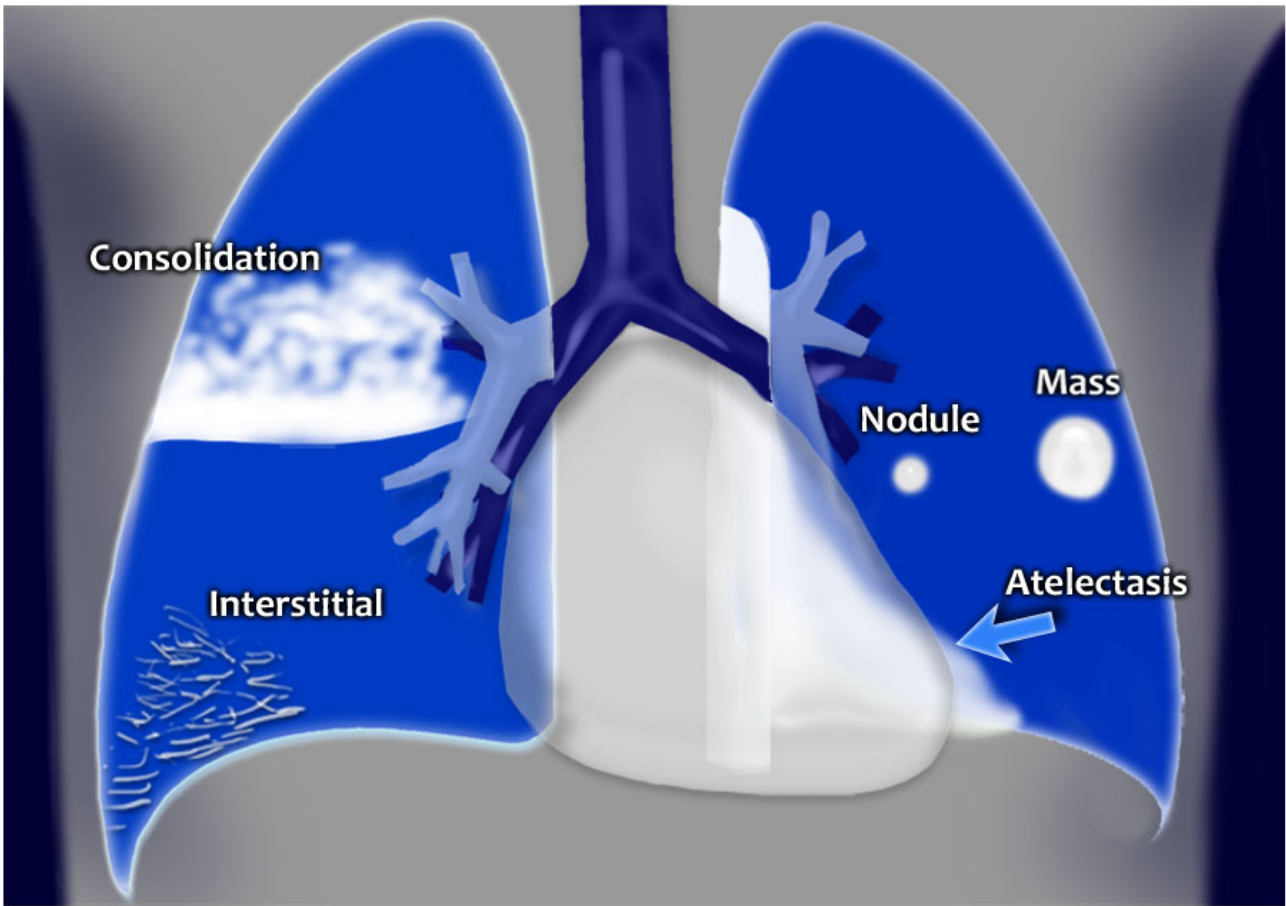


Figure 1.2: Patterns of lung abnormalities resulting in increased density on chest X-rays [4]

example of a disease diagnosis MeSH term in Table 1.1 would be ‘calcified granuloma’, as this appears as a mass on the X-ray. Another example, ‘airspace disease’, can be recognised with a pattern of consolidation and lung collapse in the lobes of the lung. The ability to recognise these patterns and their locations guides the diagnostic process, hence the 94 finding labels were grouped either under patterns or diagnoses, as per the distinctions outlined by Smithuis and van Delden [4]. The full list of MeSH terms and their groupings are listed in Appendix .1 Table 1.

A simplified subset of the IU-CX dataset was created by selecting exams with one MeSH pattern annotation, and therefore, assuming the radiologist’s interpretation is correct, only one main abnormality is present in the image(s). Statistics of the chest X-ray dataset’s free-text reports and MeSH annotations is summarised in Table 1.2. The *description* terms were also categorised into one of *anatomy/position/severity* based on a radiological medical dictionary [42]. These

categorisations of terms are only used for evaluation and not for training purposes. Sample labels and their categories are listed in Table 1.3.

normal	1357	deformity	125
opacity	512	atherosclerosis	125
cardiomegaly	347	airspace disease	124
calcinosis	332	catheters indwelling	122
pulmonary atelectasis	330	scoliosis	119
calcified granuloma	277	nodule	117
cicatrix	196	granulomatous disease	107
markings	168	surgical instruments	105
pleural effusion	161	fractures bone	93
density	129	aorta thoracic	90

Table 1.1: Top 20 most commonly occurring MeSH findings, taken from MeSH annotations of the IU-CX images [1].

	Exams	Vocab	Avg. s/e	Avg. t/e	STD t/e
IU-CX free-text	3,741	2088	5.81	44.60	22.20
IU-CX MeSH	3,741	178	2.08	7.35	7.12
IU-CX MeSH-sp-subset	2,237	131	1	3.33	1.46

Table 1.2: IU-CX [1] report vocabulary and length statistics. IU-CX MeSH-sp-subset subset of exams with one a single reported abnormal pattern. Avg. s/e refers to the average number of sentences per exam, avg. t/e is average tokens per exam, and STD t/e is standard deviation of tokens per exam.

	Diagnosis	Pattern	Anatomy	Position	Severity
Total	31	45	47	5	30
Samples	airspace disease, pleural effusion	opacity, atelectasis, degenera- tive bone, mass	diaphragm, esophagus, heart, heart ventricles	bilateral, left, pos- terior, right	acute, chronic, healed, patchy

Table 1.3: MeSH terms categorised into diagnosis, patterns, anatomy, position and severity labels.

### 1.5.2 Imperial College Local Hospital Brain DWI

The brain DWI dataset consists of 1,226 3D DWI scans and corresponding radiological reports of acute stroke patients collected from local hospitals. All the images and reports were fully anonymised and ethical approval was granted by Imperial College Joint Regulatory Office. The scans were obtained from three different scanners (Siemens) with the following acquisition

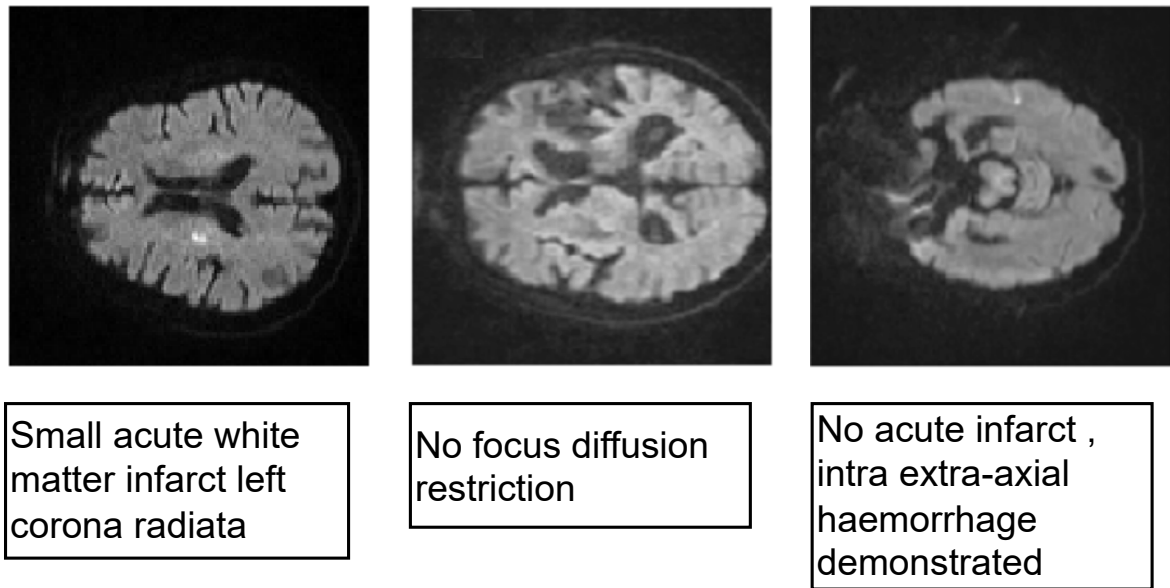


Figure 1.3: Samples of three unique brain DWI exams, shown as slices with respective filtered reports. Slices were selected based on maximum ischaemic infarct segmentation area using the segmentation network of Chen et al. [5].

parameters: field strength: 1.5-3 T; slice thickness: 5 mm; slice spacing: 1.0-1.5 mm; pixel size in x-y plane:  $1.40 \times 1.40$  or  $1.80 \times 1.80$  mm; matrix size:  $(19-23) \times (128 \times 128)$  or  $(192 \times 192)$ ; field of view:  $230 \times 230$  or  $267 \times 267$ ; echo time 90-93 ms; repetition time 3200-4600 ms; flip angle 90; phase encoding steps: 95-145. The scans were pre-processed according to the steps outlined in [5]: images were resampled into uniform pixel size of  $1.6 \times 1.6$  mm, and pixel intensities were normalised to zero mean and unit variance. The number of slices per image varies between 7 and 52, and the slice dimensions are  $128 \times 128$ .

Each report was parsed by a clinician to extract 1–2 sentences summarising the presence or absence of the pathology and its location within the brain. These filtered reports contained between 1 and 78 words, with an average of 16.7 and standard deviation 9.8. In addition, each exam was assigned a diagnosis label as part of hospital protocol: 54% were diagnosed ‘no acute infarct’, 46% were diagnosed ‘acute infarct’. The remaining, which made up a total of  $<1\%$  and included diagnoses such as ‘unknown’, ‘haematoma’, ‘tumour’, were removed, leaving a total of 1,177 exams. A sample of three brain DWI slices with their respective filtered reports are displayed in Figure 1.3.

A 2D subset of acute and non-acute (normal) slices was created from the 3D images. The acute set was created by first using the brain ischemia segmentation network developed by Chen et al. [5] to segment the images labelled with acute ischemia. The output of the segmentation network are pixel-wise probabilities. A simple heuristic was applied to the segmented slices to select the ones more likely to have a visible acute ischemia: segmentations were thresholded at 0.8, and slices were selected as having acute ischemia if the total area of the segmented ischemia was  $>10$  pixels. Most 3D images contained multiple slices that fit the criteria, and so the reports were duplicated for each slice, and each slice-report pair were treated as an instance. For the normal set, slices were sampled from the non-acute labelled images according to the same axial plane distribution as the acute set. The full subset sampling procedure is illustrated in Figure 1.4.

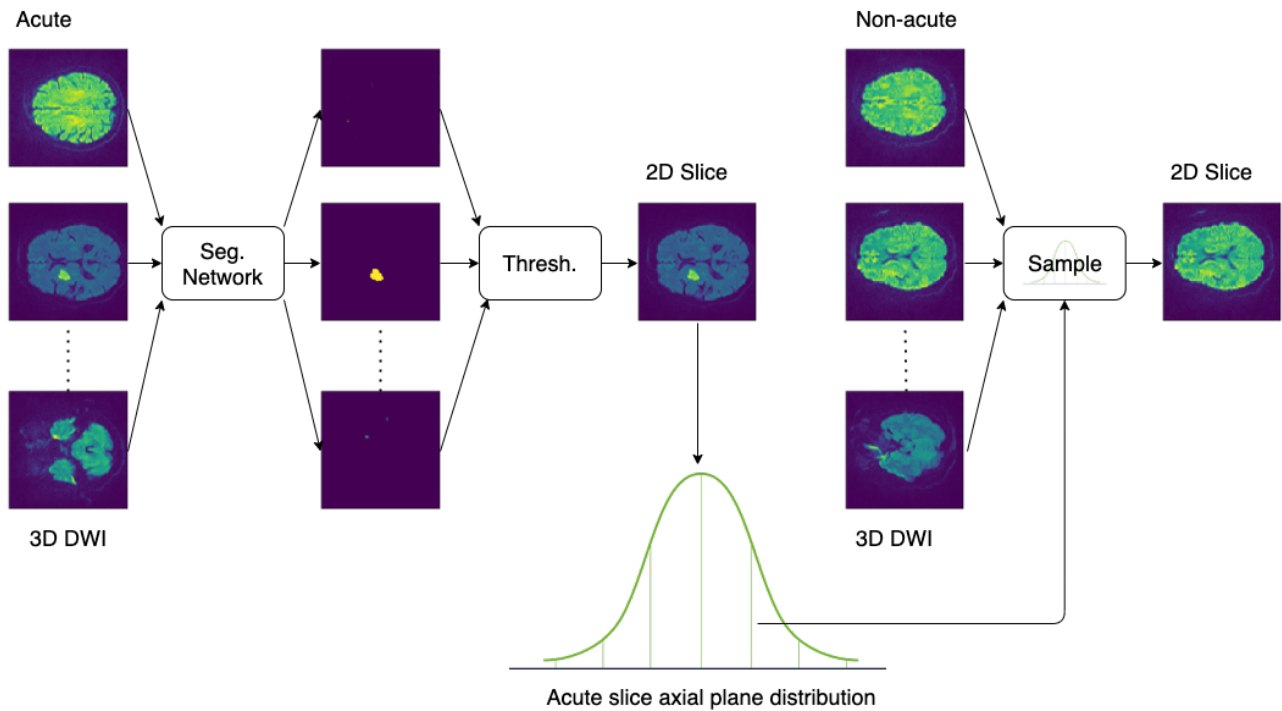


Figure 1.4: 2D Brain DWI subset procedure: 3D images are passed through a brain ischemia segmentation network of Chen et al. [5]. Slices fitting a heuristic thresholding criteria are selected as having a visible acute ischemia, and non-acute slices are sampled from normal brain images according to the same axial plane distribution as the acute slices.



## 1.6 Evaluation Metrics

As with natural image captioning tasks, the generated reports are evaluated against the true (human-generated) reports using BLEU [43] and ROUGE [44] scores, a type of modified n-gram precision and recall metrics developed originally for evaluating machine translation. These metrics measure the degree of word, or n-word, overlap between the true and predicted sentence(s), and treat all words as having equal contribution to the overall metric. It is generally considered to correlate well with human evaluation for accuracy and coherence of predicted against ground-truth sentences. When considering what is most important to accurately predict (and therefore report on) from a radiological image, it is far more important to correctly identify any markers in the image that may suggest the presence of a disease, and so not all generated words should be treated as having equal contribution to the overall metric. Therefore, in addition to evaluating the degree of overlap between the true and predicted radiology reports using established metrics such as BLEU and ROUGE, they are also evaluated on recall and precision of the predicted disease, anatomy, location and severity (DAPS) terms. These terms were categorised using radiological text tagging tools and manual checking. The DAPS metric was developed in this thesis to specifically address the challenge of assessing the quality of radiological reports irrespective of the quality of the grammar. This metrics provides a more granular assessment of what the report generation model either succeeds or fails to capture from the image.

### 1.6.1 BLEU

BLEU (Bilingual Evaluation Understudy) [43] is a type of modified precision metric that evaluates how closely a model generated text matches that of the (human generated) ground truth. It was developed for evaluating the quality of machine translation and has been reported to have high correlation with human judgement. It has since been used to evaluate image caption generation, text summarisation and speech recognition. The BLEU score is always a number between 0 and 1, with numbers closer to 1 representing a closer match to the ground truth, or reference, text. A unigram BLEU score, or BLEU-1, is the modified precision based on single

word matches, penalised by candidate sentence length in comparison to reference length. For instance, although a candidate sentence [‘the cat’] matches [‘the cat sat on the mat’] when calculating word-level precision, this would favour shorter sentences, therefore a ‘brevity’ penalty is applied to lower the score of a candidate sentence if it is shorter than all the reference sentences. Similarly, bigram, 3-gram and 4-gram individual BLEU scores can be calculated. Typically, these scores are taken cumulatively by taking a geometric mean of scores up to BLEU n-gram.

### 1.6.2 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [44] calculates n-gram precision, recall and F1 scores without modification, so is a good complement to BLEU. Whilst BLEU penalises shorter candidates, the common implementations of ROUGE do not, so it is important to consider both during evaluation. Additionally, n-gram ROUGE calculations are not cumulative, so BLEU n-gram implementations are preferred when evaluating for the overlapping of larger n-grams.

### 1.6.3 DAPS

Both ROUGE and BLEU were developed for evaluating machine translation and so have some limitations when applied to specific tasks where the quality of the translation, in our case, the translation of image encodings to text, may not have the same requirements. For instance, when generating a report from a radiological image, having an accurate pathology prediction is more important than an accurate severity prediction. For this reason, four categories were identified to be non-intersecting: disease, anatomy, position and severity, or DAPS. This is not exhaustive as some words will not fall in any category, but it provides a breakdown to the precision/recall metrics of BLEU and ROUGE. The categorisation of terms into DAPS is done in a dataset specific way with a combination of online radiological ontologies, medical text tagging tools such as MetaMap [45] and manual checking. Recall, precision and F1 is calculated in the same way as is done for classification, where each unique term is treated as

a class. A disease recall, precision and f1 score is then an average over all the scores of all the disease terms, and the same is done for terms under anatomy, position and severity, resulting in averaged precision, recall and F1 scores for each of the DAPS categories.

# Chapter 2

## Background

### 2.1 Theoretical Background

The algorithms used in this thesis for report generation, concept extraction and latent space learning are primarily neural networks. This chapter starts by covering the basic building blocks of recurrent and convolutional neural networks, which are the most common networks used to model natural language and images respectively. Image captioning is considered as a problem of encoding images using convolutional neural networks, and decoding image representations into natural language using recurrent neural networks. These are referred to as encoder-decoder architectures, and form the basis of the approaches considered in Chapter 3. Following this are the methods for medical concept extraction and representation learning used in Chapter 4 and Chapter 5, which are a combination of ontological tools, clustering algorithms and neural networks. Lastly is an introduction to representation learning, specifically through autoencoder networks as these form the basis for image representation learning and image reconstruction used in Chapter 6.

### 2.1.1 Building Blocks of Encoder-Decoder Frameworks for Image Captioning

#### Brief Introduction to Neural Networks

Introduced in the 1940's by McCulloch and Pitts [46], the ‘threshold logic’ computational model for neural networks was inspired by propagation of signals through the brain via activations of neurons. The linear threshold unit model for a neuron was expanded into a learning algorithm with the creation of the perceptron in 1958 by Rosenblatt [47] for the purpose of pattern recognition. The simplest single-layer perceptron model is illustrated in Figure 2.1. The output of the perceptron  $Y$  is calculated from the input vector  $\mathbf{x}$  as follows:

$$Y = f\left(\sum_{n=1}^N x_n \cdot w_n - \theta\right) \quad (2.1)$$

where  $w$  is a vector of weights,  $\theta$  is the bias term and  $f()$  is a step function such that:

$$f(x) = \begin{cases} 1, & \text{if } w \cdot x - \theta > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2.2)$$

In modern perceptrons, this function is more commonly replaced by the sigmoid function. Weights are randomly initialised, and the output  $Y$  is calculated for each training example  $i$  in the set. The weights are then updated as follows:

$$w_n(t+1) = w_n(t) + (Y_i - D_i)x_{n,i} \quad (2.3)$$

where  $D_i$  is the desired output.

Being a linear function of input signals, the perceptron is limited to classifying input vectors only if they are linearly separable. This observation was highlighted by [48], which, along with inadequate processing power of computers at the time, slowed the research in neural

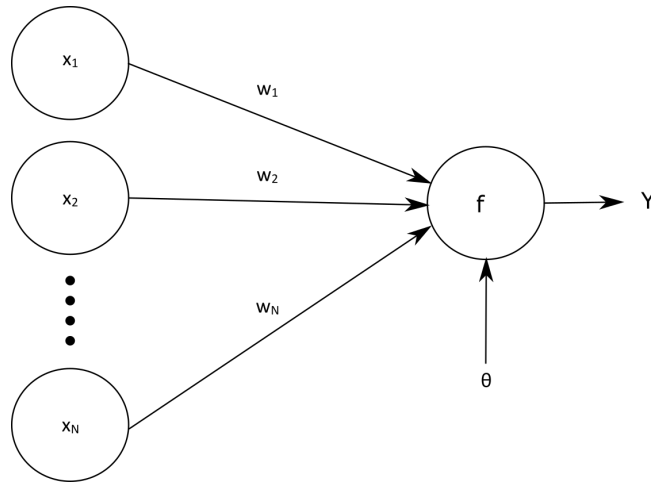


Figure 2.1: Single-layer perceptron

network. Interest picked up again with the introduction of the backpropagation algorithm by [49], which solved the problem of training multi-layer neural networks. This effectively solved the problem of non-linear separability as the universal approximation theorem [50] states that multiple neural networks stacked on top of each other (creating at least one hidden layer) can approximate a variety of continuous functions.

The backpropagation algorithm is a method for learning the weights in a neural network. Unlike in standard gradient descent where all the weights are updated as a function of error, backpropagation computes the loss function with respect to each weight by chain rule. Error on the output is calculated using a differentiable loss function, e.g. mean square loss, and is propagated backwards to the parameters. In this way, weights are adjusted based on how much they contribute to the error.

## Convolutional Neural Networks

With growing computational power through the use of graphics processing units (GPUs), and a growing digital record of images, it became possible to train multi-layer networks, or deep neural networks, for the purpose of visual recognition. The most prominent advance in image classification came about in 2012, when [51] beat the state-of-the-art image classification algorithms by large margin through the use of a deep convolutional neural network (CNN) architecture.

CNNs are multilayer neural networks consisting of one or more convolutional layers. They have been particularly successful in image recognition as they are able to learn translation invariant features by exploiting local correlation. Its design was inspired by overlapping receptive fields in the human visual cortex which act as local filters. In the same way, the convolution layer is composed of learnable filters, or kernels, which slide across the input image. The layer computes the dot product between the filter and the input in the receptive field (convolution). These convolution layers are normally followed by max-pooling (down-sampling) and rectified linear unit (ReLU) layers, ending with a fully connected layer. An example convolutional neural network with two hidden layers is illustrated in Figure 2.2.

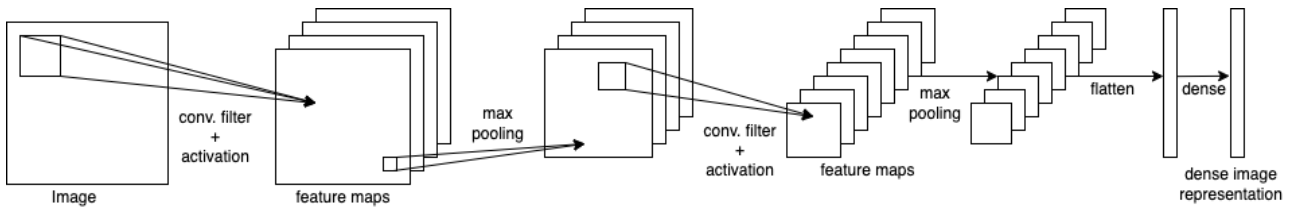


Figure 2.2: Convolutional neural network with 2 hidden layers.

Several convolutional layers, ReLUs and max-pooling layers can be stacked together into a network, such as the LeNet-5. Introduced in 1990 for the purpose of digit recognition [52] and popularised in 1998 [53], LeCun’s model was capable of classifying digits under various spacial transformations. In addition, the MNIST dataset, which they used to evaluate the performance of their model, became the standard benchmark for digit recognition tasks.

The ImageNet database was later created by [54] to benchmark natural image classification performance. In the 2012 pivotal paper, Krizhevsky et al. [51] showed that by training a deep convolution neural network on multiple GPUs and employing Dropout [55] to prevent over-fitting, they were able to almost halve the error rate from 36.7% (state-of-the-art) to 15.3% on the ImageNet dataset. Since then, convolution neural networks have been the main focus of research on image recognition tasks.

## Recurrent Neural Networks

Introduced in the 1980s, recurrent neural networks were designed with directed connections between neurons and are thus able to retain a ‘memory’ of past inputs. They are most commonly used to model temporal problems, such as language translation, speech recognition, and image captioning. Recurrent neural networks are able to model sequential information by preserving past information in an internal hidden state  $h_t$ . The hidden state and output are calculated as follows:

$$\begin{aligned} h_t &= \phi_1(W^{(hx)}x_t + W^{(hh)}h_{t-1}) \\ y_t &= \phi_2(W^{(yh)}h_t) \end{aligned} \tag{2.4}$$

where  $\phi_1, \phi_2$  are activation functions, typically logistic sigmoid or tanh, and  $W^{(hx)}$ ,  $W^{(hh)}$  and  $W^{(yh)}$  are weight parameters that are shared temporally. They are best visualised unrolled through time, Figure 2.3.

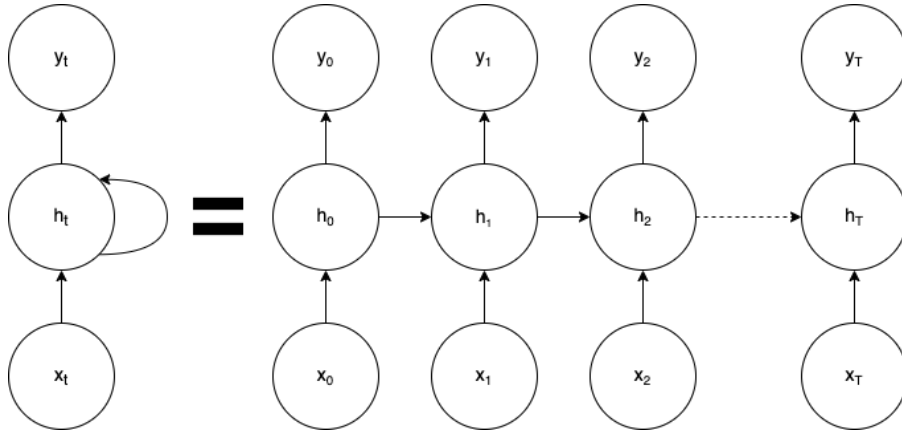


Figure 2.3: Recurrent neural network, rolled and unrolled.

Recurrent neural networks are trained by an extension to backpropagation, called backpropagation through time (BPTT) [56]. Time, in this case, is defined as a series of nested functions. Due to this process, RNNs suffer from vanishing and exploding gradients: small gradients get increasingly smaller, and larger gradients explode. One way to mitigate this problem is to use fewer time steps, or through the use of LSTMs.



**Long Short-Term Memory (LSTM)** [57] models are widely used in machine translation [58, 59, 60] and natural image and video captioning [61, 62, 10] due to their ability to capture long-term dependencies, and to reduce the problem of vanishing gradients in vanilla RNNs. Each LSTM unit has three sigmoid gates to control the internal state: *input*, *output* and *forget*. At each time step, the gates control how much of the previous time steps is propagated through to determine the output. The forget gate takes the current input and previous hidden state and decides whether to keep the information from the previous time stamp or forget it. The input gate does the same but quantifies the importance of the current input (new information). The cell hidden state  $h(t)$  is updated based on the input and forget gates, and the output gate determines the value of the next hidden state  $m(t)$ . For an input word sequence  $\{x_1, \dots, x_n\}$ , the internal hidden state  $h_t$  and memory state  $m_t$  are updated as follows:

$$\begin{aligned}
i_t &= \text{sig}(W^{(ix)}x_t + W^{(ih)}m_{t-1}) \\
f_t &= \text{sig}(W^{(fx)}x_t + W^{(fh)}m_{t-1}) \\
o_t &= \text{sig}(W^{(ox)}x_t + W^{(oh)}m_{t-1}) \\
h_t &= f_t \odot h_{t-1} + i_t \odot \tanh(W^{(hx)}x_t + W^{(hm)}m_{t-1}) \\
m_t &= o_t \odot \tanh(h_t)
\end{aligned} \tag{2.5}$$

where  $x_t$  is the input at time step  $t$ ,  $W^{(hx)}$  and  $W^{(hm)}$  are the trainable weight parameters, and  $i_t$ ,  $o_t$  and  $f_t$  are the input, output and forget gates respectively.

## Encoder-Decoder Networks

Automated caption generation draws on both computer vision and natural language processing techniques of image and text representation. RNNs have been shown to generate human-like text by training on a large corpus, such as Shakespeare [63] and Wikipedia [64]. Given a starting token (a character or a word), these models predict which tokens are likely to follow. RNN language generation models have been applied to tasks such as machine translation whereby the language generation is conditioned on a representation of a word or a sentence in another

language [58, 59, 60]. The words and sentences in one language are encoded into a single representation using one RNN network, and decoded using another.

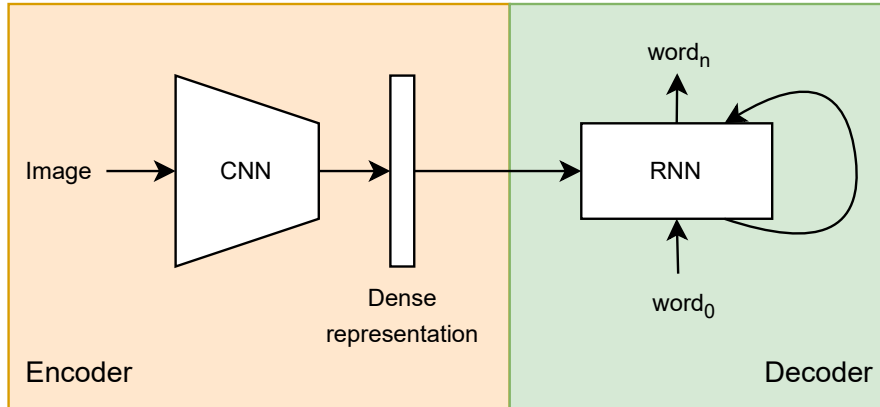


Figure 2.4: Encoder-decoder configuration for image captioning.

This idea of encoder-decoder networks has been extended to include encoding other data representations to be decoded into text: spatio-temporal data to generate pollen forecast summaries [65], weather forecasts [66], summarising electronic medical records [67], text summarisation [68], and image caption generation [69, 70, 62]. Figure 2.4 illustrates a simplified configuration of an encoder-decoder used for image captioning: a convolutional neural network is used to encode the image features into a dense representation, this representation is passed to a recurrent neural network, which decodes the information in the image and generates a textual description. The various methods and architectures of encoding the image information, passing the image features to the RNN, and generating the outputs will be discussed further in the literature review on deep learning methods for image caption generation.

### 2.1.2 Medical Concept Extraction and Word Representation Learning

As stated in the introduction, it may not be necessary to learn to automate the generation of radiology reports to match those of a radiologist, but rather, use the radiology reports to first extract medical concepts, such as pathology and location, and learn to detect these from the images. For instance, encoder-decoder networks introduced in the previous section can be used

to summarise long reports with a large vocabulary into short reports made up of key concepts and a much reduced vocabulary. However, this requires the radiology reports to be annotated with their respective task-specific summaries (for instance, if the summaries are to be used for image captioning, they must contain pathologies that can be seen in the images), and they must be done by an expert. This is typically not available for radiology reports taken directly from hospital databases as it is not part of standard reporting. Therefore, another method to consider is the use of off-the-shelf ontological tools that use pattern-matching to extract and categorise medical concepts.

Once medical concepts and phrases are extracted, they can be directly assigned to their respective radiological images as ‘labels’ for image classification, or as summary reports for report generation. The difficulty is in deciding whether to treat the concepts and phrases as single discrete labels, or continuous representations. Even with the help of ontological tools, there will be ambiguity over the meaning of certain words. For instance, whether a word is classified as a disease, a finding, or an abnormal pattern is heavily dependent on context that the tools do not have. Additionally, as described in Section 1.5.1, different diseases may present with the same abnormal patterns in a chest X-ray, in which case decisions need to be made on how to group similar concepts together. Hence, in addition to using the ontological tool MetaMap to extract medical concepts, two different methods of representing the words and phrases are used in order to group them together by meaning: term frequency-inverse document frequency (tf-idf) [71, 72] and word2vec [73, 74].

### **MetaMap: an Ontological Medical Concept Extraction Tool**

MetaMap [75] is a tool developed to map biomedical texts to the Unified Medical Language System (UMLS) Metathesaurus [76]. It was originally developed to improve retrieval of biomedical citations and abstracts from MEDLINE® (a large bibliographic database of clinical articles and journals). It uses natural language processing techniques to identify candidate phrases in the text and map them to their closest UMLS concepts, scoring each variant. It consists of several processing steps, summarised in the system diagram in Figure 2.5. Input text is first tokenised

and split into sentences, then put through MedPost - a stochastic part-of-speech (POS) tagger developed specifically for medical text tagging by [77]. The POS tagger uses a hidden Markov model where each part of speech is a state in the model and transition probabilities are based on bigram frequencies determined during training. This is followed by a lexical look-up and shallow parse where words and phrases part of the SPECIALIST lexicon (a component of the UMLS) are identified. Each word or phrase is accompanied by its acronym/abbreviation/synonym variants generated from the look-up. A candidate set of Metathesaurus strings containing these variants is retrieved and evaluated against the input text. Finally, a mapping is constructed by combining the various candidates of the separate phrases and choosing the highest scoring combined candidate mapping. Optionally, this is followed by word-sense disambiguation (WSD) [78] where candidates that are semantically consistent with the surrounding text are favoured.

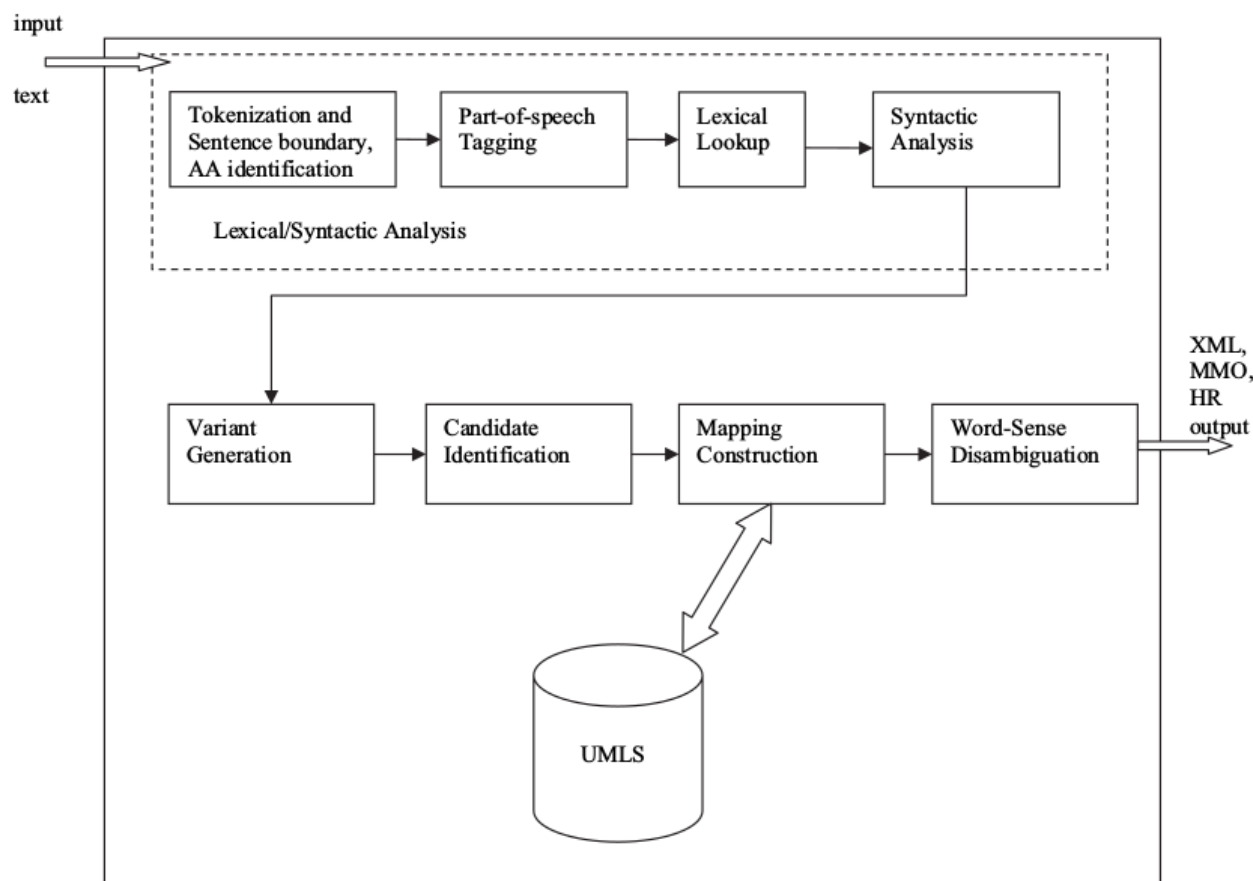


Figure 2.5: MetaMap System Diagram, reproduced from [79]

MetaMap has therefore become suitable for extracting and mapping terms in medical free text to common concepts for various purposes, such as clinical text analysis [80] and indexing and

retrieval [81, 82]. One of the main weaknesses of MetaMap is its WSD: many distinctive words or phrases in the Metathesaurus have common synonyms, for instance, *cold* can be used in the phrase *common cold* or *cold temperature*. Some of these phrases are manually suppressed during MetaMap processing, and the rest are put through the WSD algorithm, however, it still occasionally fails to determine the correct phrase or concept due to the algorithm being based on pattern-matching.

## Tf-idf

Term frequency-inverse document frequency is a method of generating weight vectors for documents by measuring how important each word is to the document in relation to the whole corpus [71, 72]. It is often used by search engine tools for scoring documents according to their relevance to a search query. The term frequency ( $tf$ ) measures the raw count of a term  $t$  in a document  $d$ , ( $f_{t,d}$ ) which can optionally be logarithmically scaled or nomalised by the frequency of the maximally occurring term  $t^{max}$  to prevent a bias towards longer documents:

$$\begin{aligned}
 tf(t, d) &= f_{t,d} \\
 &= 1 + \log(f_{t,d}) \\
 &= 0.5 + 0.5 \times \frac{f_{t,d}}{f_{t^{max},d}}
 \end{aligned} \tag{2.6}$$

The inverse document frequency term is a measure of how much ‘information’ that term provides and is calculated as follows:

$$\begin{aligned}
 idf(t, D) &= \log \left( \frac{N}{1 + n_t} \right) \\
 &= \log \left( \frac{n_{t^{max}}}{n_t} \right) \\
 &= \log \left( \frac{N - n_t}{n_t} \right)
 \end{aligned} \tag{2.7}$$

where  $N$  is the total number of documents in the corpus  $D$ , and  $n_t$  is the number of documents that contain the term  $t$ . Therefore, words that appear often in a particular document may not contribute information if they also appear in a large fraction of other documents in the corpus. The tf-idf score is calculated as a product:

$$tf_i df = tf(t, d) \times idf(t, D) \quad (2.8)$$

Hence, higher weights are assigned to words that occur frequently in the document and infrequently across the corpus.

## Word2vec

One major downside to tf-idf word representations is they fail to capture the semantic meaning of words since words are represented in a discrete way. Word2vec, developed by Mikolov et al. [73, 74], is an alternative method that creates continuous and semantically-meaningful word vector representation. It does this by training a 2-layer neural network on a large corpus of text and incorporating the context during training. The assumption is that words appearing in similar contexts will have similar meaning. Word2vec uses one of two mechanisms: continuous bag-of-words (CBOW) and continuous skip-gram. In CBOW, the context, or neighbouring words, are used to predict the target word. Bag-of-words refers to the assumption that the ordering and grammar of the context words does not affect the prediction. The skip-gram architecture predicts the context words for a given input word. The authors note that for the skip-gram architecture, increasing the range of neighbours surrounding the input word improved the quality of the word vectors, but increased computational complexity [73]. CBOW, on the other hand, was faster to train, but had worse performance on infrequent words.

An evaluation of word2vec word embeddings versus traditional count-based methods (including tf-idf) presented in [83] showed that word embeddings outperform on almost all tasks, including semantic relatedness, synonym detection, concept categorization, and analogy. Like tf-idf, they can be used to represent sentences, paragraphs and documents as fixed-length feature vectors,

but do so maintaining the order relationships between words [84].

### 2.1.3 Image Latent Space Learning through Autoencoders

Autoencoders are artificial neural networks that create dense representations of data in an unsupervised way by imposing a bottleneck in a network trained to reconstruct the data [85, 86]. They have been successfully applied in the medical imaging field for unsupervised image feature extraction [87, 88, 89], image denoising [90] and reconstruction [91, 92, 93, 94, 95, 95, 96, 97].

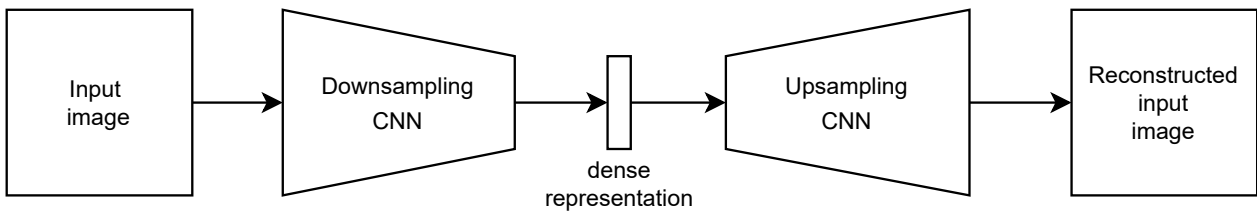


Figure 2.6: Autoencoder network for dense image representation through reconstruction.

An autoencoder is also a type of encoder-decoder network, where the encoder is used to ‘compress’ the data into a smaller representation, and a decoder follows the inverse steps in order to reconstruct the original data from the compression. A simplified illustration of an image autoencoder using CNNs is shown in Figure 2.6. The encoder portion of the network applies downsampling to the convolutional filter outputs, and the decoder portion applies the inverse upsampling such that the result at the output is the same shape as the input. By setting the loss of the full network to be the difference between the input and output image (e.g. mean squared error), a dense representation, or image latent space, is learnt at the bottleneck layer.

## 2.2 State-of-the-art in Image Caption Generation

### 2.2.1 Image Caption Generation in Computer Vision

The use of human generated visually descriptive text to infer the contents of an image has primarily been applied to image caption generation in the field of computer vision. The main

goal is a generative model that is able to describe the context of objects in an image using natural language, the way a human would. This has many potential applications, for instance, describing images or videos to people who are visually impaired. Although these applications do not necessarily translate into the medical domain, the way these models are trained using images and their free-text captions (as opposed to pixel-level labelling or classification) can give us ideas of how we can incorporate radiological reports to train models that can predict the context of diseases in radiological images.

## Template-Based Models

Earlier models of image caption generation relied on linking template-based language models to objects and spatial contexts in the image. For instance, the approach by Farhadi et al. [6] was to map images and text into an intermediate space they term ‘Meaning’: a triplet of object/action/scene, illustrated in Figure 2.7. Their problem is framed as a Markov Random Field (MRF), where nodes are object, action and scene, and edges correspond to binary relationships between nodes. Predicting these triplets in a discriminative way requires images to be manually labelled with their meaning triplets. For this purpose, they created the UIUC PASCAL<sup>1</sup> dataset. Sentences are generated by searching a pool of sentences for one that closely matches the image triplet. BLEU evaluation was only done on mapping images to meaning, and not images to sentences, but human evaluators agreed that, on average, at least one sentence generated per image was deemed ‘accurate’.

A similar approach is used by Kulkarni et al. [98] on the same dataset. Their model (‘BabyTalk’) identifies objects, modifiers and spatial relationships in an image, ‘smooths’ using statistical priors based on sample texts, and uses these results to generate sentences using an N-gram language model (a conditional probability distribution of N-word sequences) and a template with linguistic constraints. Their results showed that the N-gram language model scored higher on BLEU (25) than the template based model (15), however, the template-based model generated more coherent sentences (based on human judgement). This reveals the limitation of

---

<sup>1</sup><http://vision.cs.uiuc.edu/pascal-sentences/>



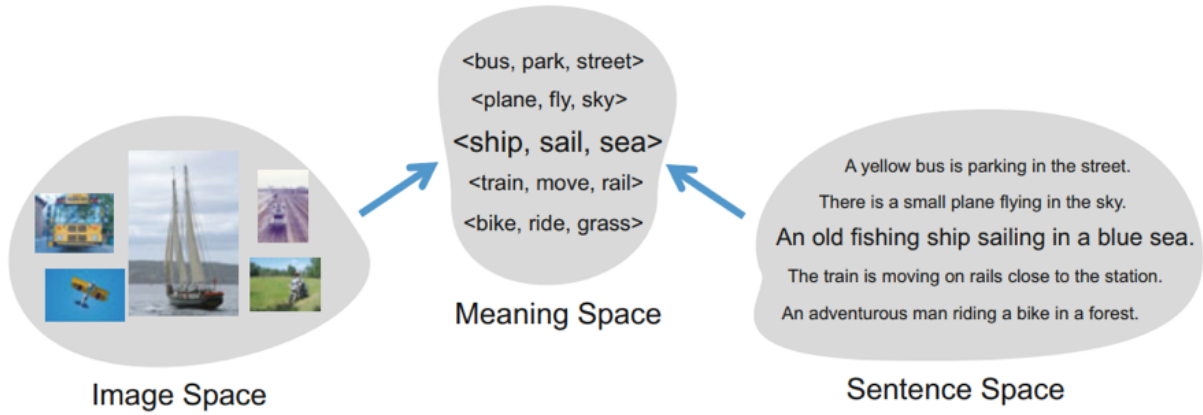


Figure 2.7: Illustration of ‘Meaning’ space triplet of object/action/scene, reproduced from Farhadi et al. [6]. Framed as a Markov Random Field, where each object, action, scene is a node and edges are relationships between nodes.

their models: although the language model is better able at generating descriptive sentences in line with what a human would describe, it is ineffective at making them grammatically correct. On the other hand, although the template-based model is generally grammatically correct, its constraints limit it to basic descriptions of object/modifier/preposition, which do not always cover the full context of the image.

In contrast to generating word captions, Fidler et al. [7] proposed a holistic Conditional Random Field (CRF) model that uses text to improve object recognition and semantic segmentation, illustrated in Figure 2.8. They introduce object detection potentials to the CRF that use text to re-rank candidate bounding boxes by, for instance, penalizing bounding box configurations that do not match the estimated cardinality from the text, and using prepositions in the text to boost certain bounding box configurations that are consistent with the spatial locations. As the goal is scene understanding as opposed to sentence generation, they evaluate their model on semantic segmentation on the UIUC PASCAL dataset using the standard VOC IOU measure and achieve a score of 36.4% (12.5% above state-of-the-art at the time). This model was developed further by the same group as part of a framework used to generate multi-sentence descriptions of images with multiple objects and complex interactions in [99].

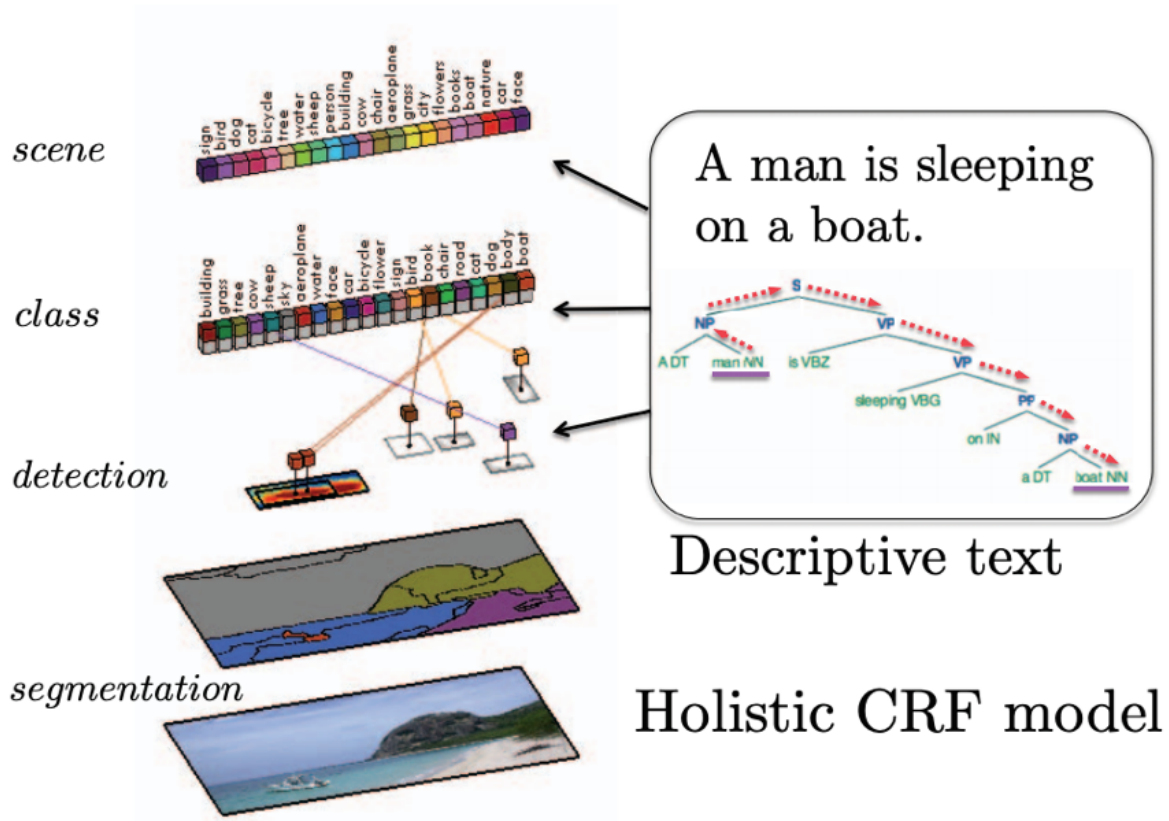


Figure 2.8: Textual descriptions to inform semantic segmentations through a conditional random field, reproduced from Fidler et al. [7]. Object relations are extracted from text and used to re-rank candidate bounding boxes for object detection.

## Deep Learning Models

More recently, interest has moved to the combined potential of convolutional neural networks and recurrent neural networks for describing images using natural language [61, 8, 100, 9, 10]. The advantage of using neural networks for caption generation is that the model is not constricted by hard-coded language templates and is able to learn more freely from the training data.

One of the first of such models was a multi-modal neural language model from Kiros et al. [101] that took inspiration from multi-modal learning. Their modality-biased log-bilinear model (MLBL-B) is a natural language model conditioned on a different modality, in this case images, by incorporating image features as an additive bias. Later, they improved on caption generation and image ranking tasks by using an encoding-decoding model inspired by machine translation [61]. A joint image-sentence embedding is learnt using an LSTM, and a structure-content neural

language model (SC-NLM) is used to decode the embeddings into captions.

A different take on the encoder-decoder model, Neural Image Caption (NIC), was proposed by Vinyals et al. [8] at the Google Brain team that used a CNN trained for image classification as an image ‘encoder’, using its last hidden layer as an input to an LSTM along with the text. The image and words are mapped into the same space, and input into the LSTM sequentially, as illustrated in Figure 2.9. The LSTM is trained by minimising the negative log likelihood of the correct word at each time step. They attribute this vast improvement over the BabyTalk model [98] primarily to the extraction of image features using deep learning.

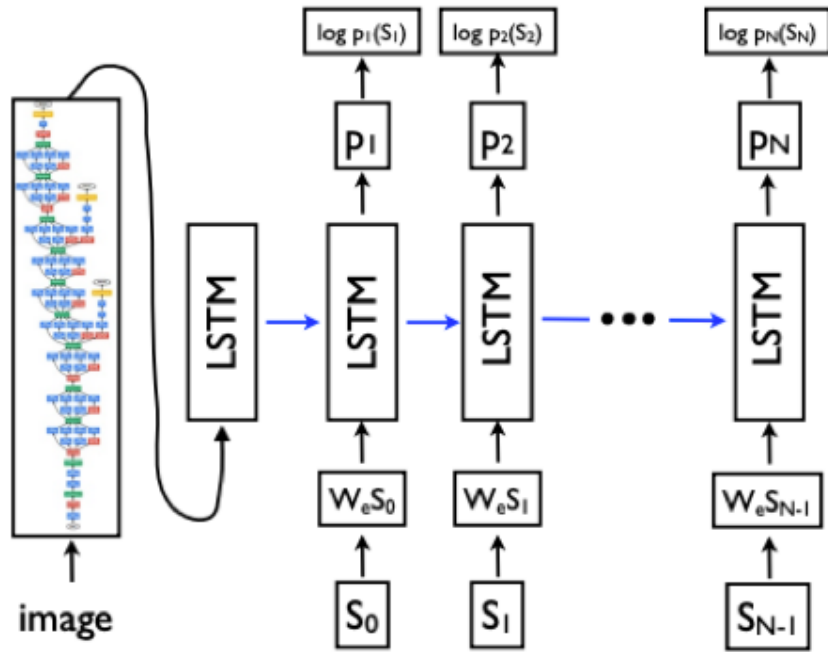


Figure 2.9: Neural image caption model, reproduced from Vinyals et al. [8]. CNN encodes image into a dense representation and input into the LSTM at time step -1. Images and words are mapped into the same embedding space.

Working in parallel, Karpathy’s model aimed to generate descriptions of image regions [100], similar in ambition to the holistic scene understanding model of [7]. However, instead of labelling images from a set of categories, they aimed for a richer understanding of image content. They first create image representations using a Region Convolution Neural Network, developed by [102]. A Bidirectional RNN is similarly used to compute word representations in the same  $h$ -dimensional vector space as the image representations. They then formulate their objective to

encourage aligned image-sentence pairs to have higher scores than misaligned pairs by defining the score between a sentence and image to be the sum over the dot products of image region vectors and word vectors. These alignments are used to create image annotations consisting of not just single words, but short, local descriptions. A Multimodal RNN is then used to generate sentences from these descriptions. For the task of image annotation and sentence generation, their model fell a little behind the NIC, which they attribute to the NIC's more powerful CNN (GoogLeNet) and more powerful sequence learner (LSTM), however, they did achieve state-of-the-art in image ranking tasks.

One major drawback of using vanilla RNNs for image captioning is that they struggle with long-term dependencies. If the image is introduced at the start of the sequence, consequent words are conditioned less and less on the image features. One solution may be to use LSTMs, which are able to 'forget' past, redundant information, but an alternative was proposed by Chen et al. [9]. Their recurrent neural network model was built on top of Mikolov's [103, 104] with an additional visual hidden layer that attempts to reconstruct visual features from previous words, and is thus able to retain a visual memory by propagating these visual features through each time step. This last layer can be ignored when generating sentences from images as the visual features are already known. Their model also allows for visual feature generation from text, though they do not propose an application to this.

Alternatively, instead of storing the visual information within the recurrent neural network, Mao et al. [69] proposed using the visual features as inputs along with every word at each time step. Each time step in their multimodal RNN consists of two word embedding layers, which encode both syntactic and semantic meanings of the words. These are: a recurrent layer, which is a multimodal layer that takes as inputs the word representation, recurrent layer output and image representation; and a softmax layer that generates a probability distribution of the next word. Their model allows for backpropagation to update the CNN part of the model as well as the RNN (though they did not have a chance to apply this due to a shortage of data). Had they done so, their CNN model may have improved with the knowledge of textual information.

For the most part, these methods use a single vector representation for the entire image. Cap-

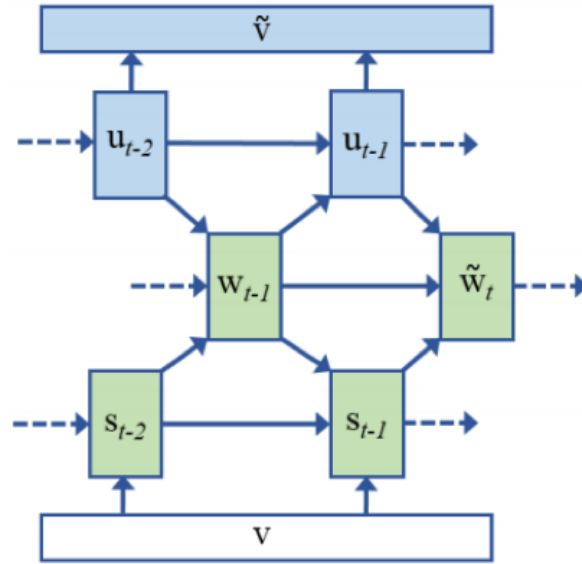


Figure 2.10: Mind’s eye: image captioning through visual feature reconstruction, reproduced from Chen et al. [9]. The green modules represent the RNN language model,  $\mathbf{v}$  is a vector of observed visual features, and  $\tilde{v}$  is the reconstruction of the visual features.

turing salient image features is a large part of image captioning, and Xu et al. proposed to use recurrent visual attention over the more descriptive, lower-level image representations for image captioning, illustrated in Figure 2.11 [10]. The Recurrent Attention Model (RAM) was first introduced by Mnih et al. [105] in order to reduce the computation required by traditional convolutional neural networks by learning to process only selected regions of interest instead of the entire image. The RAM is a recurrent neural network model that, at each time step, takes as input the ‘glimpse’ representation (given the image and location) and combines with the internal representation at the previous time step to produce the new internal state. Their model is non-differentiable, and so required the use of reinforcement learning in order to train. To tackle this, Xu et al. propose a ‘soft’ attention mechanism that, instead, learns a distribution over the vector representations, and is therefore differentiable and trainable via backpropagation [10].

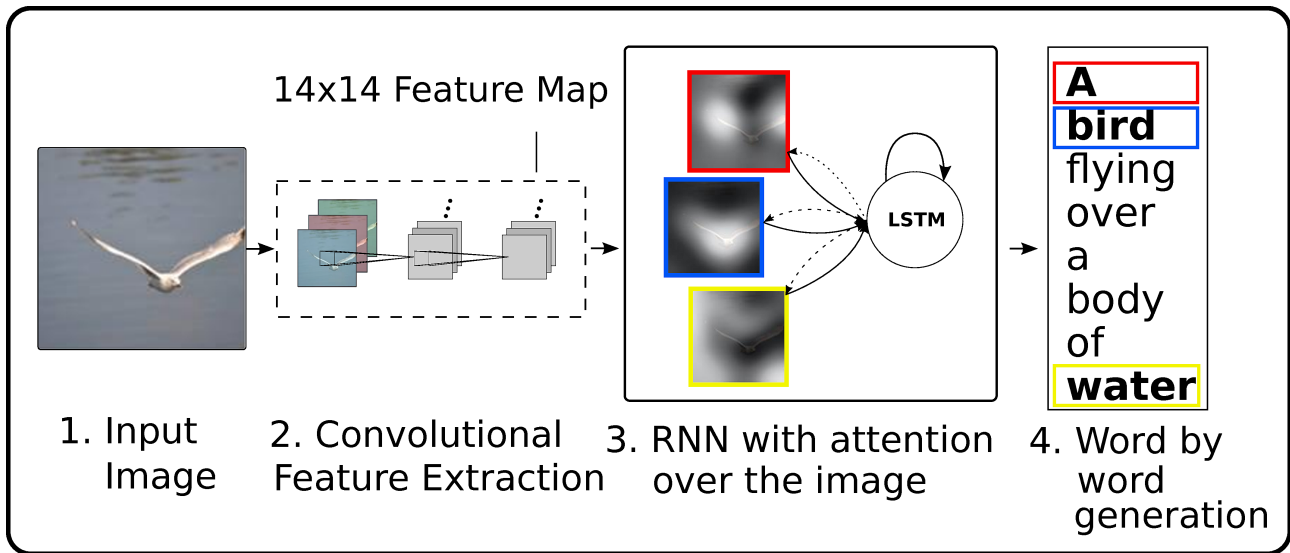


Figure 2.11: High-level illustration of recurrent attention model of Xu et al., reproduced from [10]. Instead of a single vector representation as the input to the language model RNN, they propose the use of the lower-level features maps. ‘Attention’ is characterised as the learned weights over these features, as well as the previously generated words.

## 2.2.2 Radiology Reports As Image Annotations

### Ontology-Based Concept Extraction

One common approach to using raw radiological reports in order to assist in supervised imaging tasks is text mining the reports for diagnoses and assigning them as image labels. These labels have been successfully applied to supervised multi-label classification [106] and weakly supervised localisation learning frameworks [107]. In these examples, a series of text processing techniques are applied to the reports for pathology extraction, including negation detection, and tools such as DNorm [108] and MetaMap [45], which map key words to a standardised vocabulary of clinical terms. However, other biological concepts in the reports, such as location, severity, and other visually descriptive features of the pathology are not taken advantage of.

The framework proposed by Schlegl et al. [109] attempt to take advantage of spatial semantic content of clinical reports and corresponding images for localised abnormality detection in optical coherence tomography (OCT). Image features were extracted using a CNN, and semantic target labels were extracted from the textual report using semantic parsing. Each report was reduced to  $K$  pairs of *[object class, spatial location]*. The CNN was then trained to predict

these semantic labels and tested by transforming the semantic prediction into voxel-wise class labels and comparing them to their voxel-wise labelled ground truths. They showed that the using semantic content as ‘weak labels’ to train a disease classifier results in improved performance over naive weakly-supervised learning (66.30% vs. 81.73% accuracy). They note that by training a CNN on semantic target labels, it was able to “learn abstract concepts of ‘location’”. However, their classifier performance is still not as good as one trained through fully supervised learning using voxel-level annotations (which achieved 96.98% accuracy).

## **Statistical Text Mining**

Statistical mining approaches have also been applied to extract labels for classification. For instance, Shin et al. [110] applied latent Dirichlet allocation for topic and sub-topic extraction to then be used for the classification of images into sub-topics. Sub-topics were made up of a collection of key-words which included pathologies, anatomy, imaging modality and severity, and so could be used to auto-generate key-word ‘reports’. Wang et al. [111] proposed clustering image embeddings and grouping their associated text reports. In both cases, sub-topics and cluster groups are only implicitly defined and depend on the number of topics/groups providing the lowest perplexity score, which can be a range of values. In addition, these are not generative models, therefore reports can only be selected based on nearest-neighbour methods from ones present in the training sets. Therefore, a generative learning approach in a similar style to image captioning has also been considered for radiology report generation

## **Automated Radiology Report Generation**

Image captioning models and learning frameworks are, to a lesser extent, being applied to medical images and their reports: from learning to automate MeSH annotations for chest X-rays [3], to leveraging reports in a dual-attention framework to improve features used for classifying histopathology images and to provide interpretability to the classification [112, 113]. In the latter paper, a structured report output is explored as a potential application of their dual CNN-LSTM classification network. In all these examples, manually created structured

reports are used for supervision as they are short and only contain visually-relevant information that can be extracted using a suitable CNN. These are, for instance, the presence of localised pathologies such as lesions and masses in the case of chest x-rays used in [3], and cell appearance such as crowding and mitosis in the case of the histopathology images used in [113].

The structured reports used by Shin et al.[3] were the MeSH annotations of the Indiana University chest X-ray dataset, described in Section 1.5.1. Their recurrent neural cascade approach starts by first training an RNN to generate the MeSH annotations conditioned on a dense image representation appended to the start. They then mean-pool the hidden state vectors of the RNN to obtain what they term a ‘joint image/text context vectors’. These vectors are grouped together with k-means to represent new labels, which are then used to fine-tune the CNN to create new image representations, and those representations are used to fine-tune the RNN. The training sequence is illustrated in Figure 2.12.

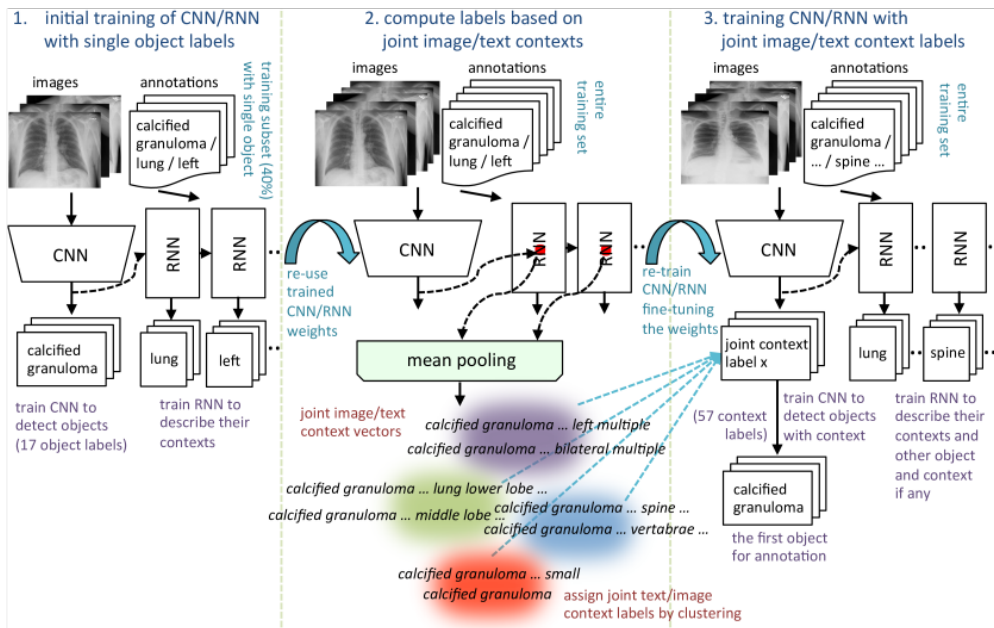


Figure 2.12: Recurrent neural cascade training sequence, reproduced from Shin et al. [3].

Learning a generative captioning model from raw textual radiology reports is a more difficult task due to the reporting of non-visually informative information, such as negation or uncer-



tainty over the presence of pathologies. Jing et al. [114] demonstrated how long textual reports with partially-correctly identified medical concepts can be generated by first training a multi-label CNN on the images and the Medical Text Indexer (MTI) tags identified in the original raw reports of the Indiana University chest x-ray dataset. The image features from the CNN and tag embeddings are used as inputs to a hierarchical LSTM for topic (sentence) and word learning.

Yuan et al. [115] use a similar technique of first training a multi-label CNN on extracted disease labels (though it is not mentioned how these are extracted), and then training a hierarchical encoder-decoder RNN network to generate the raw reports of IU-CX. They go a step further and use the medical concepts extracted from the reports using a UML extraction tool called SemRep to inform the report generation directly. The hierarchical decoder RNN is made up of two layers: one trained to generate sentence hidden states from the visual features (incorporating attention), and a second trained to generate word hidden states from the sentence hidden states and SemRep medical concept embeddings. They achieve s.o.t.a. BLEU in report generation when evaluated against previous image captioning techniques. However, reports can be very long and heterogeneous, and as BLEU treats  $n$ -grams equally, it may not be an appropriate metric for determining whether a pathology has been identified as present in the image, which is a far more important task to achieve if it is to be used in a clinical setting to assist radiologists.

## 2.3 Summary

The approaches to creating valuable CAD tools from radiology image-report datasets depend on the desired output. If the goal is to be able to automatically generate reports from new radiological images in the style of a radiologist, the most closely related approach in literature is that of image captioning. These approaches can be summarised by a general encoder-decoder framework that takes as input an image representation (or multiple representations), and decodes that representation into natural language. Convolutional neural networks and recurrent neural networks are uniquely suited to model images and text respectively, and so are the most

commonly used encoders and decoders.

On the other hand, if the goal is to automate the detection of diseases, without the need to frame them in a report (and therefore dispense the need for learning a language model), then there is a required intermediate step of extracting disease labels from the textual reports. There are many off-the-shelf ontological tools capable of this, though none designed with the specific task of extracting labels that can then be used for image classification. Therefore, a more common approach in literature has been a more data-centric one: the reports are mined using statistical methods in order to define the ‘disease’ clusters, and group images accordingly. This has the disadvantage that disease clusters are only defined implicitly, and require choices in heuristics that then make it difficult to apply the same methods to other datasets.

## Chapter 3

# Report Generation for Single and Multi-View Radiological Images

### 3.1 Introduction

Within the UK National Health Service (NHS), when a patient has a radiology exam such as an X-ray, computed tomography (CT), ultrasound or magnetic resonance imaging (MRI), a radiologist will make a report on the image or set of images. The purpose of the report is to summarise diagnostic findings and possibly recommend treatment, interventions or follow-ups. These reports are then stored, in their free-text format, within the hospital PACS or RIS along with the images. The PACS or RIS is therefore a valuable resource of large volumes of radiological images that have been interpreted by experts. This chapter investigates whether radiological exams gathered directly from a PCAS or RIS can be used as part of a supervised learning framework to predict pathological information from new, unseen radiological images.

A common approach in literature is to treat this as an image captioning task [3, 112, 116, 115] as there are many parallels: images are human-annotated using unstructured, natural language, and there is no limit to the vocabulary and no hard rules on number or type of concepts that are commented on. When writing radiology reports, there are still clinical protocols and standards radiologists must follow in terms of language and structure, but these vary across

institutions and are still subject to an individual’s training and interpretation. Hence, one major component of image captioning is learning a language model that is capable of capturing this highly variable and complex structure.

Another aspect of image captioning is the choice of image representations. Convolutional neural networks require large amounts of images for training, as many as thousands per class, something not available for radiological images. This number can be reduced through the use of transfer learning. A common method of generalised image feature extraction in literature is the use of a deep CNN pre-trained on a large and diverse image dataset, such as ImageNet. It has been shown that such transfer learning is effective for use on radiological images, even if the network has been trained on natural images [3, 116, 115]. Hence, this method is used to extract task-agnostic image features from radiological images in this chapter.

In the work of the cascaded report generation model of Shin et al. [3], chest X-ray feature vectors were extracted from a CNN network pre-trained on ImageNet, and pre-appended as a ‘word’ to the radiology report, made up of manual annotations using Medical Subject Headings (MeSH). A recurrent neural network was then trained to predict this sequence with iterative fine-tuning. Their model was trained and tested on the IU-CX MeSH dataset, filtered to contain only a single MeSH disease/description per image. The first half of this chapter uses this as a baseline technique that makes use of image representations from a pre-trained network and builds on it by:

1. Exploring single-view image-text encoder-decoder configurations to achieve the optimal diagnostic predictions in terms of both language (BLEU, ROUGE) and content (DAPS).
2. Improving training performance by incorporating image and text augmentation techniques, weighted sampling, dropout over image representations and word embeddings, and training end-to-end.
3. Extending the architecture to incorporate multi-view images by combining image view feature vectors through mathematical operations such as sum, max and concatenate.

4. Evaluating the optimal encoder-decoder configurations on more complex clinical data consisting of multi-view images and unstructured, free-text reports with multiple cases of pathology per image.

The second half of the chapter builds on this further by incorporating dynamic attention over the multi-view convolutional image features when generating the report. This approach has been successfully used for image captioning to capture local, salient features at each time-step during the generation process [10, 116, 115]. In the original natural image captioning implementation of Xu et al. [10], the attention weights are conditioned on the previous hidden state, and they interpret this as the networking learning ‘where to look next’ depending on the previously generated words. This method can be translated to multi-view radiological images since each generated word in a report depends on the features in a specific location within the image, possibly only seen in one of the image views. This additionally provides interpretability to the words being generated as the attention weights can be visualised over the images, and hence give some indication as to which locations in the image resulted in the prediction of abnormalities.

## 3.2 Datasets

In order to compare the performance of the radiological report generation architectures on single and multi-view exams, the models were trained and evaluated on the IU-CX and the ICH-Brain-DWI datasets. The IU-CX MeSH-sp-subset was made up of chest X-ray exams where the MeSH annotation of the PA view images reported the presence of a single abnormal pattern. This simplified dataset was used to first determine the optimal encoder-decoder configuration and hyperparameters based on a combination of BLEU, ROUGE and DAPS metrics. Then, the best performing model was trained and evaluated on the more complex IU-CX MeSH, IU-CX free-text and ICH-Brain-DWI-2D. The creation of the 2D subset of ICH-Brain-DWIs is described in Section 1.5.2. By evaluating the report generation model only on the 2D subset, the challenge of creating a specialised 3D brain DWI image encoder is circumvented, but will be addressed

later in Chapter 6.

### 3.2.1 Preprocessing - Chest X-rays

**Image preprocessing** All exams consisted of at least one image, however, image metadata did not contain image view labels. An image view classifier was made using a fixed-weight ResNet50 CNN model [117] pretrained on ImageNet [54] with a binary classification output instead of the multi-class output. It was trained on 40 randomly selected and manually labelled chest X-ray images, 20 PA and 20 lateral. Another 5 PA and 5 lateral view images were used for validating the trained model. The model was trained using early stopping on the validation loss, and was trained for 42 epochs. Final accuracy, precision and recall were 100% on the validation set. The trained model was then used to predict the views of the remaining chest X-ray images.

**Text preprocessing** Of the 3,955 X-ray exams introduced in Section 1.5.1, 876 are missing findings, 10 are missing impressions and 40 are missing both. The ones missing both were removed from the dataset, and for simplicity, findings and impressions were combined under one textual report, and indications were removed as they were not visual descriptions of features in the images. As the negation of pathologies was generally standard across the free-text reports, negation removal was performed using NegEx [118]: a processing package using regular expression (regex). The package identifies and tags negation signifiers such as ‘[tag]no[tag] [disease] present’ or ‘[tag]without evidence of[tag] [disease]’ and so a regex rule was made to remove phrases beginning with [tag] and ending in full stops.

Several preprocessing steps were then done on the reports and MeSH annotations. This involved lower-casing, removal of punctuation (except full-stops and commas), and non-alpha-numeric character removal. Stopwords such as ‘and’ and ‘the’ were removed from the reports and MeSH annotations, and remaining words were tokenised i.e. split into units of words separated by spaces and punctuation. A vocabulary was then created from the tokens that captured 99 percent of the content, with a unique vocabulary created for the free-text reports, the MeSH

	Vocab	Avg. s/e	Avg. t/e	STD t/e
IU-CX free-text	1309	3.0	23.9	16.7
IU-CX MeSH	118	2.1	7.2	7.0
IU-CX MeSH-sp-subset	95	1.0	3.3	1.5
ICH-Brain-DWI	1021	1.4	10.8	6.3

Table 3.1: IU-CX [2] statistics of free text reports (IU-CX free-text and ICH-Brain-DWI), MeSH (IU-CX MeSH) and MeSH single-pattern subset (IU-CX MeSH-sp-subset) annotations after processing. Avg. s/e refers to the average number of sentences per exam, avg. t/e is average tokens per exam, and STD t/e is standard deviation of tokens per exam.

annotations and the MeSH single-pattern subset annotations. Tokens not in the respective vocabularies were removed. This method of capturing 99 percent of the content meant that rare words, typos and misspellings were filtered out.

### 3.2.2 Preprocessing - Brain DWI

After some initial data clean-up outlined in Section 1.5.2, the brain DWI dataset was reduced to 1,177 exams, each consisting of a re-sampled 2D image of dimensions 128x128, a binary diagnosis of presence/absence of acute infarct, and 1–2 sentences summarising the findings in the image that pertain directly to the presence or absence of an infarct. Pre-processing steps on the reports were the same as for the chest X-rays: lower-casing, removal of non-alpha-numeric characters, tokenization and vocab reduction. Statistics of the processed reports of the IU-CX and ICH-Brain-DWI are listed in Table 3.1.

## 3.3 Static Image Embedding Models

### 3.3.1 Related Work

There are multiple image captioning frameworks that are well suited for this type of task, for instance, the Neural Image Caption (NIC) [8] model and the attention-based captioning model [10]. The NIC model uses a pre-trained deep CNN network to encode the image into a single, static, feature vector, which is used as the initial ‘word’ in the caption sequence. Subsequent

words are embedded into the same embedding space. An LSTM network is then trained to generate this sequence. This forms the basis of the recurrent neural cascade model of Shin et al. [3], where the CNN network used to generate image embeddings is fine-tuned first on selected disease labels, then again on new ‘labels’ based on the joined image-context vectors. These are computed by mean-pooling the hidden state vectors of the LSTM at each time step over the entire sequence, and then selecting new disease-context labels using dimensionality reduction and k-means clustering. This method has the following main disadvantages: errors in the initial training stages are propagated to the final training stages, PA and Lateral-view images are considered as separate instances even when they are part of the same exam, and the model cannot be trained end-to-end. Hence, a simplified NIC framework was considered as a baseline that can be trained end-to-end and extended to incorporate multi-view images.

### 3.3.2 Model Architectures

The group of models described here will be referred to as the Static Embedding Report Generation, or SERepGen models. The baseline model was trained as follows: a pre-trained CNN network was used to extract image features from a single view, and then pre-pended to the sequence of words in the report. An LSTM network was then trained to generate this sequence by predicting the next word. This model is referred to as the SERepGen-init as the report generation model is initialised with the static image embedding.

Initialising the model with the image embedding means treating it as the initial ‘word’ in the sequence, and so subsequently generated words are conditioned less and less on the first word. One option to mitigate this is to inject the image embeddings at the input to the LSTM at every time step, here referred to as the SERepGen-inject model. In this model, the LSTM is acting as both an encoder of the linguistic and visual information. Alternatively, the image embedding can be merged (through concatenation for instance) at the output of the LSTM such that the LSTM is acting only as an encoder of reports, and the decoder is a dense layer taking as input this merged, multi-modal representation. This model will be referred to as the SERepGen-merge. There has been some work to suggest that the role of the recurrent



neural network in a caption generator is as an encoder rather than a generator [119], and this is explored more in this chapter.

This approach has several advantages over the recurrent cascade training method of [3]: each of the SERepGen models can be extended to incorporate multi-view images by combining the feature vectors of multiple views into a single vector through mathematical functions such as sum, max and concatenate. The models can also be trained end-to-end by training unique instances of the CNN image encoder network, one for each view. Finally, I demonstrate how the inject and merge approach differ from initialisation, and that understanding the role of the recurrent neural network is important when trying to model more diverse language, such as in the case of free-text radiology reports.

For all the SERepGen models, an image embedding,  $\mathbf{im}_i = \text{CNN}(I)$  where  $\mathbf{im}_i \in \mathbb{R}^g$  is extracted from the final spatial-average pooling layer of a pre-trained CNN. The words in the report sequence are fed through an embeddings layer of the same dimension. An LSTM RNN is used to model the report word sequence. In order for the report generation model to be conditioned on the input image, three static embedding report generation (SERepGen) architectures are considered:

1. SERepGen-init: The image embedding is projected into the same embedding space as the word embeddings via a dense transition layer:  $\mathbf{im} = \text{relu}(W^{(dg)}\text{CNN}(I))$ . The image embedding is concatenated with the word sequence and thus treated as the initial ‘word’ in the report sequence.
2. SERepGen-merge: The image embedding is projected via a dense transition layer into a fixed embedding width and combined with the output of the recurrent layer through either concatenation or summation operation, and passed to the decoder  $dec$ :

$$\mathbf{dec}_t = \text{relu}(W^{(z)}(o_t * \text{relu}(W^{(dg)}\text{CNN}(I)))) \quad (3.1)$$

where  $*$  represents concatenation or summation and  $W^z$  are the weights of the decoder.

3. SERepGen-inject: The image embedding is projected via a dense transition layer into a fixed embedding width and combined with the input of the recurrent layer through either concatenation or summation operation, and passed to the encoder  $enc$ :

$$enc_t = \text{relu}(W^{(a)}(x_t * \text{relu}(W^{(dg)}\text{CNN}(I)))) \quad (3.2)$$

where  $W^a$  are the weights of the encoder.

The model architectures are illustrated in Figure 3.1. For all models, the decoder outputs are passed to the prediction layer  $s(t) = f(W^T x_t)$  where  $f$  is the softmax function.

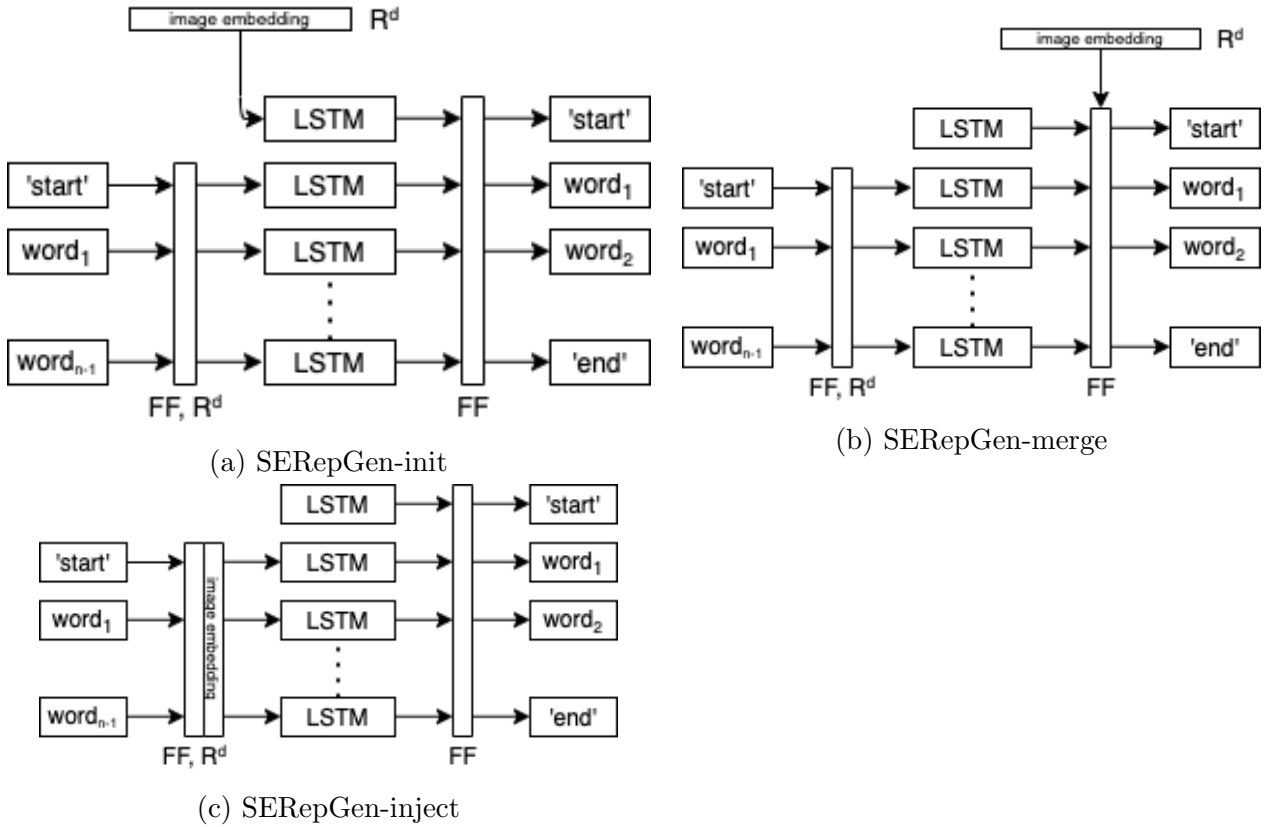


Figure 3.1: Image-report learning architectures using a single static image embedding as an aggregate representation of all input image views.

The extension to multi-view is the same for all SERepGen models. For an exam consisting of multiple views  $[V_1, V_2, \dots, V_K]$ , a mathematical function of all the features is aggregated across the image views to create a fixed-size input  $im_i = f(\text{CNN}(V_1), \text{CNN}(V_2), \dots, \text{CNN}(V_K))$ . The multi-view static image embedding computation is illustrated in Figure 3.2.

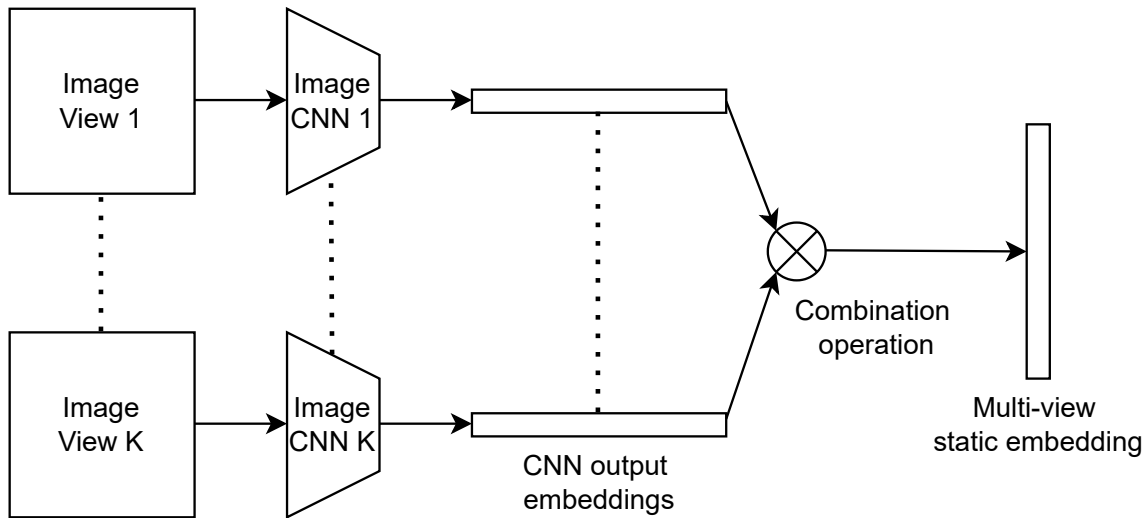


Figure 3.2: Multi-view static embedding. All multi-view images from a single exam are passed through a pre-trained classification CNN where the classification layer has been removed. The outputs are then considered the image representations and are combined into a single static vector through a combination operation, such as max, concat, sum.

Augmentation	Parameters
Image rotation	$0.01 \times 2\pi$
Image translation	$h \times 0.05, w \times 0.05$
Image cropping	$224 \times 224$
Sentence shuffling	random

Table 3.2: Image and text augmentation parameters.

### 3.3.3 Experiments

**Data balancing and augmentation** Image augmentation was done on-the-fly with random rotation, random translation, and random cropping, parameters listed in Table 3.2. Text augmentation, unlike image augmentation, must be very domain-specific if we are to maintain the original meaning. Randomly shuffling sentences in the text reports will maintain the context surrounding the pathologies and provide extra instances where pathologies are listed in different order. During training, sentences in the text reports were randomly shuffled during sampling. The multi-pattern MeSH annotation *disease/description* pairs were also randomly shuffled during sampling. For the single-pattern MeSH captions, the training samples were weighted by the inverse of the total frequency of the finding/pattern label in the caption.

**Encoding** The text reports were cropped/padded to mean + 1 std + ‘start’ + ‘end’ tokens: 7 for IU-CX MeSH-sp-subset, 16 for IU-CX-MeSH, 43 for IU-CX free-text and 19 for ICH-Brain-DWI. Word indices were one-hot-encoded and passed through a word-embedding layer. The image encodings were extracted from the ResNet50 [117] CNN architecture, pre-trained on the ImageNet dataset [54]. Static image features of each X-ray image view were extracted from the last spatial average pooling layer ( $\mathbb{R}^{2048}$ ). The sum of all the features was aggregated across the image views to create a fixed-size input to the RNN of dimension  $\mathbb{R}^{2048}$ , which was then passed through the image transition layer.

**Training** For all experiments, the same 10% of the dataset was held out as test, which equated to 357 unique exams. The remaining 3,221 exams were used for training the sequence models. The LSTM model was trained for report generation conditioned on the combined image features by minimising the negative log-likelihood between the output and true sequence:

$$L(S, I) = - \sum_{t=0}^T \log p(P_t = T_t | \text{CNN}(I), P_0, \dots, P_{t-1}) \quad (3.3)$$

where  $p$  is the probability that the predicted word  $P_t$  equals the true word  $T_t$  at time step  $t$  given image features  $\text{CNN}(I)$  and previous words  $P_0 \dots P_{t-1}$ , and  $T$  is the LSTM sequence length. At training time, loss was minimised over the training set using stochastic gradient descent, and parameters were updated using Adam [120] optimisation. Training was terminated when loss on validation no longer decreased. Since it is time-consuming to evaluate BLEU/ROUGE/DAPS scores during training as it requires sampling from the model, metrics of accuracy, recall and precision over output words were used to determine whether the model was converging to an optimum. This was still not an ideal metric as the output space is equal to the vocab size, which is in the thousands for the free-text reports. For instance, relatively high recall/precision can be achieved by the simple prediction of the end token as these appear in all the reports. Hence, BLEU/ROUGE/DAPS evaluation that was performed post convergence was not necessarily being performed on the optimum model. Optimising directly for BLEU/ROUGE/DAPS performance is not possible because they require sampling, and are therefore non-differentiable.

However, this approach of minimising negative log-likelihood and terminating based on recall and precision is adequate for the purpose of tuning hyperparameters. Results were averaged over 5-fold cross-validation splits and hyperparameters were chosen based on the optimal average BLEU/ROUGE/DAPS performance over all the validation metrics.

**Inference** During inference, the image features were extracted from the pre-trained CNN and combined to the ‘start’ token embedding to create the input to the SERepGen-inject and SERepGen-merge models. For SERepGen-init, the image features are passed through the transition layer and used as the first word input. A prediction of the next word is sampled at the output and subsequently used as the input for the next prediction. Words are sampled until an ‘end’ token is reached.

### 3.3.4 Results

**Model comparison** Hyperparameters, such as the dimension of the word embedding layer, image transition layer and the LSTM hidden layer, were tuned for all models using 5-fold cross validation studies trained and validated on the IU-CX MeSH-sp (single-pattern) and single-view (PA) subset of the IU-CX dataset. BLEU-1, BLEU-4, Rouge-1 F1, and the F1 scores of groupings of disease, anatomy, position and severity terms are reported for combinations of hyperparameters in Figure 3.3.

In each static embedding model, the internal hidden and memory states of the LSTM attempts to capture different dependencies and hence each model requires specific tuning. For the SERepGen-init, the image transition layer forces the image embedding into the same embedding space as the word embedding in order for the LSTM to treat it as a pre-appended word. The word embedding layer is simultaneously learning dense word representations based on its position (since we are predicting the next word in the sequence). The balance is between keeping the maximum features of the image representation, but not overfitting the word embedding layer as the vocabulary is relatively small (95 for the IU-CX MeSH-sp subset). From the four hyperparameter studies under SERepGen-init in Figure 3.3, the optimal transition

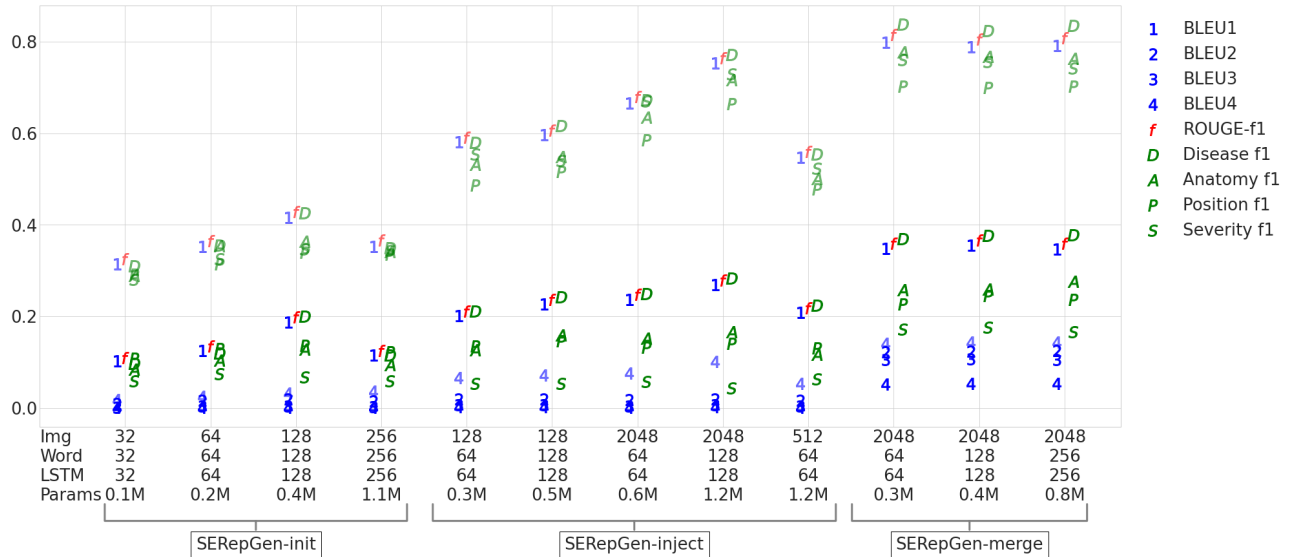


Figure 3.3: 5-Fold averaged cross validation hyperparameter studies of static image embedding models SERepGen-init, SERepGen-inject and SERepGen-merge. All models were trained and evaluated on the IU-CX MeSH-sp PA view subset. Reported metrics are BLEU-1, BLEU-2, BLEU-3, BLEU-4, Rouge-F1, and the F1 scores of groupings of disease, anatomy, position and severity terms. Results on training set are at half-transparency and results on validation set are opaque. The first four experiments in the figure are the hyperparameter studies of SERepGen-init, the next five are SERepGen-inject, and the last three are SERepGen-merge.

layer/word embedding dimension was 128, after which it begins to overfit.

For the SERepGen-inject model however, the image representation and word embedding dimensions are not tied together, and so the word embedding layer can be kept relatively small and the full image representation can be used. However, the hidden state of the LSTM is now required to learn the time dependency of the image together with the input words at each time step. In effect, the LSTM is being trained on more instances as each word now has an associated image and so overfitting is less of a problem. As demonstrated in the five hyperparameter studies under SERepGen-inject in Figure 3.3, keeping the full image representation as output by the CNN achieved the best results, with word embedding and LSTM hidden dimension set to 128.

The SERepGen-merge, like the SERepGen-inject, also doesn't require that the image and word representation have the same dimensions and so the full image representation can be used. However, in contrast to the SERepGen-inject, the image representations are combined at the output with the output of the LSTM at the final time-step. The LSTM hidden state is no

longer modelling image-word dependencies and hence the learn-able parameters at the input can be reduced. This is reflected in the results in Figure 3.3 as fewer parameters are required to achieve the same results as the SERepGen-inject model. Increasing the word embedding and LSTM hidden dimensions did not have a large effect on the overall performance of the SERepGen-merge, so fewer parameters are required to encode the language structure of the reports.

**Comparison of model performance on MeSH generation** Each static embedding model was re-trained on the full training set of the IU-CX MeSH-sp PA-view subset using the optimal hyperparameters, and evaluated on the test set. The BLEU scores are compared with the performance of the recurrent cascade model of Shin et al. [3] in Table 3.3. The DAPS metrics are compared for the three SERepGen models in Table 3.3.

	BLEU-1		BLEU-2		BLEU-3		BLEU-4	
	tr	te	tr	te	tr	te	tr	te
Recurrent Cascade [3]	97.2	<b>79.3</b>	67.1	9.1	14.9	0.0	2.8	0.0
SERepGen-init	41.6	18.8	18.6	1.9	8.2	0.0	3.2	0.0
SERepGen-inject	75.3	27.0	38.8	8.1	20.4	0.1	10.1	0.0
SERepGen-merge	78.3	34.8	40.1	<b>13.3</b>	28.5	<b>10.7</b>	12.6	<b>5.2</b>

Table 3.3: BLEU n-gram scores of static image embedding models train (tr) an test (test) metrics, evaluated on the IU-CX MeSH-sp PA-view subset, compared with recurrent cascade model of [3].

	D-F1		A-F1		P-F1		S-F1	
	tr	te	tr	te	tr	te	tr	te
SERepGen-init	73.5	27.8	62.7	17.6	60.2	13.7	64.3	6.6
SERepGen-inject	99.8	35.0	99.7	18.4	99.8	14.3	96.3	6.9
SERepGen-merge	99.9	<b>38.2</b>	99.8	<b>25.4</b>	99.8	<b>16.9</b>	97.6	<b>10.2</b>

Table 3.4: Disease, anatomy, position and severity (DAPS) averaged F1 scores of static image embedding models train (tr) an test (test) metrics, evaluated on the IU-CX MeSH-sp PA-view subset.

The recurrent cascade model uses a similar framework to SERepGen-init where the image embedding is used as the initial ‘word’ in the report. However, their approach to iteratively train the model significantly improves the model’s ability to produce the first word, but struggles to maintain visual correspondence in generating subsequent words, hence the steep reduction in

higher n-gram precision. This can be explained by the fact that the image CNN in the recurrent cascade model is first trained to classify images into disease annotations (by their definition of a disease annotation), and then fine-tuned to classify into clusters of joint image-context labels (which they identify through k-means). This approach relies heavily on assumptions made about the annotations and what is considered a non-overlapping disease label. If the initial assumptions of the disease labels, and later the image-context clusters, were incorrect, the joint image-text features are less meaningful and hence may result in poorer predictions. For instance, ‘calcified granuloma lung lower lobe left’ and ‘calcified granuloma lung base right’ have the same disease annotation, but their image features should differ based on the fact that the disease occurs in different locations. However, according to their results, these two annotations were found in the same embedding cluster. This may explain why the recurrent cascade model struggled to produce higher n-gram overlaps with the true reports.

Additionally, the definition of what is a disease annotation is subjective, and their approach mines the labels for least overlap, and removes a large number of cases (60%) where the MeSH annotations are too complex. The creation of the IU-CX MeSH-sp-subset had a similar approach, but removed only 40% of the exams for containing more than one disease pattern (described in more detail in Section 1.5.1).

The SERepGen-inject and -merge solve these problems by conditioning the word generation process on the image features at each time-step and by being trained end-to-end. In this way, image features are tuned in such a way as to retain semantic information pertaining to the disease and descriptions during training, and no assumptions are made about the meaning of the words in the annotations. Both models therefore achieve higher BLEU-2,3,4, and generalise better based on the improved performance on the test set. SERepGen-merge performed slightly better on all the metrics, which signifies that not only is it not necessary to input the image into the RNN at each time-step, but that it has a negative effect on the RNN’s ability to learn sequential dependence. By using the RNN to encode purely the linguistic features and a CNN to encode the image features, the dense decoder is able to use the meaning of the MeSH annotation and the learned image features to make a prediction on the next word.



**Qualitative evaluation** Examples of the MeSH annotations generated by SERepGen-merge on the test dataset are displayed in Figure 3.4. BLEU n-gram calculations are reported for individual predictions to give an indication of how ‘correct’ and ‘incorrect’ predictions are scored. It is evident that there are some limitations to using BLEU scores to evaluate the quality of the reports. For instance, reports shorter than 4-gram will automatically suffer a penalty for BLEU less than or equal to 4. An example is the prediction of ‘normal’ and true report ‘normal’ having no 2, 3, and 4-gram overlaps, contributing to a lower BLEU-2,3,4 score, even though the report is correct.

Another limitation of BLEU scores is in the interpretation of an individual score. It is not necessarily the case that a higher score correlates with better predictions of the correct findings, since all words are treated equally. Hence, it is also necessary to look at the DAPS recall and precision scores, which tally up the total predicted true positives, false positives and false negatives per MeSH category. The batch DAPS scores of the samples in Figure 3.4 give a better breakdown of how good the model is at predicting the correct disease vs the correct descriptions of the disease findings (anatomy, position, severity). DAPS recall and precision is calculated on a per-batch basis by counting how many of the disease/anatomy/position/severity terms in the MeSH annotations are correctly identified by the predictions. For instance, in the case of disease terms, 4 out of 8 are identified correctly (true positive rate = 0.5), and 4/8 are identified incorrectly (false positive rate = 0.5, false negative rate = 0.5).

**Multi-view and free-text extension** For extension into multi-view end-to-end training, the image encoder of the SERepGen-merge was modified to combine the outputs of two instances of the image CNN: one for encoding the PA-view images, and one for the L-view images. The model was trained and evaluated on its ability to produce reports from more complex chest X-ray exams where the images presented at least one abnormal pattern (IU-CX MeSH dataset), and for its ability to learn from and generate more complex reports where the vocabulary is more diverse: the IU-CX free-text and ICH-Brain-DWI 2D datasets.

From the BLEU, Rouge-F1 and DAPS-F1 scores in Table 3.5, it can be observed that, overall,









			
True Predicted B1 / B2 / B3 / B4	cardiomegaly, mild cardiomegaly, mild 100.0 / 100.0 / 50.1 / 31.6	normal normal 100.0 / 31.6 / 25.1 / 17.8	opacity, lung, hilum, right, round opacity, lung, lower lobe, left 38.94 / 31.79 / 18.52 / 13.23
			
True Predicted B1 / B2 / B3 / B4	calcinosi, lung, hilum, lymph nodes, left calcified granuloma, lung, middle lobe, right 19.47 / 7.11 / 7.53 / 6.26	deformity, clavicle, right lung, hypoinflation 0.00 / 0.00 / 0.00 / 0.00	opacity, lung, base, right calcified granuloma, lung, hilum, right 50.00 / 12.91 / 11.92 / 9.55
Batch DAPS Recall and Precision:			
D-R A-R P-R S-R      D-P A-P P-P S-P			
0.5 0.44 0.2 0.33      0.5 0.57 0.4 1.0			

Figure 3.4: Sample MeSH annotation prediction generated by SERepGen-merge on test set of IU-CX-sp PA view subset. Per-sample BLEU n-gram scores are reported beneath each prediction based on the original MeSH annotation. The disease, anatomy, position and severity recall and precision are reported on the batch at the bottom of the figure.

the best performance was achieved when training from and predicting the IU-CX MeSH-single-pattern. This is expected as it is a simplified dataset with one observed abnormal pattern and one corresponding manual MeSH annotation per exam. When training with exams that consisted of multiple MeSH annotations, the IU-CX MeSH, the BLEU-1 performance suffered considerably. Including the lateral view improved the BLEU-1 score and DAPS-F1, however, the disease label F1 score (31.3) was still lower than for IU-CX MeSH-sp (38.2). This implies that there were fewer correctly identified disease labels when training with multiple diseases per exam. There may be different reasons for this: multiple disease annotations do not necessarily correspond to multiple disease patterns in a one-to-one mapping. As mentioned in Section 1.5.1, MeSH annotations can be a combination of abnormal patterns and the disease diagnosis they represent. Additionally, some MeSH disease labels share abnormal patterns, such as ‘calcified

	View	B-1	B-2	B-3	B-4	R-F1
Random init.	-	21.2	5.2	1.0	0.2	23.8
IU-CX MeSH-sp	PA	34.8	13.3	10.7	5.2	37.6
IU-CX MeSH	PA	35.4	3.9	1.1	0.4	12.0
IU-CX MeSH	PA+L	35.4	3.0	1.9	0.6	36.4
IU-CX free-text	PA	25.2	13.2	7.0	3.6	32.9
IU-CX free-text	PA+L	24.5	14.7	8.4	4.0	36.9
ICH-Brain-DWI 2D	Axial	20.8	12.9	5.1	1.6	31.6

Table 3.5: BLEU n-gram scores of SERepGen-merge trained and evaluated on the IU-CX with PA- and L-view images and ICH-Brain-DWI dataset with axial slice images. Reported performance on test set. Results are compared against initialising the trained report generation model on random noise (in the shape of the image, with matched mean and variance to the image space).

	View	D-F1	A-F1	P-F1	S-F1
IU-CX MeSH-sp	PA	38.2	25.4	16.9	10.2
IU-CX MeSH	PA	36.9	13.5	20.3	21.9
IU-CX MeSH	PA+L	31.3	31.3	15.2	10.6

Table 3.6: Disease, anatomy, position and severity (DAPS) averaged F1 scores of SERepGen-merge trained and evaluated on the IU-CX MeSH-single-pattern subset and full IU-CX MeSH dataset with PA- and L-view images. Reported performance on test set.

granuloma’ and ‘nodule’. This could explain the reason why the disease F1 score is lower, but the anatomy, position and severity scores are similar to that of the model trained on IU-CX MeSH-sp: the patterns are identified incorrectly, but their context is the same.

The free-text reports could not be evaluated for DAPS since it would require categorising a much larger vocabulary, with no official ontological tool as there was for MeSH annotations. Hence, judging on purely the BLEU and Rouge-F1 scores, it can be seen that the performance of predictions on the IU-CX free-text and ICH-Brain-DWI is lower on average. Training on free-text reports suffers from the same obstacles as the MeSH annotations: exams that consist of multiple abnormal patters will have a complex report that lists not just the abnormal patterns, but potential diseases. As mentioned previously, different diseases can present with the same patterns, and potentially we do not want to predict a diagnosis, but rather limit the prediction to a qualitative description of the abnormalities. In addition to this, free-text reports do not have a fixed vocabulary for the same concepts, which introduces even more potential textual outputs for the same image feature inputs.

There are suggestions in literature on how to tackle training on long, multi-disease free-text reports, for instance, the use of a hierarchical RNN decoder [115] that conditions the generation of the sentence hidden states on the image features, and the word hidden states on medical concept embeddings extracted from the original reports. In this way, each generated word is informed by not only the image features and previous words, but also on chosen medical concepts. For instance, these concepts can be extracted disease labels, or abnormal patterns, which could help boost the generation of the correct disease predictions at the output.

### 3.3.5 Summary

Three different configurations of encoder-decoder architectures were trained and evaluated on their ability to generate increasingly complex radiological reports from single and multi-view images. The best performance judging by BLEU, ROUGE and DAPS scores was achieved by the SERegGen-merge, where the image features are combined at the output of the text encoder RNN, and the decoder is simply a fully-connected dense layer. Training the SERepGen-merge model on increasingly more complex datasets, including multi-view images, multi-disease/pattern reports and free-text reports revealed a considerable drop in BLEU, ROUGE and DAPS performance. Free-text reports have a more diverse vocabulary, negation, uncertainty and references to multiple diseases, all of which make it harder for a language model to learn the more variable structure. This also makes it harder for the image-text decoder to learn a mapping from image features to text when large portions of the text are no longer directly related to the images (specifically negation and uncertainty).

## 3.4 Dynamic Image Embedding

As seen in the previous section, the BLEU and ROUGE performance of the static embedding model dropped significantly when trained on full MeSH annotations and the more complex and varied free-text reports. In both cases, the reports now had references to multiple diseases in multiple locations, and a static embedding model is no longer suitable at capturing all of these in one vector representation. Hence, a dynamic embedding model, one where the text generation is conditioned on different image locations at each time-step, can potentially improve the predictions. In this section, the attention mechanism introduced by Mnih et al. [105] and used by Xu et al. [10] for image captioning is applied in the task of multi-view free-text report generation on the chest X-ray and brain-DWI images.

### 3.4.1 Dataset

For comparison with the SERepGen models, the same IU-CX free-text and multi-view dataset is used for training and evaluation of the dynamic attention model. The attention model is also evaluated on the 2D single-view subset of the brain-DWI dataset, and a ‘multi-view’ brain-DWI set which is simply the full 3D DWI images where each axial slice is treated as a view. Each brain-DWI multi-view exam is therefore made up of between 7 and 52 axial slices, which are sampled and padded to each consist of 20 slices (mean+1std). Exams with more than 20 slices are sampled with even distribution, and exams with fewer than 20 slices are padded with even distribution.

### 3.4.2 Related Work

The attention model builds on the NIC by taking lower-level CNN features corresponding to parts of the image and learning the ‘context vector’: a dynamic representation of the relevant part of the image at each time step, also trained by using an LSTM in order to generate the caption sequence. Attention mechanisms have been successfully used in machine translation

[121], image classification [105] and image captioning [10] in order to learn to attend to parts of the input: words in text, image regions, or both simultaneously. The benefit of using attention as opposed to a single vector representation for an image is that the lower-level CNN features retain richer, contextual information that is lost in the final layer through pooling of features. It also allows the sequence generation to focus on different, relevant parts of the image at each time step, similar to human image captioning. Additionally, the learned attention weights can be used to visualise the salient parts of the image used to generate a word at each time step, providing interpretability to the generation process. The soft attention mechanism of Xu et al. [10] demonstrated on an image captioning task is used by Zhang et al. [112] in the same way, but they additionally propose auxiliary attention sharpening which uses the implicit class-specific localisation property of global-average pooling, adding an extra layer of supervision to the attention weights. A co-attention mechanism is used by Jing et al. [116] where they compute attention over the image convolution features as well as the CNN predicted multi-label tags. They additionally propose the use of a hierarchical sentence-word LSTM as a better alternative for modelling longer, multi-sentence reports. Yuan et al. [115] use a similar approach but have the attention mechanism in both the sentence decoder and word decoder, as well as incorporating multi-view image fusion through a multi-view CNN encoder (also incorporating attention).

### 3.4.3 Model Architectures

Attention is learned over image regions by computing a context vector  $\mathbf{c}_t = \phi(\{\mathbf{a}_i\}, \{\alpha_i\})$  which is a dynamic representation of the relevant parts of the image at time step  $t$  for each location  $i$ , where  $\alpha_i$  are the weights of each image annotation vector  $\mathbf{a}_i$ . For 2D  $L \times L$  images, these annotation vectors are taken from a lower convolutional layer of a CNN. A recurrent neural network processes these inputs at each time step, learning a sequential internal representation of locations based on the prediction task. As per the formulation in [121] and [10], at each time step, a scalar score  $e_{ti}$  is computed for each location  $i \in 1 \cdots L \times L$ :

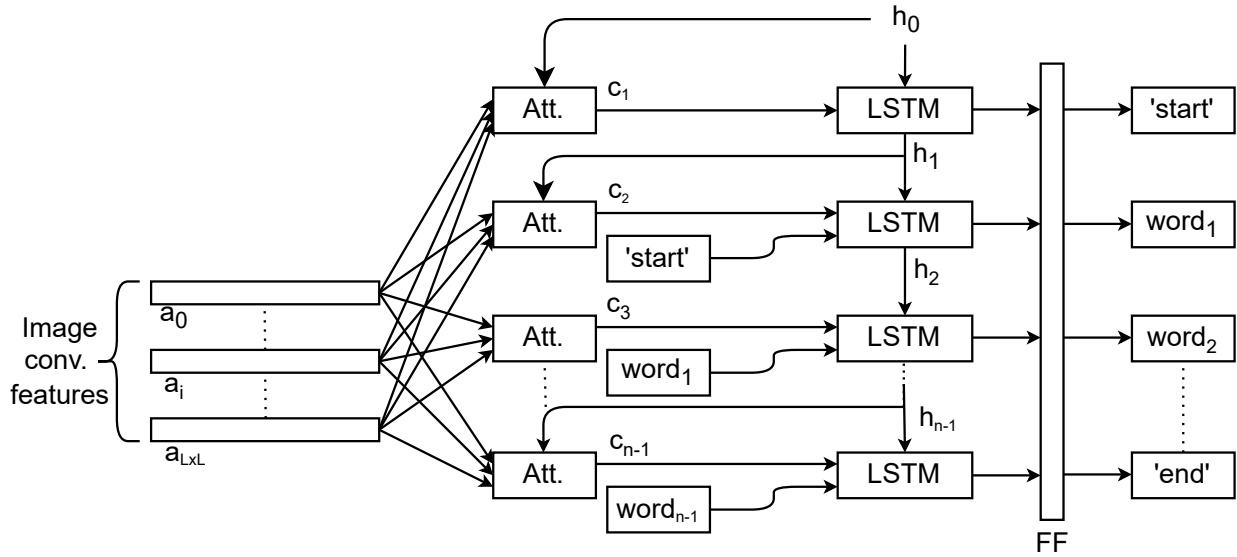


Figure 3.5: An illustration of soft-attention image captioning, reproduced from model description in [10]

$$e_{ti} = \text{FF}(\mathbf{a}_i, \mathbf{h}_{t-1}) \quad (3.4)$$

where FF is a feed-forward neural network,  $\mathbf{h}_{t-1}$  is the hidden state of the RNN at the previous time step. The score is effectively a similarity measure between the encoded states of the image and the encoded states of the text. An alternative to a FF network is the simpler dot product. Soft-attention weights  $\alpha_i$  can be computed using the softmax function over the score  $e_{ti}$ :

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_i \exp(e_{ti})} \quad \text{such that} \quad \sum_i \alpha_{ti} = 1 \quad \text{and} \quad \alpha_{ti} \geq 0 \quad (3.5)$$

The location  $i$  for the next time step can be found by sampling from this softmax (hard attention), or by computing the expectation over the feature slices (soft attention), the advantage of soft attention being that it is differentiable and can be learned through back-propagation. The soft-attention image captioning architecture adapted from [10] is illustrated in Figure 3.5. The weighted sum combination of inputs  $\mathbf{c}_t$ , or context vector, is then fed to the RNN:

$$\mathbf{c}_t = \sum_i^{L \times L} \alpha_{ti} \mathbf{a}_i \quad (3.6)$$

$$\mathbf{h}_t = \text{RNN}(\mathbf{h}_{t-1}, [y_{t-1}, \mathbf{c}_t]) \quad (3.7)$$

where  $y_{t-1}$  is the previous predicted word. The attention module of the DARepGen is illustrated in Figure 3.6

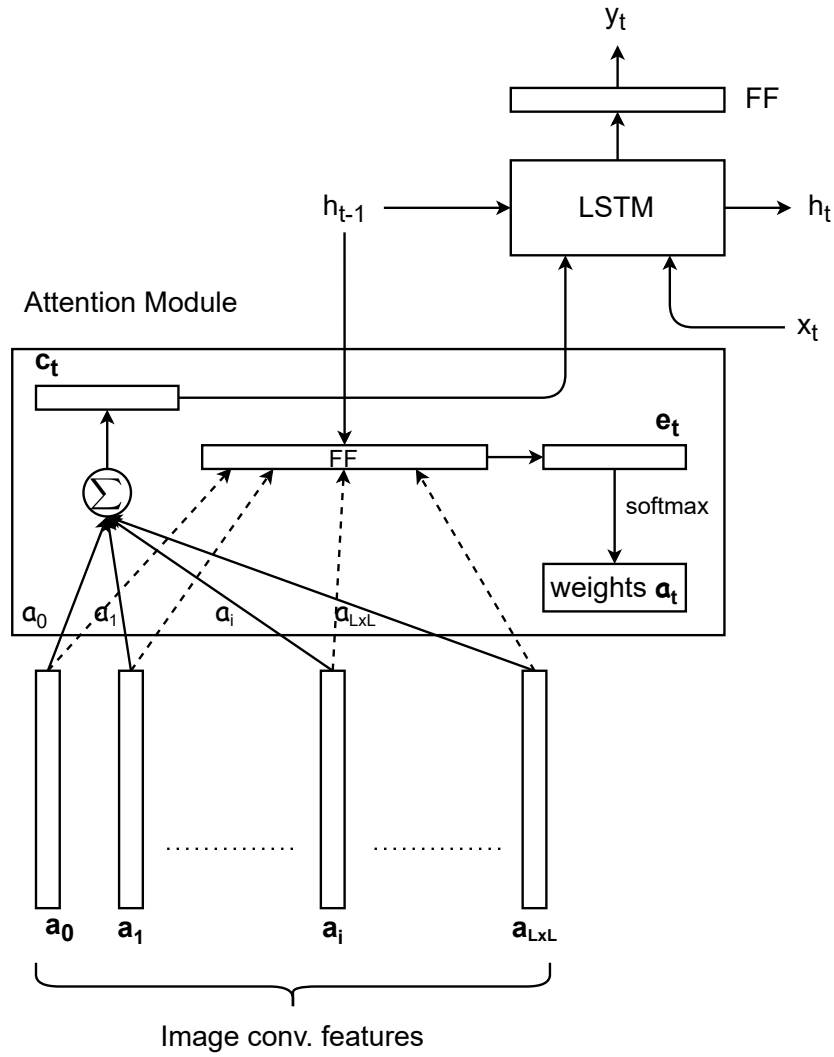


Figure 3.6: Dynamic attention report generation (DARepGen) module where attention is computed using the previous hidden state of the LSTM. Image convolutional features are a concatenation of conv features from either a single or multiple views.

To make this work for multi-view images, as with the static embedding models, a dedicated



image CNN network is trained to produce image features for each view.

### 3.4.4 Experiments

The dynamic attention model was trained to reduce the categorical cross-entropy loss in the same way as for the static embedding model:

$$L(S, I) = - \sum_{t=0}^T \log p(P_t = T_t | \text{CNN}(I_0) \dots \text{CNN}(I_K), P_0 \dots P_{t-1}) \quad (3.8)$$

where  $\text{CNN}(I_0) \dots \text{CNN}(I_K)$  are the feature vectors extracted from a convolutional network from image views  $[V_1, V_2, \dots V_K]$ . Data balancing, image and text augmentation and text encoding is performed using the same techniques as for static embedding model training, including image rotation/translation, sentence shuffling, pre-pending and appending ‘start’ and ‘end’ tokens. Convolutional image features were extracted from the final convolutional layer of ResNet50 pre-trained on ImageNet.

The method for training on the 20-slice brain-DWI set was slightly different to training on the single-view DWI subset and chest X-rays: attention was calculated over each individual slice, as opposed to locations within slices. Feature vectors were taken from the last spatial average pooling layer of the ResNet50 [117] CNN architecture, pre-trained on the ImageNet dataset [54]. The intention was to contrast this with having location-specific convolutional features taken from each slice, but this was not completed due to time constraints.

### 3.4.5 Results

BLEU and ROUGE metrics evaluated on test sets of each dataset are displayed in Figure 3.7. The use of attention significantly improved the performance of the text generation when training on the chest X-rays, but had less of an effect when training on the brain DWI exams. This may be because having location-specific features is more important when the reports are

made up of multiple diseases in multiple locations (as is the case for chest X-rays), but less important when looking for only one disease in one location (as is the case when examining a brain DWI for the presence of ischemic stroke).

Sample report predictions from the model trained on PA-view chest X-rays are displayed in Figure 3.7. The learned attention weights are overlaid on top the images at each time step, giving an indication of where the model is looking in order to generate each word. All predictions are of varied length since the word generation process was terminated after the first appearance of the end-token. Below each sample prediction is the original report and the BLEU n-gram scores as calculated for that specific prediction and true report. These samples demonstrate that the reports are coherent in terms of language and structure (i.e. they sound as if a radiologist has written them), but do not give much of an indication whether the generated words are conditioned on image features. This is especially evident when looking at the BLEU scores of the sample with highest BLEU performance (third sample): the original report notes the presence of emphysema and a calcified granuloma, but the rest of the report is a tick-list of anatomic locations of normal appearance. The predicted report also notes emphysema and granuloma (though in a much less coherent way), but also notes ‘degenerative changes in thoracic spine’, ‘thoracic kyphosis’ and ‘biapical pleural parenchymal scarring’, none of which are present in the original report. The high BLEU scores are achieved not only by matching the correct diseases and locations, but also on correctly matching what is ‘normal’ in appearance, which is typically a large number of things if the radiologist is following a systematic way of reporting on all anatomical locations, as suggested by The Basic Interpretation guide by Smithuis and van Delden [4].

Examples of brain DWI report predictions for the single axial slice dataset are presented in Figure 3.8. Attention weights are overlaid over the slice at each time step, and the corresponding generated word is printed above the slice. For the 20-slice brain DWI dataset, the generated reports and selected slices based on attention weights are displayed in Figure 3.9. The generated sequence of words and selected slices are highlighted in orange. The slices highlighted in blue are the ones selected by the 2D subset selection procedure that used the brain ischemia segmentation network developed by Chen et al. [5], outlined in Section 1.5.2. The

segmentation masks are displayed alongside the DWI slices.

Model	View	B-1	B-2	B-3	B-4	R-F1
IU-CX free-text						
SERepGen	PA	25.2	13.2	7.0	3.6	32.9
SERepGen	PA+L	24.5	14.7	8.4	4.0	36.9
DARepGen	PA	35.6	19.2	16.6	<b>5.3</b>	38.4
DARepGen	PA+L	<b>37.4</b>	<b>20.5</b>	<b>17.8</b>	5.0	<b>40.1</b>
ICH-Brain-DWI						
SERepGen	1-axial	20.8	12.9	5.1	1.6	31.6
DARepGen	1-axial	21.5	<b>13.4</b>	<b>6.8</b>	<b>2.3</b>	31.8
DARepGen	20-axial	<b>23.4</b>	12.5	5.5	2.1	<b>33.2</b>

Table 3.7: BLEU n-gram and ROUGE-1 F1 scores of DARepGen trained and evaluated on the IU-CX free-text single (PA) and multi-view (PA+L) datasets, and ICH-Brain-DWI single axial slice and 20 axial slice datasets. SERepGen-merge results displayed for comparison. Reported performance on test set.

### 3.5 Conclusion

Both static and dynamic image embeddings can be used as image representations for report generation in an encoder-decoder framework. When using static image embeddings, it is important to consider the purpose of the encoder: whether it is learning the time-dependency of a sequence initialised with an image, learning the language structure, or a combination of image-word time dependency. Optimal performance, based on a combination of BLEU, ROUGE and DAPS metrics, was achieved when the image CNN was considered as the encoder of image features, the LSTM as an encoder of textual features and the decoder as a dense layer that took both encodings as inputs to generate a probability distribution over words. The implication being that RNNs were designed to model sequential dependencies, which are inherent in language but not necessarily between image-word sequences. The use of the attention mechanism allows the LSTM to additionally learn a sequential traversal over the image when generating words, which bears some similarity to the way humans describe images by focusing sequentially on different parts of the image. In addition to learning attention over image regions at each time step, this model can be improved with attention over other inputs to the RNN decoder, for instance multi-label class predictions from a separate network, or other types of feature encodings

taken from the images from a separate but related task, for instance, image segmentation.

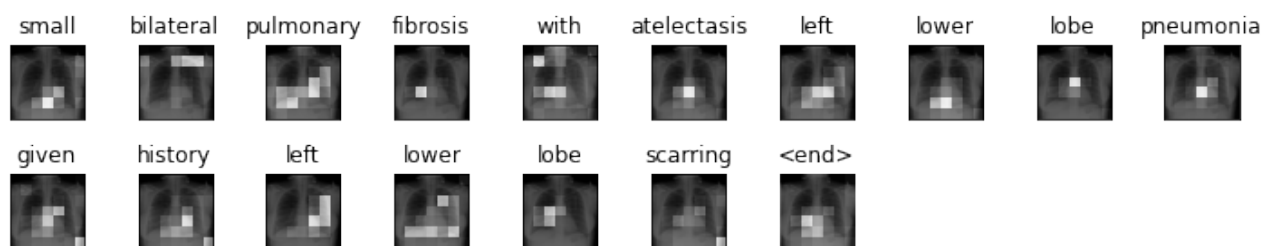
Training and evaluating this framework on hospital data was considerably more difficult than on curated data where reports were annotated with medical subject headings. This is as expected since the radiological reports of the hospital data (IU-CX free-text and ICH-Brain-DWI) were unstructured and had a much more varied vocabulary. This means that even if two images present the same abnormal patterns, two different radiologists may describe them using different words that both map to the same MeSH annotations. One way to address this is, instead of directly learning from free-text reports, to first learn a mapping of free-text to a vocab-controlled thesaurus, such as MeSH, and train using image-MeSH annotations.

In the following chapter, the combination of a medical concept extraction tool MetaMap [79] and statistical and machine learning methods of word representation and clustering, including tf-idf [71, 72], word2vec [73, 74] and k-means clustering [122], are used to extract and group concepts that can then be used as image labels.

## 3.6 Related Publications

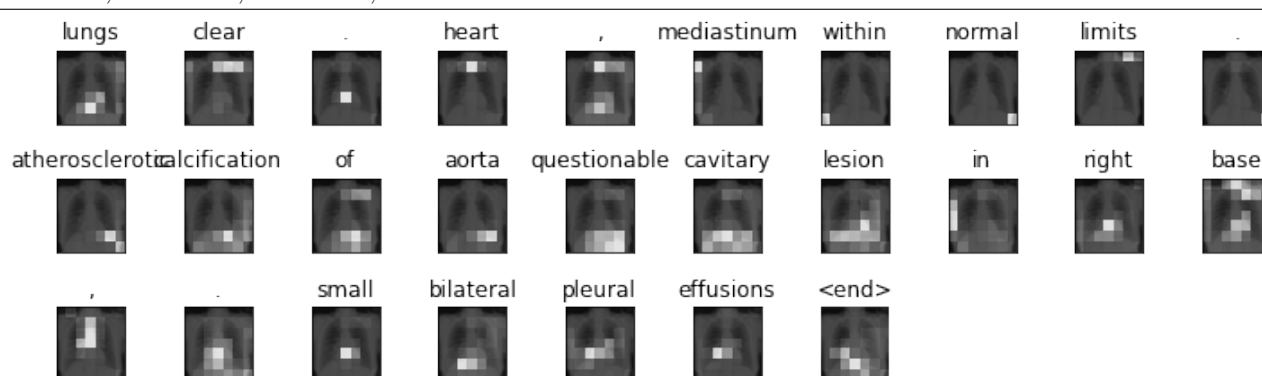
Gasimova, A. (2017). Automated Knee X-ray Report Generation, In NeurIPS Workshop on Machine Learning for Health, 2017.

Gasimova, A. (2019). Automated enriched medical concept generation for chest X-ray images. In Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support (pp. 83-92). Springer, Cham.



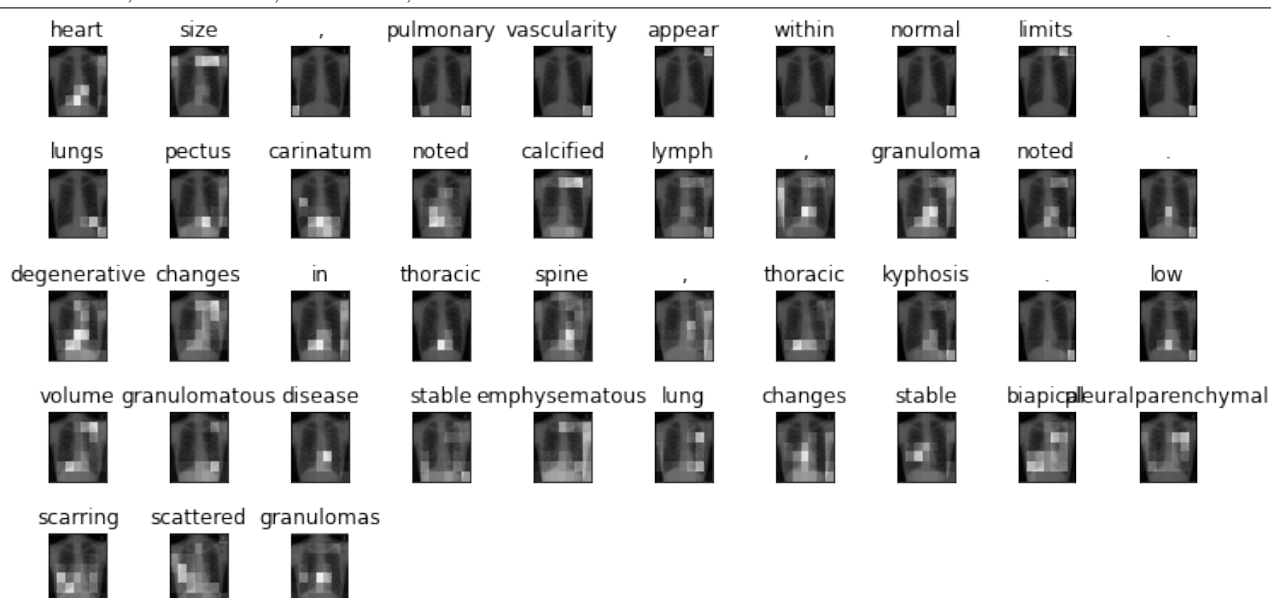
True report: ‘small focal opacity in left upper lobe, differential diagnosis includes subsegmental atelectasis, small infiltrate, scarring, followup recommended. heart size within normal limits. mediastinal, left hilar calcifications suggest previous granulomatous process.’

B1 = 9.7, B2 = 1.3, B3 = 1.0, B4 = 0.0



True report: ‘lower cervical, upper thoracic spinal fixation. multiple sternotomy. bilateral calcified granulomas, degenerative change in spine. lungs appear clear.’

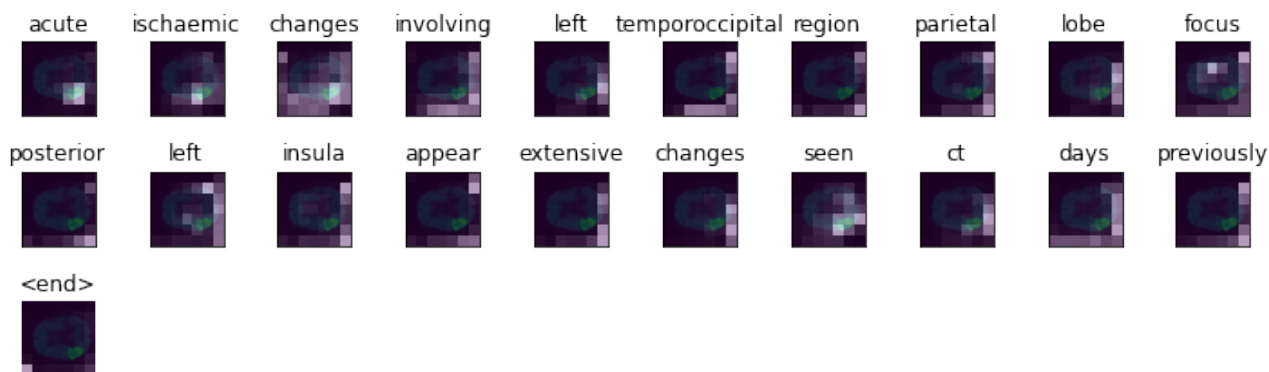
B1 = 33.3, B2 = 11.3, B3 = 5.2, B4 = 2.2



True report: ‘lungs hyperexpanded consistent with emphysema. pectus carinatum noted. heart size , pulmonary vascularity appear within normal limits. lungs. calcified granuloma noted. vascular calcification noted. hyperexpanded lungs consistent with emphysema. pectus carinatum.’

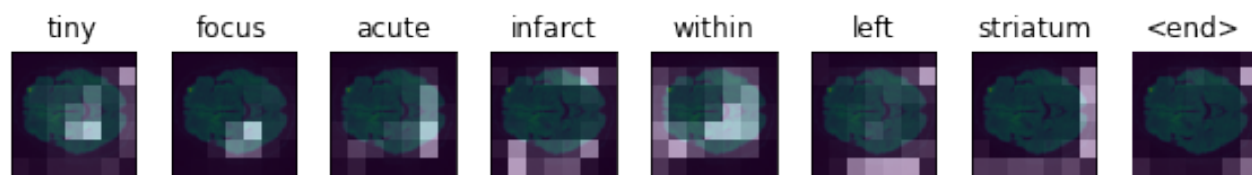
B1 = 44.2, B2 = 38.4, B3 = 37.9, B4 = 29.8

Figure 3.7: IU-CX free-text sample test report predictions and attention maps of DAREpGen.



True report: ‘acute left mca territory infarct involving left parietal lobe superior lateral left occipital lobe extending anteriorly left temporal lobe left operculum posterior left insula.’

B1 = 33.0, B2 = 23.9, B3 = 16.6, B4 = 6.0



True report: ‘recent acute/subacute less 10 days old infarcts within left cerebellum superior cerebellar artery territory.’

B1 = 11.8, B2 = 8.9, B3 = 5.1, B4 = 2.8



True report: ‘no acute infarction diffusion weighted sequences.’

B1 = 28.6, B2 = 21.8, B3 = 12.4, B4 = 7.0

Figure 3.8: ICH-Brain-DWI sample test report predictions and attention maps of DAREpGen.

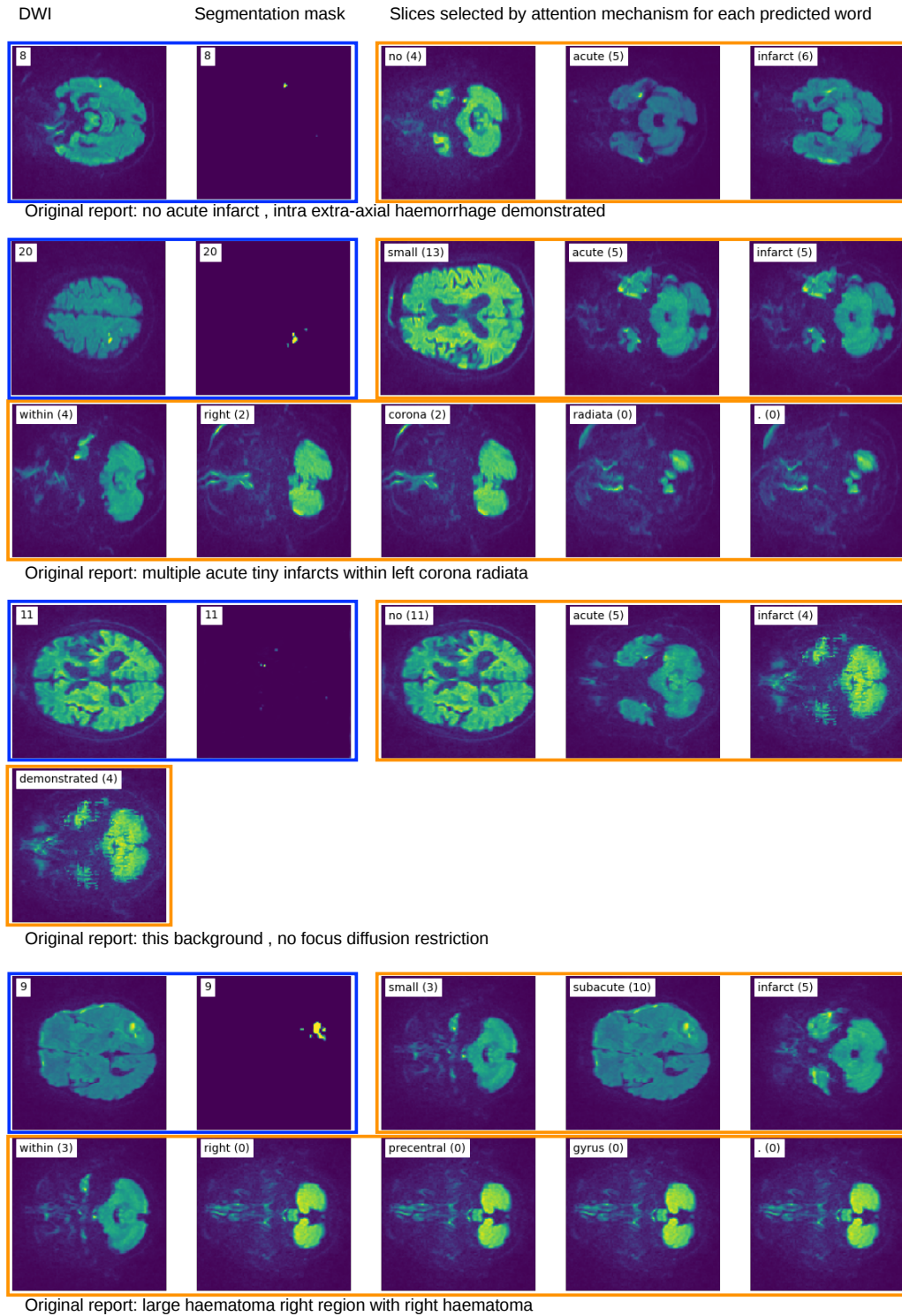


Figure 3.9: 20-slice brain DWI exams with predicted reports. Slices highlighted in blue on the left are the slices selected by the 2D brain-DWI subset selection procedure described in Section 1.5.2. They are displayed for comparison in order to highlight which slice (if any) contains the ischemia according to the segmentation network of Chen et al. [5]. Slices on the right, highlighted in orange, are selected based on max attention weights at each time-step. The generated word and selected slice number for that word are displayed in the top left corner of each image. The original reports are displayed underneath each sample.

# Chapter 4

## Medical Concept Extraction from Free-Text Diagnostic Reports

### 4.1 Introduction

As seen in the previous chapter, training report generation models on free-text reports are very difficult due to the unstructured nature of the reports. The reports may contain non-visually descriptive information (such as negation, which is difficult to correlate to image features), redundancy, spelling errors, inconsistent use of language (medical and non-medical). It is additionally difficult to evaluate the predictions for clinically-relevant concepts, for instance, correct identification of pathologies. On the other hand, training and evaluating using the medical subject heading annotations of the IU-CX dataset proved more successful as the vocabulary is controlled and consistent for abnormal patterns, and less noisy with errors. There is therefore a strong motivation to be able to map free-text reports to vocabulary-controlled structured medical concepts that correspond directly to visually-relevant image features.

Aside from radiology report generation, generating or extracting phrases that summarise visual image features from free-text reports has applications in automated image labelling for weak-label supervision and image retrieval tasks. For instance, radiology reports from past examinations have been used to extract classification labels for large chest x-ray training datasets



[107, 123, 124]. A variety of methods have been used to extract disease labels: rule-based [124], ontology-based tools such as DNorm [108] and MetaMap [79] that map terms to Unified Medical Language System [107] concepts, and manual expert-based annotation for supervised multi-label classification using convolutional and recurrent neural networks attention models [123].

The main challenges all these methods face are that free-text radiology reports are typically long ( $>50$  words), potentially consisting of references to multiple pathologies or lack thereof, with only small number of sentences or phrases containing key information summarising the disease, its location and severity. In addition, there is typically inconsistency across radiologists in the way they refer to findings in the images, with many phrases mapping to the same disease concept. Language inconsistency primarily affects rule-based and ontology-based approaches since they rely on pattern matching of a finite set of terms. On the other hand, neural network classification models have the challenge of generating meaningful report representations that are able to capture the disease concepts, and discard extraneous information. In addition, for tasks such as image retrieval and automated radiology report generation, single-class labels do not capture the full context of the pathology, such as severity and location, and therefore a generative approach is more suitable.

This chapter focuses on exploring and comparing manual, ontological, and a combination of statistical and ontological approaches to medical concept extraction of free-text radiological reports. An ontological tool is first used to extract phrases and concepts related to disease findings, and then these phrases and concepts are grouped together based on their meaning using a statistical approach (k-means clustering). The reason for merging the two approaches is that ontological tools, specifically MetaMap, do the majority of the work in filtering reports and extracting disease findings. However, as it is an extractive approach, these disease phrases and concepts can have a lot of overlap in terms of meaning. Additionally, the use of the ‘Diseases or Syndromes’ tag alone misses a large proportion of disease concepts. Additional MetaMap tags can be used, but must be carefully chosen to capture more phrases and not over-eagerly capture ambiguous phrases unrelated to disease findings. Both ‘Finding’ and ‘Pathologic Function’ capture a wider range of disease phrases and concepts, as well as some phrases that

are unrelated to disease findings. Using word-vector representation and k-means clustering, these phrases are grouped together based on their meaning, and hence a large vocabulary of findings is mapped to a much smaller, more consistent vocabulary, which is more suitable to use as image labels. These image labelling technique is then evaluated by performing multi-class classification on the images. The main contributions are as follows:

1. The ontological medical concept extraction tool MetaMap is used to tag and extract phrases and concepts related to disease findings in free-text radiological reports. Three different tag combinations are compared: ‘Disease or Syndrome’, ‘Disease or Syndrome’ + ‘Finding’, and ‘Disease or Syndrome’ + ‘Finding’ + ‘Pathologic Function’.
2. The tag combinations that resulted in the best correspondence of ‘normal’ cases according to the ground-truth MeSH annotations are chosen as new labels, but required grouping according to meaning due to many disease phrases referring to the same disease.
3. Two methods of grouping these concepts and phrases are compared: tf-idf and word2vec. This is done by creating disease phrase vectors and then clustering them using k-means.
4. Optimal clusters were chosen using the silhouette scores and manual checking of the clusters for semantically similar phrases. Clusters were then assigned as labels to the images, instead of the original extracted disease phrases.
5. These two phrase clustering methods of labelling are compared with the disease labels extracted directly from the MeSH annotations, and with labels extracted using MetaMap’s tag ‘Disease or Syndrome’.
6. The four labelling methods are compared by performing image classification and evaluating the accuracy, precision and recall of the predictions.

All of these techniques are tested and evaluated on the IU-CX dataset, Section 1.5.1, since they can be compared with the manual annotations provided by MeSH. The brain DWI dataset, described in Section 1.5.2, does not have manual annotations of the disease phrases in the reports, though it does have binary acute/no acute ischemia labels for the images. Additionally,

the reports are filtered to the 1-2 sentences describing the presence/absence of ischaemia, and so are much simpler than the reports for chest X-rays (which can report on any number of diseases in the images, or lack thereof). This means a slightly different approach needs to be used to group and assign disease labels that is more suited to this particular dataset. In the second half of this chapter, a combination of manual extraction and a hierarchical brain ontology is used to extract and group brain regions referenced by the reports to contain acute ischemia. These annotations, together with the binary labels of acute/no acute ischemia, and evaluated by assigning them as multi-label vectors to the images and performing multi-label classification.

## 4.2 Datasets

The Indiana U. Chest X-ray Collection (IU-CX) is the most suitable for comparing ontological/statistical and machine learning concept extraction methods because it consists of 3,955 individual exams, where each exam consists of X-ray images, the radiological report, and manual Medical Subject Heading annotations (MeSH<sup>®</sup>). The MeSH annotations can be treated as the gold-standard in concept extraction since they are made by expert radiologists, and summarise the disease findings in the reports using vocabulary-controlled MeSH concepts.

<p><b>Findings:</b> Normal cardiomedastinal silhouette. Interval improvement in lung volumes bilaterally. Improved aeration of the right and left lung bases. Bilateral small pleural effusions and left base atelectatic change, with interval improvement. Visualized XXXX of the chest XXXX are within normal limits.</p> <p><b>Impression:</b> Interval improvement in aeration of lung bases and pleural effusions. Residual small left effusion and questionable small right pleural effusion.</p>
<p><b>MeSH:</b> 1. Pleural Effusion/bilateral/small 2. Pulmonary Atelectasis/base/left</p>

Figure 4.1: Sample report and Medical Subject Heading (MeSH) annotations. Highlighted are phrases in the report that contribute the most to the MeSH annotations.

The free-text reports of IU-CX are made up of *Indication*: presenting symptoms, *Findings*:

visual descriptions of the x-ray scan that cover all presenting/non-presenting pathologies, and *Impression*: comment on presence/absence of specific pathology for which the scan was initiated. Each exam contained at least 1 MeSH annotations, with 43 % having at least 2. A sample of the free-text report and MeSH annotations, with manually highlighted disease findings, is presented in Figure 4.1.

## 4.3 Statistical and ontology-based concept extraction of chest X-ray reports

### 4.3.1 Related work

The burden of ground truth image label generation primarily rests with experienced clinicians, but as automated concept extraction from clinical reports is applied to more and more modalities, researchers are looking at ways in which to automate the process of ground truth label generation. Considerable work has been done in text mining radiological reports for concept extraction for the purpose of query retrieval [125, 126, 127], clinical support services [128, 129, 130], and coding of reports for administrative and analytical purposes [131].

A statistical approach to latent topic extraction was proposed by Shin et al. [110]. They proposed the use of Latent Dirichlet Allocation (LDA) to extract a hierarchy of latent topic labels from radiology reports. LDA was first presented as a probabilistic graphical model for topic learning by [132] that models ‘topics’ as having a distribution over a vocabulary and ‘documents’ as having a distribution over topics, where the topic distribution is assumed to have a Dirichlet prior. This distribution can be learned using Bayes inference methods, for instance, collapsed Gibbs Sampling [133]. Shin et al. [110] chose the number of topics by evaluating the ‘perplexity score’ of each model: the log-likelihood on a test set. A lower perplexity score corresponded to a better model. The main problem with this approach is that topics are not well-defined and are not explicitly disease findings.

Wang et al. [107] propose a primarily ontological approach, together with hand-crafted nega-

tions and uncertainty detection. They merge the results of ‘Diseases or Syndromes’ and ‘Findings’ concepts extracted by MetaMap [75] and DNorm [108] and use them to assign eight common disease labels to images for weakly-supervised multi-label classification and localisation.

### 4.3.2 Methods

**Disease terms extraction using MetaMap** The NegEx algorithm created by [118] was used to identify negatives from a ‘negation phrase list’ in the raw reports. After they were identified, a regular expression parser was used to remove all the negated phrases. These processed reports were then analysed by MetaMap [75].

The total number of tags that MetaMap was able to extract from the IU-CX reports was 107, the top 20 of which are listed in Table 4.1 along with their frequency of appearance. Visually-relevant parts of the report, such as pathology, severity, location, have been detected under tags such as *Finding*, *Disease or Syndrome*, *Qualitative Concept*, *Spatial Concept* and *Body Part, Organ, or Organ Component*. At least one *Finding* concept appears in 93% of the reports, and at least one *Qualitative Concept* appear in 92% of the reports.

Note, each identified phrase within the report may have multiple candidate mappings, each with varying scores determined by MetaMap. For instance, the phrase ‘interval development of bandlike opacity’ had 18 candidate mappings, two of which as displayed in Figure 4.2. In this example, the word ‘opacity’ is mapped to both *Finding* and *Pathologic Function*.

<b>Phrase: interval development of bandlike opacity</b>	
<b>Meta Mapping (696):</b>	
593	Interval [Temporal Concept]
760	development (development aspects) [Functional Concept]
593	OPACITY (Decreased translucency) [Finding]
<b>Meta Mapping (696):</b>	
593	Interval [Temporal Concept]
760	development (development aspects) [Functional Concept]
593	Opacity (Abnormally opaque structure (morphologic abnormality)) [Pathologic Function]

Figure 4.2: MetaMap sample output: identified phrase ‘interval development of bandlike opacity’ and two out of 18 candidate mappings, with corresponding MetaMap score.

Table 4.1: Top 20 Tags Assigned to the X-Ray Reports by MetaMap

Tag	Frequency	% of reports w/ at least 1 ap- pearance
Qualitative Concept	59311	92
Spatial Concept	57882	70
Finding	40167	93
Body Part, Organ, or Organ Component	37559	88
Quantitative Concept	31643	84
Intellectual Product	21531	38
Functional Concept	19283	46
Body Location or Region	18027	56
Pathologic Function	16713	33
Disease or Syndrome	14464	35
Idea or Concept	8264	22
Temporal Concept	7696	19
Therapeutic or Preventive Procedure	5544	12
Inorganic Chemical	4742	19
Organ or Tissue Function	4488	9
Medical Device	4476	10
Health Care Activity	4448	9
Manufactured Object	3070	8
Body Substance	2961	7
Tissue	2749	22

The concepts under the tag *Disease or Syndrome* appear at least once in 35% of the reports, and have the potential to make suitable labels for the extraction of valuable image features through a classification task. However, there are a few problems with simply using this tag. To begin with, reports where no *Disease or Syndrome* tag was identified can either be assumed to be ‘normal’, or excluded from training altogether. If they are assumed to be normal, then 64.7% of the reports would be labeled as ‘normal’, compared with 36.3% as identified by the manual MeSH annotations. This suggests a number of disease terms that were identified by radiologists were not identified by MetaMap. Alternatively, some diseases may have been classified under a different tag, such as *Finding* or *Pathologic Function*.

Including terms or phrases under *Finding* and *Pathologic Function* tags increased the proportion of exams with identified diseases, and increased the vocabulary of the identified disease terms (Table 4.2). A full list of phrases and terms extracted from *Disease and Syndrome*, *Finding* and *Pathologic Function* that appear in at least 30 reports is presented in Appendix .2, Table

MetaMap tag	Total vocab	% normal
Disease or Syndrome	344	64.7%
Disease or Syndrome + Finding	854	42.5%
Disease or Syndrome + Finding + Pathologic Function	949	37.9%

Table 4.2: Total vocabulary and percentage of ‘normal’ cases when using multiple MetaMap tags to extract disease terms and phrases.

2 and the top 20 in Table 4.3. These tags also captured words and phrases that may not be related to pathologies, as well as phrases that state no findings, or describe normal appearance, for instance ‘clear’ and ‘normal heart size’. Therefore, to determine the proportion of ‘normal’ cases, reports in which every extracted phrase contained words in [‘clear’, ‘normal’, ‘intact’, ‘negative’, ‘well’, ‘limited’, ‘unchanged’, ‘negative’] were considered normal, which came to 37.9% of the exams, similar to the proportion identified by the ground-truth MeSH.

As more terms have been extracted from each report, there is considerable overlap in appearances between phrases. For instance, from Table 4.3, the word ‘normal’ has 113% overlap because it appears together with other phrases, and additionally within the other phrases such as ‘normal heart size’. The phrase ‘cardiomegaly’ appears alongside another phrase in 93% of cases. Terms and phrases with high percentages of overlap are not particularly useful as image labels if the ultimate goal is to use CNNs for image classification and train by minimising cross-entropy. One option is to perform multi-label classification, however, it is also evident that the phrases are not independent. In some cases, they have the same meaning: ‘normal heart size’ and ‘heart size normal’. In other cases, phrases have enough semantic similarity that they can be grouped together under one label, for instance, *thoracic spine degeneration* and *degenerative spine* are referring to the same pathology, with *thoracic* referring to position; similarly with *pleural effusion* and *pleural effusions bilateral*. Therefore, an approach is needed to cluster phrases together based on their meaning, and treat the clusters as independent image labels.

**Creating disease phrase vectors** There are several choices for clustering algorithms, but they first require a choice of semantic similarity measure. Text similarity measures can be split into two main types: string-based and knowledge-based. String based similarity measures

Table 4.3: Top 10 Metamap extracted ‘Disease or Syndrome’, ‘Finding’ and ‘Pathologic Function’ terms and their percentage overlap with other phrases. Percentage is calculated as total appearance as a single phrase divided by appearances with and within other phrases. For instance, the word ‘normal’ appears as single Metamap term 2512 times, but also appears within other Metamap terms such as ‘normal heart size’.

Disease/Finding/Pathology phrase	Total appearances	Appearances w/ and within other terms or phrases	Overlap %
clear	3350	2651	79
normal	2512	2848	113
heart size	1812	2434	134
intact	711	664	93
normal heart size	703	506	72
atelectasis	630	696	110
cardiomegaly	532	494	93
opacities	479	477	100
heart size normal	397	359	90
unchanged	342	319	93
thoracic spine degeneration	336	306	91
opacity	327	387	118
disease	319	678	212
degenerative spine	259	230	89
pleural effusion	244	343	140
crowding	165	159	96
normal breast	163	159	98
emphysema	157	169	107
tortuous aorta	157	153	97
tortuous	157	288	183



various forms of ‘distance’ between strings for approximate string matching, for instance, cosine similarity of term frequency-inverse document frequency (tf-idf) vectors. Knowledge similarity is measured based on similarities obtained from semantic networks, such as WordNet [134] (a large lexical database of English where words are organised and linked according to their semantic relations). Additionally, cosine similarities can be obtained from word embeddings such as word2vec [74] which have been trained on a large corpus. Both tf-idf and word2vec can be created directly on the corpus, though they both greatly benefit from having a large and diverse set of documents.

In application to radiological reports, one document is the set of words within a phrase extracted by MetaMap under the tags ‘Disease or Syndrome’, ‘Finding’ and ‘Pathologic Function’, and the corpus is the full set of phrases extracted from all reports.

Since both architectures make use of the context surrounding a word to determine its meaning, keeping the context surrounding the extracted disease terms is fundamental for training. Therefore, the full textual reports (post negation removal) were used to train the word2vec model to create word representations, and disease phrase representations were created by averaging the word vectors within the phrase.

**Clustering disease phrase vectors** K-means clustering was performed on tf-idf weight vectors and word2vec phrase representations. K-means minimises the within-cluster variance, or squared Euclidean distance, but cosine similarity is a more appropriate measure for tf-idf and word2vec representations. Hence, the vectors are first normalised so that the Euclidean distance is connected linearly to the cosine distance.

To start with, a tf-idf matrix was generated for the 949 individual disease phrases in the reports. Each phrase consisted of a vector of tf-idf scores, one discrete value score for each word in the vocabulary. When building the vocabulary, words that appeared in 99% of the reports are discounted, as well as words that appeared in less than 1%. This gave a total vocabulary, and therefore dimensionality, of 39 words. The tf-idf matrix ( $\mathbb{R}^{949 \times 39}$ ) was then clustered using k-means. To determine the optimal number of clusters, the mean silhouette score was plotted

for each  $k$ , displayed in Figure 4.3a. The silhouette score is a measure of how separated the clusters are, and is calculated by  $(b - a) / \max(a, b)$  where  $a$  is the mean intra-cluster distance and  $b$  is an average of the distances between each sample and a cluster it is not part of. The range of values is between -1 and 1, 1 indicating that the point was placed in the correct cluster. The average silhouette score reached a maximum at around 175 clusters, a sample of which is visualised in 2 dimensions using SVD in Figure 4.3b.

For comparison, a word2vec matrix was created by first fitting the CBOW word2vec model on the reports (without negation), and then averaging the disease phrase vectors. The feature space was chosen to be 100. K-means clustering was performed on the word2vec phrase representations, and the optimal number of clusters was chosen by also plotting the silhouette scores, displayed in Figure 4.4a. In the case of word2vec, the silhouette score had still not converged after 300 clusters, meaning a better score can be achieved by increasing the number of clusters. However, this may not be desirable since, according to the ground-truth MeSH annotations, there are only 31 unique diagnosis labels (Table 1.3 in Section 1.5.1).

It is difficult to determine the quality of the clustering from the silhouette scores and visualised SVD clusters alone. Looking at the two distributions of terms per cluster in Figure 4.5, it is evident that a single cluster contains the largest proportion of disease phrases. For tf-idf, roughly 40% have been grouped under one cluster, and for word2vec, roughly 13% (the lower proportion could also be due to the larger number of clusters used to perform k-means). Investigating the phrases under largest cluster of tf-idf and word2vec, a sample of which is listed in Table 4.4, it can be seen that, semantically, these words are not similar. This may be due to the way in which the tf-idf matrix is constructed: words that appear very rarely across the documents are left out from the vocabulary, and so if a disease phrase only contains words that are not in the vocabulary, the entire vector is zeros. These vectors naturally become grouped under one cluster. These words can be incorporated back into the vocabulary, however, this is evidence of a slightly larger problem with the clustering system. When each ‘document’ (in this case, each disease term or phrase) is small and vocabulary is large, the tf-idf matrix is very sparse. This means that phrases may become grouped together if they only share one word (as long as that word is not especially common across the phrases). For instance, terms

Table 4.4: Sample of disease phrases extracted from largest cluster of rare words from tf-idf and word2vec k-means.

Tf-idf	Word2vec
hiatal hernia	retracted
opacity	collapsed
osteopenia	infections
clear	hematoma
tortuous	pass
ectatic	pleuritis
cardiomegaly	pancreatitis
consolidation	obese
shunt	avn
increased	gallstones
mas	stigmata
opacities	cholelithiasis
combination	short
thickening	degenerated
round	bronchiolitis, viral

*small airways disease, granulomatous disease, diseases of the joints, bullous disease* are under one cluster presumably because they all contain the word *disease*, and the other words in the phrase are either unique, or very rare. The same problem is present when using word2vec due to the same reasons: words that are rare are left out, and if a phrase only contains rare words, the phrase vector is zeros. If all rare words are kept in the vocabulary, there will many instances of phrases containing unique or rare words which are distinct enough that they do not belong in any cluster. It would require very careful tuning of the k-means clustering algorithm to make sure clusters are distinct.

Another way to evaluate the quality of the clusters is to assign them as image labels and evaluate the performance of image classification. The performance will indicate whether there is correlation between the disease phrase clusters and visual features in the radiological images. The phrases with zeroed vectors were removed from both tf-idf and word2vec, and the matrices re-clustered using  $k=175$  and  $k=300$  respectively. Phrases were then replaced with their respective cluster indices, and multi-class classification was performed on the PA-view images using the ResNet50 architecture, pre-trained with ImageNet [51], with the final classification layer being replaced with a dense layer of dimension equal to the number of class labels. This was compared to the classification performed on images labeled with the ‘Disease or Syndrome’

extracted terms and phrases, which were not clustered. For all experiments, labels that appeared less than 30 times were removed, and exams with no labels were dropped from training. To create a single image label, only the most common disease labels were used per each exam. The final number of labels and exams used for training the different experiments is outlined in Table 4.5. The same set of 300 exams was used for testing all of the experiments.

Table 4.5: Multi-class experimental set-up for extractive and clustering labelling methods.

Labelling method	Num. labels	Total exams
MeSH disease	15	2369
MetaMap D/S tag	16	3138
tf-idf 175	31	3197
word2vec 300	54	3587

Table 4.6: Classification comparison of labelling methods using micro-precision, -recall, -F1 and macro-precision, -recall and -F1. Results reported on test set. Micro-P/R/F1 refers to the average over classes, and macro-P/R/F1 are averaged over samples.

Labelling method	Acc.	micro-P	micro-R	micro-F1	macro-P	macro-R	macro-F1
MeSH disease	22.9	24.5	22.9	23.7	4.7	10.2	6.4
MetaMap D/S tag	10.3	69.9	10.3	18.0	8.8	13.4	10.6
tf-idf 175	63.8	41.8	63.8	50.5	3.4	5.2	4.1
word2vec 300	9.8	57.2	9.8	16.7	7.9	6.7	7.3

These labelling methods are compared with using disease labels extracted from manual MeSH annotations in Table 4.6. Classification results indicate that, even when using manual MeSH annotations to provide disease labels, precision and recall performs very poorly. There may be various reasons for this. Firstly, there is an inherent problem in assigning a single disease label when many diseases tend to co-occur. A multi-label classification approach may be more appropriate, but would require carefully tuning the loss to balance the distribution of labels per sample as well as what these labels should be.

Secondly, even MeSH disease annotations may not be ideal as image labels, specifically for chest X-rays since the same image features, or patterns, are associated with different diseases (as explained in Section 1.5.1), which may make convergence difficult since the same image feature inputs are being mapped to different outputs. The MeSH annotations consist of a mix of abnormal patterns (general image features, such as areas of opacity) and diagnoses, such as ‘calcified granuloma’. Extracting the patterns from MeSH annotations as described in

Section 1.5.1 was a manual process, and therefore not ideal as an automated approach to image labelling.

MetaMap ‘Disease or Syndrome’ tag achieved the best macro-F1. Macro scores are averaged over classes rather than instances, so the MetaMap labels achieved the best average per-class performance. Although MetaMap ‘Disease or Syndrome’ tag has the downside that it mis-categorises, therefore misses, some disease phrases from the reports, results show that the ones it does not miss can be used for image labelling. However, based on the recall and precision performance, it is still not an adequate method for automated diagnosis of chest X-ray images.

## 4.4 Anatomical brain region mapping using manual extraction and a hierarchical ontology

The radiological findings of the ICH brain DWI dataset of patients with stroke-like symptoms are structured in a similar way to the chest X-rays: presence/absence of pathology/lesion, anatomical location, severity and any visually descriptive features of the lesion. The problem of extracting disease labels is somewhat simplified for this dataset as clinicians are primarily interested whether the patient has suffered an ischemic stroke, and if so, its location. As detailed in Section 1.5.2, each exam is assigned a diagnosis label as part of reporting, which is treated as a binary presence/absence of acute infarct. The descriptions of the brain regions, on the other hand, are far more diverse and therefore more difficult to categorise. This section proposes a semi-automated method of extracting and grouping brain regions from text reports using a hierarchical brain ontology, and evaluates the labelling technique by attempting multi-label brain-DWI image classification. The classification network is inspired by attention-guided multilabel video classification tasks [135, 136], where each slice in the DWI is treated as a frame and attention weights are learned over the individual 2D slices. In this way, the problem of having to model a 3D brain volume is avoided and, instead, pre-trained 2D convolutional neural networks can be used to extract image features.

### 4.4.1 Method

A combination of manual and automated annotation was necessary in order to extract terms relating to brain regions from the clinical reports. A hierarchical brain region ontology available from the Allen Institute [137] was used to manually extract the terms, and then automatically assign these terms to larger, parent regions in the hierarchy. For instance, ‘left middle temporal gyrus’ is located within, and therefore reassigned to, the ‘temporal lobe’. Regions that occurred less than 3 times across the entire corpus were excluded. In this way, 356 unique regions were reduced to 42. The 20 most common regions and their frequency of appearance are listed in Table 4.7.

Table 4.7: Top 20 classes of brain regions after re-assignment based on a hierarchical ontology.

Brain region	freq.	Brain region	freq.
frontal lobe	100	left cerebellar hemisphere	17
basal ganglia	99	cerebellum	15
parietal lobe	74	centrum semiovale	13
corona radiata	72	posterior cerebral artery	12
middle cerebral artery	68	medulla oblongata	11
occipital lobe	45	midbrain	11
pons	43	superior cerebellar artery	11
insular cortex	41	perirolandic region	9
thalamus	39	thalamocapsular region	8
temporal lobe	36	right cerebellar hemisphere	8

An example of the extraction and mapping is shown in Figure 4.6. In this way, all of the free-text reports are mapped to a binary presence/absence of ischemic infarct and the corresponding brain region labels. This labelling technique was evaluated by encoding the labels into k-hot vectors, and performing multi-label classification on the brain image slices.

As with the chest X-rays, a pre-trained CNN network, in this case VGG-16 [138], was used to extract dense image features from the last average-pooling layer. This was compared with using recurrent attention over the final convolutional layers of the individual slices, inspired by attention-guided multilabel video classification tasks [135, 136]. In video-labelling, the RNN is used to model the temporal dependencies of frames and attention is learnt over locations within individual frames. The same approach is adopted for the brain slices where an RNN is used to model the sequential dependencies of slices within the DWI and trained to produce a

multilabel output.

For each 3D DWI, the annotations vectors of each slice are taken from the last convolution layer of a CNN:  $a = \{\mathbf{a}_1, \dots, \mathbf{a}_N\}$ ,  $\mathbf{a}_i \in \mathbb{R}^{L \times L \times D}$ . The weights can be thought of as the probability distribution of the relevancy of each location encoding to the output (diagnosis+location). The input at the next time step  $x_t$  is then the expectation of features at different locations:

$$x_t = \sum_i^{L^2} \alpha_i \mathbf{X}_i \quad (4.1)$$

The complete model is illustrated in Figure 4.7. The dimension of the hidden state of the LSTM is set to  $\mathbb{R}^{512}$ . The LSTM is unrolled up to 19 time steps for the average number of slices. Images with fewer than 19 slices were re-distributed and padded with intervening slices. The LSTM model was trained by minimising the cross-entropy loss:

$$L(S, I) = - \sum_{t=0}^K \sum_{c=0}^C y_{t,c} \log \hat{y}_{t,c} + \lambda \sum_i W_i^2 \quad (4.2)$$

where  $y_t$  is the k-hot vector of labels at time step  $t$  and  $N$  is the LSTM sequence length,  $\lambda$  is the weight decay coefficient, and  $W$  are all the model parameters. Exams were split into 80%/10%/10% for training, testing and validation respectively. At training time, loss was minimised over the training set using stochastic gradient descent (batch size 16, learning rate 1e-5, 10 epochs), and parameters are updated using Adam [139] optimisation.

#### 4.4.2 Results

The performance of the models is evaluated using accuracy, precision, recall and Hamming Loss [140], comparing the results of the VGG-16 model trained with and without attention. Table 4.8 summarises the quantitative results. The accuracy of the ‘infarct’ class is reported on its own, and the mean average accuracy, precision and recall (mAA, mAP, mAR) and Hamming Loss (HL) is reported over all the classes (including the ‘infarct’ class).

Table 4.8: ICH Brain DWI multi-label classification results using extractive labelling technique. The accuracy of the ‘infarct’ class is reported separately as well as part of the mean average accuracy, precision and recall (mAA, mAP, mAR) and Hamming Loss (HL) of all classes.

	Acc. ‘infarct’ (%)	mAA (%)	mAP (%)	mAR (%)	HL (%)
VGG-16, central slice	59.3	97.3	29.3	17.1	2.7
VGG-16, max-agg.	55.1	<b>97.9</b>	24.6	14.2	<b>3.1</b>
VGG-16+att.	<b>68.0</b>	97.7	<b>39.0</b>	<b>19.6</b>	2.2

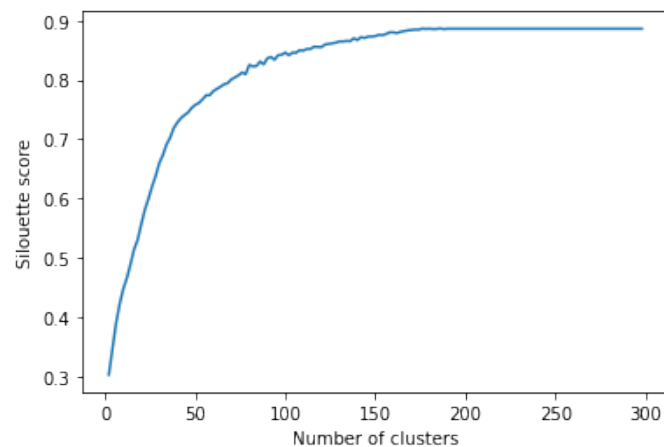
From the performance results, it can be seen that taking the central slice as input performed better than taking all of the slices and max-aggregating the features. This is as expected since lesions will only be present in a small number of slices, and taking all the slices as inputs introduces a lot of noise. On the other hand, taking the central slice is the more naive approach: a lesion reported in the text may not be present in the central slice. Taking an expectation over all the locations across the slices provided a compromise: the entire image was explored, but the input was more localised at each time step. The accuracy of the ‘infarct’ class achieved better performance using attention, and in addition, an improvement is seen in mean average recall, which is especially important in the medical domain as we want to identify all patients with high-risk lesions (for further examination).

Although the mean average precision and recall is relatively low (39.0 and 19.6 respectively), it does show that brain regions can be predicted alongside the presence/absence of an infarct to some extent through the use of manual labelling techniques and models that use attention over image slices. Manual labelling has some advantages over automated extraction using MetaMap: for instance, all brain regions were extracted and mapped no matter their spelling or word order, and word disambiguation is less of a problem for human annotators since they are able to interpret the context of words far better than MetaMap. However, manual annotation of exams is a very time-consuming process, and can introduce human errors, especially if it is not performed by a clinician. These errors can contribute to making the labels noisy, and therefore more difficult to learn from.



## 4.5 Conclusion

Using general ontological tools to extract disease findings from radiological reports proved to have several disadvantages: extracted disease terms or phrases had several potential mappings, some of which were not under the obvious tag of ‘Disease or Syndrome’, but the more general ‘Finding’, many disease phrases were semantically similar and required further statistical processing, and the ambiguity of context meant that many words and phrases were miscategorised. Grouping the phrases into clusters of similar meaning was also challenging as tf-idf and word2vec phrase representations were not suitable for modelling rare words, especially with such a small training corpus. The results of the classification performed on the clusters showed that the labels were not suitable for extracting image features that correlate with the findings in the clusters.

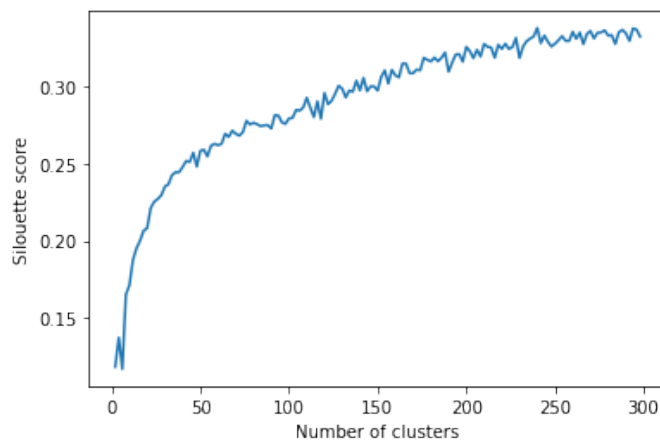


(a) K-mean average silhouette.

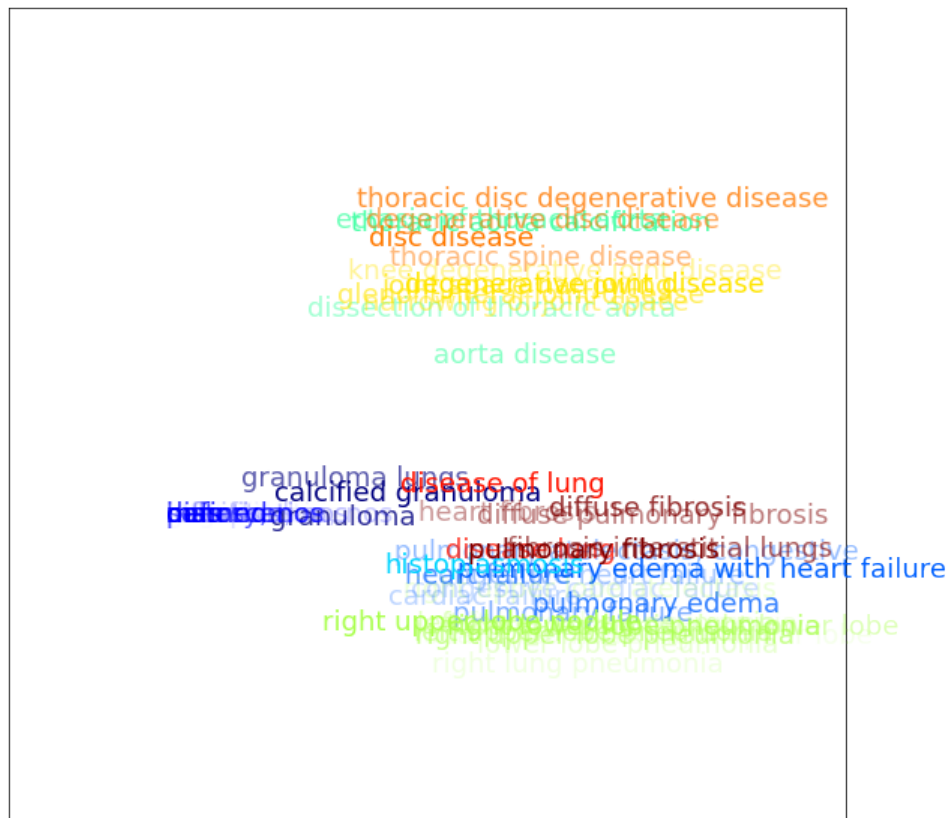


(b) Sample of disease phrase vectors and their clusters, reduced with SVD.

Figure 4.3: K-means clustering performed on tf-idf representations of disease phrases extracted by MetaMap.

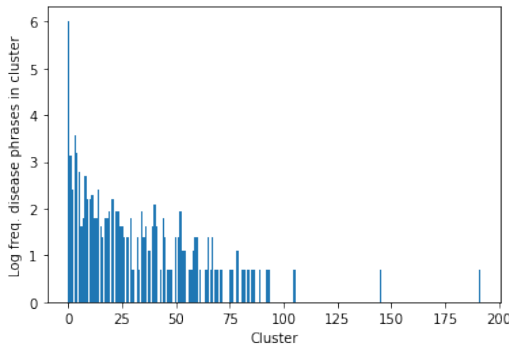


(a) K-mean average silhouette.

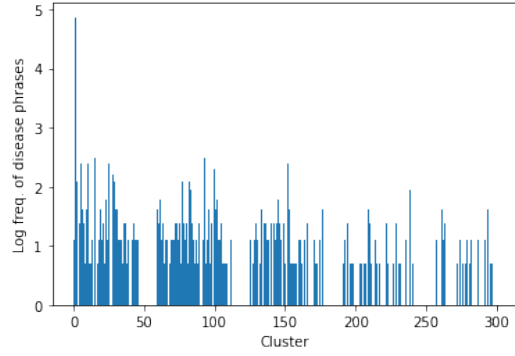


(b) Sample of disease phrase vectors and their clusters, reduced with SVD.

Figure 4.4: K-means clustering performed on averaged word2vec representations of disease phrases extracted by MetaMap.

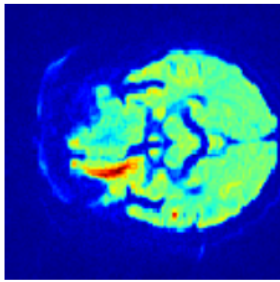


(a) Tf-idf



(b) Word2vec

Figure 4.5: Log of disease phrase frequency distributions over k-means clusters performed on tf-idf and word2vec representations.



**Clinical Report:** There is a small focus of acute ischaemia in the right corona radiate, and a tiny focal cortical infarct in the left middle temporal gyrus.

**Clinical diagnosis:** Acute infarct

**Manually extracted regions:** right corona radiate, left middle temporal gyrus

**Region mappings:** corona\_radiata, temporal\_lobe.

Figure 4.6: Central slice of sample DWI exam with corresponding clinical report, clinical diagnosis, manually extracted brain regions and region mappings.

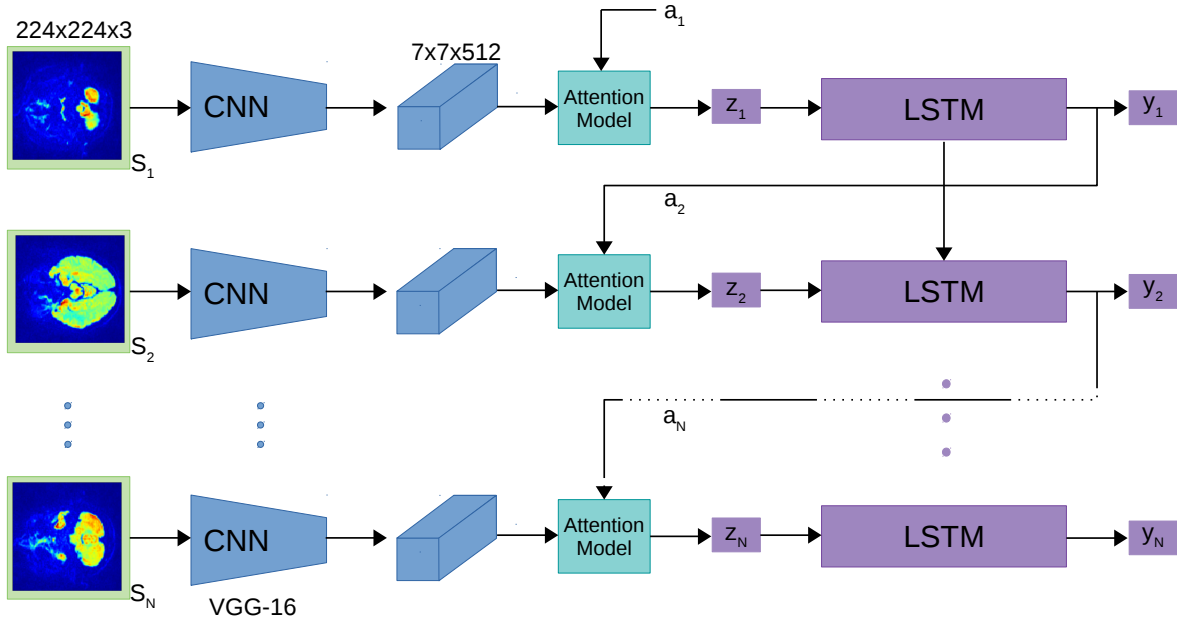


Figure 4.7: Attention-guided clinical report generation model.

## Chapter 5

# Abstractive Concept Extraction from Free-Text Diagnostic Reports

### 5.1 Introduction

In the previous section, it was shown that using ontological tools to extract terms and phrases from free-text reports that relate to disease findings results in a very diverse set of phrases that are very difficult to group by meaning. Extractive summarisation techniques like this have the downside that phrases are selected directly from the original text, rather than mapped to their common meaning. This is fine when creating a summary of disease findings for human readability, but not suitable for creating image labels that can then be used for image classifiers or automated report generation models. Additionally, it was shown that using manual annotation and mapping is incredibly time-consuming and still results in poor image classification. On the other hand, abstractive text summarisation models generate a paraphrase of the main concepts in the original text, potentially even using a different, reduced, vocabulary set. Given that free-text reports have a very large and diverse vocabulary, and MeSH annotations summarise the findings using a much smaller set of key terms, abstractive text summarisation techniques have the potential to be applied to medical text for auto-generating MeSH annotations for radiological images.

In this chapter, a supervised abstractive text summarisation technique was used in order to generate vocabulary-controlled phrases summarising pathology findings from free-text radiology reports. Reports were encoded using convolutional neural networks (CNNs), and pathology summaries (pathology, severity, anatomical location concepts) were encoded using a recurrent neural network. During decoding, an attention model was used over the word-level feature maps, phrase-level feature maps, and the output of the summary encoder at each time step to provide a joint context, that was then used to predict the next word in the summary sequence.

### 5.1.1 Related work

The current s.o.t.a. methods for abstractive text summarisation use recurrent neural network (RNN) [68, 141, 142] where a typically long document is mapped into a short summary using encoder-decoder frameworks. RNNs model documents sequentially and therefore past information does not contribute equally to the document’s encoding. However, important information is not necessarily found in any particular section of the document. This is addressed partially by using LSTMs to ‘forget’ past, unimportant information, and additionally by the use of hierarchical attention [68]. In hierarchical attention, the encoding of the document is made up of word-level attention contexts and sentence-level attention contexts, and the decoder learns to attend to sections of the document when generating the output at each time-step.

However, it is not necessarily essential to model document-level sequential correlation for certain tasks, and may even be more beneficial to capture local word-level correlation using local convolution. CNNs have successfully been applied to document classification [143, 144, 145] and as a way of encoding sentences into features used as inputs to an RNN model [146]. Deep CNNs are able to capture n-gram patterns irrespective of their position in the source text and are therefore more suitable at modelling longer documents with scattered short, key phrases. CNNs are particularly suitable for modelling radiological reports where key terms summarising the visual features of the pathology are found within short phrases within much longer reports consisting of multiple sentences which do not refer to the pathology at all (and may instead refer to the absence of pathologies).

This section is focused on the examination of applying RNN encoder-decoder abstractive text summarisation techniques to generate vocabulary-controlled disease summary outputs from free-text radiological reports. These methods are then compared with a purely CNN encoder to demonstrate the capability of CNNs of capturing word phrase-level features in order to improve the generated summaries.

### 5.1.2 Method

#### Clinical Report Encoder

**Word-level encoder:** Let  $\mathbf{w}_i \in \mathbb{R}^d$  be the  $d$ -dimensional word vector for the  $i$ -th word of report  $\mathbf{r}$ . The text report is thus represented as a concatenation of word embeddings:

$$\mathbf{r} = [\mathbf{w}_1, \dots, \mathbf{w}_i, \dots, \mathbf{w}_N] \in \mathbb{R}^{Nd} \quad (5.1)$$

where  $N$  is the maximum number of words in each report. Filters  $\mathbf{m} \in \mathbb{R}^{d \times kd}$  of multiple kernel widths  $k = k_1, k_2, k_3$  are convolved with a window of  $k$  words to produce a new feature  $c_i$ :

$$c_i = f(\mathbf{m} * \mathbf{x}_{i:i+k-1} + b) \quad (5.2)$$

where  $f$  is a non-linear activation function and  $b$  is a bias term. Multiple filters are applied consecutively to every  $k$ -word window in the report, generating feature maps  $\mathbf{c}_k^{(w)} = [c_1, \dots, c_i, \dots, c_N] \in \mathbb{R}^{Nd}$ . Reports are padded such that, for any kernel width, the length of feature maps is equal to the length of the input sequence, and can therefore be concatenated into the combined feature maps  $\mathbf{c}$ :

$$\mathbf{c} = [c_{k1}^{(w)}; c_{k2}^{(w)}; c_{k3}^{(w)}] \in \mathbb{R}^{N(3d)} \quad (5.3)$$

**Phrase-level encoder:** Much like in image CNNs, successively applying convolutional layers captures higher-level visual features. In our case, a second convolutional layer aims to capture phrase-level semantic structure. Taking inspiration from GoogleNet’s Inception Module [147], we perform multiple filter-width convolutional operations and concatenate with max-pooling. The phrase-level feature maps are obtained in the same way as word-level feature maps:  $\mathbf{c}_k^{(p)} = [c_1, \dots, c_i, \dots, c_N] \in \mathbb{R}^{Nf}$  where  $f$  is the number of feature maps. Stride-1 max-over-time pooling [148] is applied over each of the word-level convolutional feature maps:  $\hat{c}^{(w)} = \max(\mathbf{c}^w)$ . The concatenated output is then a combination of word-level feature maps and phrase-level features maps and max-pooling outputs:

$$\mathbf{p} = [c_{k1}^{(w)}; c_{k1}^{(p)}; c_{k2}^{(p)}; c_{k3}^{(p)}; \hat{c}^{(w)}] \quad (5.4)$$

### Summary Sequence Attention Decoder

We use Bahdanau attention, in a similar implementation as the dynamic attention of Chapter 3, first introduced for neural machine translation [121]. The report summary is modeled using an LSTM. Each LSTM unit has three sigmoid gates to control the internal state: ‘input’, ‘output’ and ‘forget’. For an input summary sequence  $\mathbf{x} = [x_0, \dots, x_t, \dots, x_M]$ , the internal hidden state  $h_t$  and memory state  $c_t$  are updated as follows:

$$\begin{aligned} \mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} \\ &+ \mathbf{i}_t \odot \tanh(W^{(cx)}\mathbf{x}_t + W^{(ch)}\mathbf{h}_{t-1} \\ &+ W^{(cz)}\mathbf{z}_t) \\ \mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \end{aligned} \quad (5.5)$$

where  $\mathbf{z}_t$  is the context vector,  $W^{(cx)}$ ,  $W^{(ch)}$ ,  $W^{(cz)}$  are the trainable weight parameters, and  $\mathbf{i}_t$ ,  $\mathbf{o}_t$  and  $\mathbf{f}_t$  are the input, output and forget gates respectively. The context vector is calculated as the weighted sum of the input sequence *annotations*  $\mathbf{p}$  generated by the report encoder,



computed at every decoder time-step:

$$\mathbf{z}_t = \sum_{n=0}^N \bar{a}_i p_i \quad (5.6)$$

The weights  $\bar{a}_i$  are computed as the normalised score given by function  $f(h_{t-1}, p_i) \mapsto \alpha_i \in \mathbb{R}$ :

$$\bar{a} = \text{softmax}(a) \quad (5.7)$$

The function  $f(x)$  is chosen to be a feed-forward MLP. The initial hidden and memory state of the LSTM are taken as the average over the annotations. The weight  $\bar{a}_i$  can be interpreted as the relative importance of annotation  $p_i$  in generating the next word in the summary sequence  $x_t$  given the previous hidden state  $h_{t-1}$ . The output word probability is then computed by a dense decoder layer:

$$p(x_t | x_0, \dots, x_{t-1}, z_t) = g(x_{t-1}, h_{t-1}, z_t) \quad (5.8)$$

where  $g(x)$  is a single-layer dense MLP with softmax activation. The full schematic of the encoder-decoder network is illustrated in Figure 5.1.

### 5.1.3 Experimental Settings

**Preprocessing** After removal of reports with missing finding and/or impressions, there remained a total of 3,740 exams, 300 of which are used for training and validation each. Pre-processing involved lower-casing, punctuation and non-alpha-numeric character removal, removal of common ‘stopwords’ (‘and’, ‘the’), and words that fell outside of the 99th percentile. Following processing, the average number of words per report was  $21.5 \pm 14.2$  std with a vocabulary of 1,319. For the MeSH annotations, the average number of captions per report was 2.1, with average number of total terms per exam being  $5.2 \pm 5.5$  with a final vocabulary of 126.

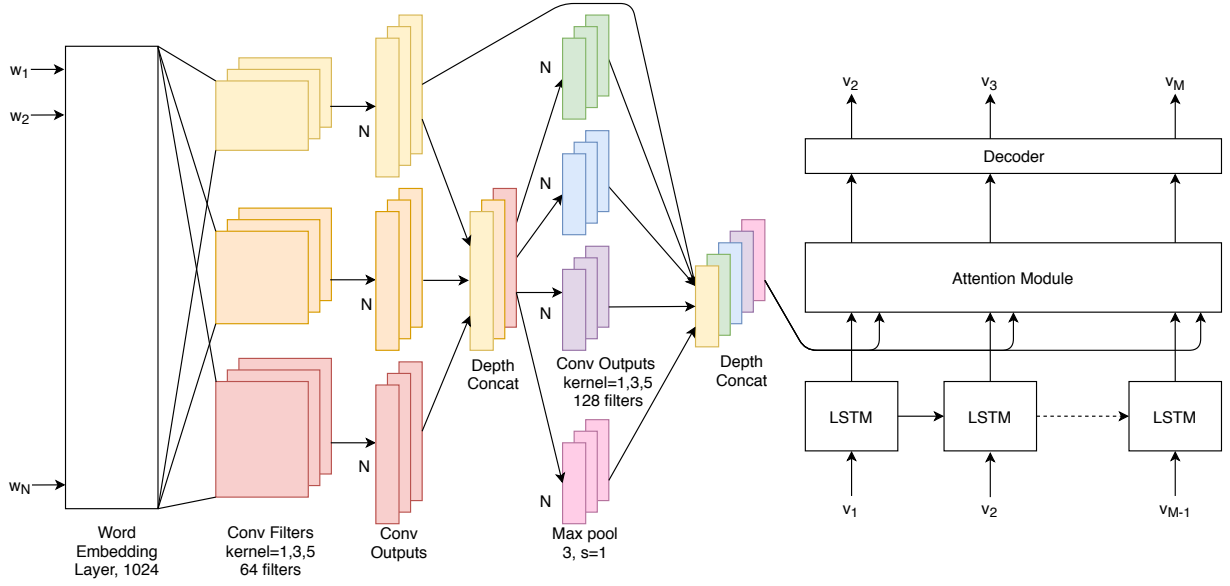


Figure 5.1: TextCNN2Seq-Att model schematic. Best viewed in colour.

**Augmentation:** As ‘normal’ cases made up 37% of the exams, the ‘abnormal’ cases ( where the MeSH caption is not ‘normal’) were augmented by randomly shuffling the sentences of the reports. This results in 5,748 report-MeSH annotation pairs used for training.

**Encoding** Reports and MeSH annotations were cropped/padded to length 37 and 13 respectively (mean+1std+start-token+end-token+unknown). Both the report words and MeSH terms were one-hot-encoded before being passed through the learn-able word embedding layer. For both the word-level and phrase-level convolution, kernel filter widths are set to  $k_1 = 1, k_2 = 3, k_3 = 5$ , with 64 filters each for word-level and 128 phrase-level. Max-pooling on word-level feature maps is done with width 3 filters and stride=1 with padding. The final feature map output of the report encoder was therefore 320. The report word embedding dimension was set to be 1024, the LSTM hidden dimension to be 320 and decoder dimension to 128.

**Training** The model was trained by minimising the categorical cross-entropy between the generated summary sequence and true sequence. At training time, loss is minimised over the training set using SGD (batch size 128, learning rate  $1 \times 10^{-5}$ ), and parameters were updated using Adam [120] optimisation.

**Inference** During inference, first the word- and phrase-level feature maps are generated using the trained CNN. To predict the first word, the LSTM hidden and memory states are initialised with an average over the feature maps, and the start-token is input into the LSTM. Greedy search is performed over the output probabilities i.e. the word with the highest probability is selected as the output. This word is then appended to the LSTM input sequence, the hidden and memory states are updated, and the process repeated until the sequence is of length  $M = 17$ .

#### 5.1.4 Results

The quality of the generated summaries was evaluated by measuring BLUE [43] and ROUGE [44] scores averaged over all the summaries, Table 5.1. The scores of the TextCNN2Seq model are compared against a LSTM seq2seq model with and without attention, and a partial implementation (excluding the pointer/generator) of the hierarchical seq2seq model with attention of [68] as their code was unavailable. The baseline summaries are those extracted by MetaMap under the tags ‘Disease or Syndrome’, ‘Pathologic Function’, ‘Finding’, ‘Qualitative Concept’, ‘Spatial Concept’, ‘Body Part, Organ, or Organ Component’.

The CNN encoder model performs better on all metrics when compared to the sequences encoder models, and MetaMap extracted phrases, and qualitative evaluation (Figure 5.2 reveals that it is capable of attending to short, key phrases in the report in order to generate summaries. Results are averaged over 5 training procedures. It should also be noted that the TextCNN2Seq approached convergence to the optimal Rouge/Bleu metrics with approx 50% of the parameters as the HierSeq2Seq+Att model.

Model	Rouge-1	Rouge-2	Rouge-L	B-1	B-2	B-3	B-4
MetaMap	29.0	2.5	22.6	20.7	3.6	0.3	0.0
Seq2Seq	66.5	24.5	68.9	61.2	19.9	11.0	4.0
Seq2Seq + Att	69.2	27.9	71.4	64.4	23.3	13.8	6.3
HierSeq2Seq + Att	79.2	36.0	79.9	73.2	31.8	22.2	12.3
TextCNN2Seq + Att	<b>81.2</b>	<b>38.7</b>	<b>81.8</b>	<b>74.5</b>	<b>33.4</b>	<b>24.0</b>	<b>13.8</b>

Table 5.1: TextCNN2Seq performance comparison with seq2seq models. Reported are Rouge unigram bigram, and longest sequence F1 scores, and BLEU 1-4-gram scores. All metrics are reported on the test set.

**Report:** interval cabg . sternotomy appear intact . stable , mild degenerative disc disease thoracic spine . visualized bony structures otherwise unremarkable appearance . atherosclerotic calcifications thoracic aorta . clear lungs . peripheral vascular disease

**True MeSH:** atherosclerosis aorta\_thoracic thoracic\_vertebrae degenerative mild

**Pred. Mesh:** atherosclerosis aorta\_thoracic thoracic\_vertebrae degenerative thoracic\_vertebrae

Figure 5.2: Sample output MeSH summary from the TextCNN2Seq+Att with largest and second-largest attention weights highlighted in colour.

## 5.2 Conclusion

Abstractive text summarisation techniques can be applied to generate vocabulary-controlled disease concept summaries, and a purely CNN document encoder improves on previous sequence-encoder approaches. These generated summaries correlate very well with the ground-truth MeSH term annotations created by radiologists, and so can be used successfully as image annotations for report generation, as done in Chapter 3. One of the limitations of this approach is the need for high-quality annotations, such as the MeSH annotations used here. As they are made from drop-down lists, the annotations are limited in vocabulary and are consistent across radiologists. Creating such annotations is a laborious task for radiologists, hence, if some domain knowledge can be injected into the process through some ontological tools, we can potentially be able to use fewer annotations to achieve the same performance. For instance, ontological tools can be used to annotate the text words and phrases in the reports with labels such as 'pathology' and 'anatomy', which can then be encoded and used as additional inputs into the attention network. Additionally, using word2vec to generate the word embedding layer could improve training, as word and phrase vectors will already have contextual meaning prior

to training. Knowledge-based similarity measures require training on a large corpus of radiological reports and clinical text, therefore it can benefit from a larger corpus of unlabeled text reports (if such exist in the public domain).

## 5.3 Related Publications

Gasimova, A. (2019). Automated enriched medical concept generation for chest X-ray images. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support* (pp. 83-92). Springer, Cham.

# Chapter 6

## Image Latent Space Learning for Diagnostic Report Generation

### 6.1 Introduction

Patients that have suffered the symptoms of a stroke have a very short time frame in which to be effectively treated; therefore, it is imperative that radiologists determine the cause of the symptoms in order to provide the appropriate treatment. The majority of strokes are caused by cerebral ischaemia, which can be characterised as reduced blood flow to the brain, causing poor oxygenation that can lead to permanent brain cell death. Both computed tomography (CT) and multi-modal magnetic resonance imaging (MRI) are effective in assessing brain ischaemia, but diffusion-weighted MRI (DWI) is particularly advantageous as it provides highest sensitivity to early ischaemic lesions. In comparison to CT, typical DWI has a much longer acquisition time (1-2 hours vs 20 around minutes) which additionally makes the scans more susceptible to patient motion and subsequent unwanted imaging artefacts. Furthermore, requiring patients to be still without any motion for long periods of time may lead to discomfort. A well-explored approach for accelerating scan-time is through *undersampling* whereby fewer scanner measurements are taken, violating the Nyquist-Shannon sampling theorem and thus introducing aliasing artefacts into the reconstruction of the image. Several studies are focused on the dealiasing of such

images, validating undersampled MRI as an accepted acceleration technique [91, 92, 93, 94, 95, 96, 97].

Assessing the quality of the MR image reconstruction is typically focused on calculating similarity metrics such as peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index between the dealiased reconstruction and the fully-sampled image [149]. This does not, however, guarantee the retention of pathological features necessary for a diagnosis, especially at more aggressive acceleration rates. Therefore, a complementary way of reviewing extremely accelerated images is through the use of real-time diagnostic tasks such as segmentation and classification [150]. In this study, done jointly with Gavin Seegoolam, we explore the automated generation of radiological text reports containing relevant diagnostic and contextual information. The logging of diagnostic reports generated by qualified radiologists is standard hospital protocol. As a result, datasets for studies involving automated text report generation can be acquired directly from hospital archives. In contrast, segmentation and classification tasks require non-standard time-consuming manual annotations. In addition, DWI diagnostic reports typically detail contextual information as well as the presence/absence of an acute lesion, such as anatomical location and severity of the lesion, and being able to auto-generate them will expedite the process of identifying and documenting acute ischemia.

The motivation for learning a pathology-preserving latent space through reconstruction is therefore two-fold: to provide an alternative method to evaluating accelerated images, and to preserve the anatomical structure of the brain in order to improve the report generation task. Previous work on medical image report generation, in this thesis as well as other studies [3, 112, 116, 151, 115], make use of transfer learning from a pre-trained convolutional neural network, trained on natural images. We show that the image representations taken from the pathology-preserving reconstruction network, trained only on brain DWI images, outperform the image embeddings taken from a pre-trained natural image network on the task of report generation.

To this end, we have developed a pipeline that 1) learns an implicit context-preserving manifold of brain DWIs that captures both spatial and pathological information, 2) enforces a latent

code for the accelerated DWIs that performs in a similar fashion to the fully-sampled images 3) utilises these accelerated brain DWI image representations to learn to automatically generate reports using a recurrent neural network. To our knowledge, this is the first demonstration of deep latent space learning for the retention of semantic feature information required for report generation, and the first demonstration of learning to auto-generate reports from brain DWI images.

## 6.2 Related Work

### 6.2.1 Latent space learning of accelerated MRI

Previous work has shown the use of deep latent space learning for performing tasks such as segmentation and reconstruction in the context of accelerated MRI [95, 150]. Accelerated MRI data acquisition is centred around the ability to reconstruct image data in a typically ill-conditioned inverse regression problem. However, certain tasks will only require certain parts of information from the sensor space, called ‘k-space’. For example, approximate motion estimation from cardiac cine MRI can be performed with acceleration rates of 51.2 which corresponds to only 5 lines of k-space [91]. [150] shows that cardiac segmentation can be performed by a single line acquisition in k-space. Inspired by this we explore the use of deep latent space learning for learning diagnostically-relevant contextual image embeddings. Whilst [150] shows that deep latent space learning provides a manifold that can be robust to different undersampling patterns, they also show that at extreme acceleration rates, deep latent space learning can outperform conventional approaches.

The accelerated acquisition of brain DWI has been previously studied in the context of image reconstruction [152, 153, 154, 155]. However, in our study, we explore its use for automated text report generation. We demonstrate how the latent space learned by the accelerated reconstruction network captures both spatial semantic and pathology information required in order to learn to generate reports.



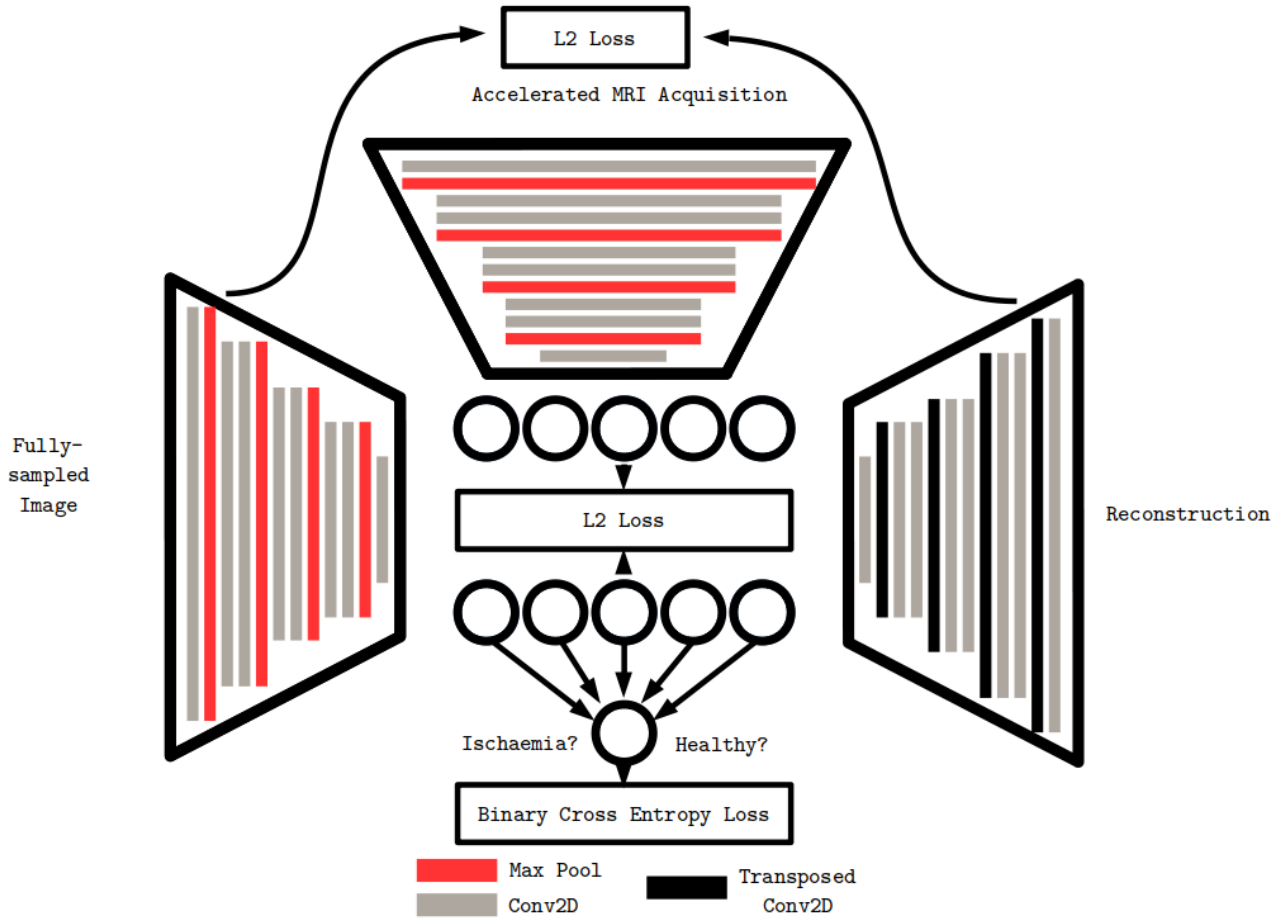


Figure 6.1: An autoencoder is trained to reconstruct the fully-sampled image through an L2 loss. The latent space is conditioned to encode pathological information by performing a classification of ischaemia, trained with a binary cross-entropy loss. The latent space encoding learned at the bottleneck is used as a training target for the encoding branch which only sees the accelerated image.

## 6.3 Methods

Our study accelerates DWI acquisition through aggressive variable-density Cartesian under-sampling as has been studied in several previous works such as [91, 150]. In our study, we start with attempting a zero-fill reconstruction whereby the lines in k-spaces that are not acquired are filled with zeros. An example of a fully sampled image and a corresponding undersampled image is shown in Figure 6.2. For all acceleration rates, we always sample the two most central lines in k-space whilst the other lines are acquired following a Gaussian distribution centered at the point of highest energy in k-space. During training, undersampling masks are generated on the fly and images are also augmented with additional rotations and translations.

### 6.3.1 Latent space learning

In our approach, we use an autoencoder network that takes as input the original fully-sampled DWI brain MRI. The purpose of this is to learn a latent space at the bottleneck that contains spatial and contextual information that may be useful for a text report generator. In particular, we manipulate the embedding manifold toward one more suitable for text report generation by introducing an ischaemia-classification loss as a regulariser. This loss can be summarised by equation (6.1) where an Adam optimiser with learning rate  $1.0 \times 10^{-5}$ ,  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  was used.

$$L(x, y) = ||D(E(x)) - x||_2^2 - \gamma(y \log C(E(x)) + (1 - y) \log(1 - C(E(x)))), \quad (6.1)$$

where  $E$ ,  $D$  and  $C$  are the encoder, decoder and classifier networks (from figure 1) respectively,  $x$  is our fully-sampled image,  $y$  is a binary classification label for ischaemia and  $\gamma = 8000$ . We can measure the performance of the latent space learnt as a combination of reconstruction error (in particular of the ischaemia) and of the classification error.

Along side this, we use a structurally-identical encoding branch to learn a latent space for the accelerated MRI acquisition. We use the approach of performing a zero-fill reconstruction whereby after convolutional layers can be used to identify aliasing artefacts as directly relevant image features themselves. The latent space is trained against the bottleneck of the autoencoder using an L2 loss and another Adam optimizer with the same optimizer parameters. This is summarised in Figure 6.1 and in equation (6.2). Note, for each acceleration rate used in our study, a unique encoder is learned to generate the required latent space. An advantage of deep latent space learning is that we can train the specific encoder associated with different acceleration rates towards the same manifold which avoids the need for retraining of the text report generator model.

$$L(x, x_{acc}) = ||E(x) - x_{acc}||_2^2, \quad (6.2)$$

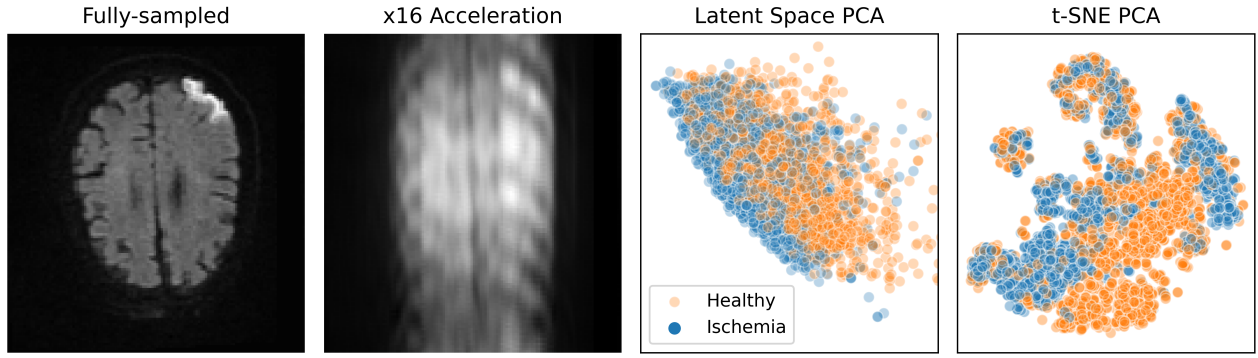


Figure 6.2: Left to right: (1) An example of a brain with ischaemia (2) The corresponding x16 accelerated image is zero-fill reconstructed from k-space using a 2D Fourier Transform. Note that this image suffers from heavy aliasing artefacts. (3) A projection of the first two principle components in a PCA analysis of the latent space. Some clustering can be seen (4) a t-SNE projection of the latent space showing clear clustering.

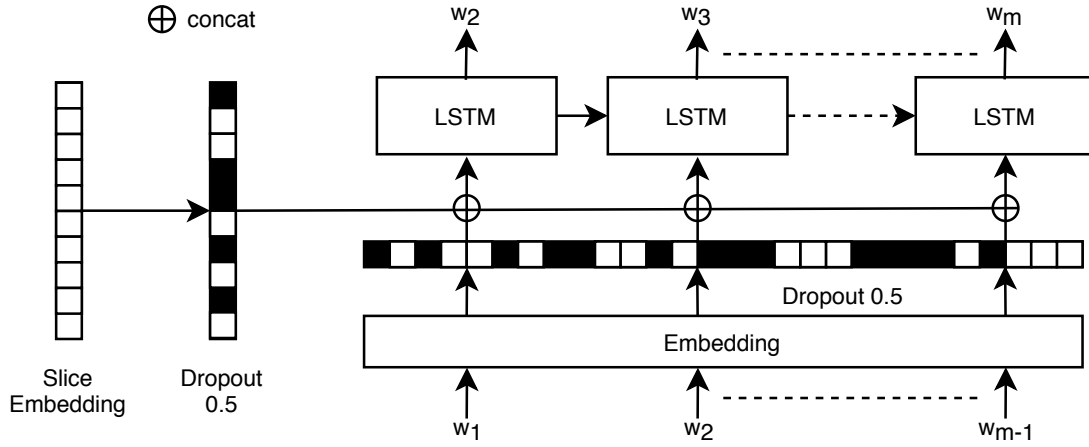


Figure 6.3: Clinical report generation model from accelerated image latent space embeddings.

where  $x_{acc}$  is our accelerated, aliased image and  $E_{acc}$  is our encoding branch for the accelerated images.

We can measure the performance of the ‘accelerated’ latent space by seeing how well it reconstructs images (in particular of the ischemia) by using reconstruction part of the autoencoder, and then separately the classification loss from the manifold classification regulariser. The latent space is verified by reconstructions and through decompositional projections shown in Figure 6.2.

### 6.3.2 Report generation model

We use a report generation model based on [156] where the report word sequence is modelled using the Long Short-Term Memory (LSTM)[57], and conditioned on image embeddings at each time step through concatenation at the input to the LSTM. At each time step, the input, output and forget gates control how much of the previous time steps is propagated through to the output. For an input embedding sequence  $\{x_1, \dots, x_n\}$  where  $x_i \in \mathbb{R}^D$ , the internal hidden state  $h_t \in \mathbb{R}^h$  and memory state  $m_t \in \mathbb{R}^m$  are updated as follows:

$$\begin{aligned} h_t &= f_t \odot h_{t-1} + i_t \odot \tanh(W^{(hx)}x_t + W^{(hm)}m_{t-1}) \\ m_t &= o_t \odot \tanh(h_t) \end{aligned} \tag{6.3}$$

where  $x_t \in \mathbb{R}^D$  is the concatenation of the latent space image embedding and word embedding at time step  $t$ ,  $W^{(hx)}$  and  $W^{(hm)}$  are the trainable weight parameters, and  $i_t$ ,  $o_t$  and  $f_t$  are the input, output and forget gates respectively. The model architecture is illustrated in Figure 6.3. We additionally add Dropout layers after image and word embeddings to force the model to condition on both thus regularising training.

## 6.4 Dataset

The dataset consists of 1226 DWI scans and corresponding radiological reports of acute stroke patients. All the images and reports were fully anonymised and ethical approval was granted by Imperial College Joint Regulatory Office. The scans were obtained from three different scanners (Siemens) with the following acquisition parameters: field strength: 1.5-3 T; slice thickness: 5 mm; slice spacing: 1.0-1.5 mm; pixel size in x-y plane: 1.40×1.40 or 1.80×1.80 mm; matrix size: (19-23)×(128×128) or (192×192); field of view: 230×230 or 267×267; echo time 90-93 ms; repetition time 3200-4600 ms; flip angle 90; phase encoding steps: 95-145. The scans were pre-processed according to the steps outlined in [5]: images were resampled into uniform pixel

size of  $1.6 \times 1.6$ mm, and pixel intensities were normalised to zero mean and unit variance. The number of slices per image varies between 7 and 52, and the slice dimensions are  $128 \times 128$ .

Each report contains between 1 and 2 sentences summarising the presence or absence of the pathology, a visual description, and its location within the brain. In addition, each exam is assigned a diagnostic label as part of hospital protocol: 54% were diagnosed ‘no acute infarct’, 46% were diagnosed ‘acute infarct’. The remaining, which made up a total of  $<1\%$  and included diagnoses such as ‘unknown’, ‘haematoma’, ‘tumour’, were removed for the purpose of training. Processing was done on the reports to remove words outside the 99th percentile, exams with empty reports were removed, leaving a total of 1104 exams, total vocab length 1021, mean words per exam 10.8, std. 6.3.

In order to simplify the problem, we created a 2D dataset of acute and non-acute (normal) slices from these images. For the acute set, we used the brain ischemia segmentation network developed by Chen et al.[5] to segment the images labelled with acute ischemia, thresholded at 0.8, and selected slices where the total area of ischemia was  $>10$  pixels. For the normal set, we sampled slices from the non-acute labelled images according to the same axial plane distribution as the acute set.

## 6.5 Experiments

The accelerated latent space model was trained to reduce the sum of the reconstruction and classification losses defined in Section 6.3.1. Training was terminated when validation loss no longer decreased. Once trained, the encoder was used to generate embeddings for accelerated images, with augmentation in the form of rotation. These embeddings were then used to train the report generation model.

Reports were padded with ‘start’ and ‘end’ tokens to length 19 (mean + 1std. + ‘start’ + ‘end’). The word embedding layer maps one-hot encoded word embeddings into a 256 dimensional space. The LSTM hidden state is also set to dim 256, and the LSTM units are unrolled up to 19 time steps. We train the model on non-accelerated latent embeddings and their associated

reports by minimising the categorical cross-entropy loss over the generated words. All models are trained with batch size 128, using Adam optimisation [139], learning rate=0.0001 for a maximum of 300 epochs, with early termination of training based on validation loss.

## 6.6 Results

The intermediate evaluation of the accelerated latent space model was performed by calculating the F1, mean-squared-error (MSE) and peak signal-to-noise ratio (PSNR), reported in Table 6.1. The scores on the fully-sampled images are reported alongside those of accelerated images since the models were re-trained for each acceleration rate, and drop in performance from fully-sampled to accelerated images is relative to the individual model. However, the F1, MSE and PSNR scores are fairly consistent for all but two of the accelerations: x4 and x64. The models trained with x4 and x64 acceleration both had significantly lower MSE of 26.49 and 26.66 respectively, and much higher PSNR of 38.11 and 38.69 respectively. Their F1 validation scores are, on the other hand, relatively lower. This could mean that these two models converged to a different optimum whereby the reconstruction was favoured over the classification. Excluding these two, we can see a general trend of reduced F1 of the accelerated images in both the train and validation data. However, there is no significant drop in MSE or PSNR of the reconstructed images. This implies that it becomes more difficult to retain pathologic information within images with increased acceleration as opposed to structural information.

For the report generation task, inference was performed by first sampling from the LSTM using a ‘start’ token concatenated with the accelerated embeddings, and consequently appending the output word embedding to the input and sampling until an ‘end’ token was reached. The quality of the generated reports was evaluated by measuring BLUE [43] and ROUGE [157] scores averaged over all the reports, report in Table 6.2. We observe that the both the BLEU and ROUGE scores decrease with increasingly accelerated images, as expected. We note that there is a significant reduction in performance between the x4 and x8 accelerated images possibly due to some contextual information not being captured by the latent space.

We also assess the sampled reports qualitatively in Figure 6.4. We observe fairly coherent reports for all accelerations, with x2 and x4 correctly identifying the presence/absence of ischemia as well as the location. Note: the last example shows a text report that was ischemic but was classified as healthy. This is likely to have confused the latent code for this example resulting in poor text report generations.

Table 6.1: F1, MSE and PSNR metrics of ground truth and accelerated MRI. F1 score is taken at the output of the classifier module and MSE and PSNR metrics from the output of the reconstruction module of the autoencoder. All metrics are reported as average over samples.

	Fully-sampled MRI						Accelerated MRI					
	F1		MSE		PSNR		F1		MSE		PSNR	
	train	val	train	val	train	val	train	val	train	val	train	val
Acc. $\times$ 1	0.89	0.78	148.03	171.23	34.85	29.98	0.89	0.78	148.03	171.23	34.85	29.98
Acc. $\times$ 2	0.87	0.83	151.53	173.77	34.68	29.92	0.84	0.77	151.53	173.77	34.68	29.92
Acc. $\times$ 4	0.86	0.68	26.49	26.37	43.54	38.11	0.78	0.66	27.87	24.56	43.29	38.42
Acc. $\times$ 8	0.85	0.72	151.91	173.77	34.69	29.92	0.73	0.69	151.91	173.77	34.69	29.92
Acc. $\times$ 16	0.85	0.75	152.28	173.77	34.66	29.92	0.74	0.75	152.28	173.77	34.66	29.92
Acc. $\times$ 64	0.86	0.77	26.66	23.09	43.64	38.69	0.74	0.67	26.45	22.38	43.70	38.82

Table 6.2: BLEU1,2,3,4-gram and ROUGE1 F1, precision (P) and recall (R) metric comparisons on increasingly accelerated image embeddings.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	R-1 F1	R-1 P	R-1 R
ResNet embedding models							
SERepGen-merge	20.8	12.9	5.1	1.6	31.6	44.2	26.7
DARepGen	21.5	13.4	5.8	2.3	31.8	43.1	27.6
Latent space models							
Acc. $\times$ 1	38.12	27.26	20.28	15.59	47.10	52.89	44.96
Acc. $\times$ 2	34.07	23.31	15.55	11.57	44.00	51.86	40.68
Acc. $\times$ 4	31.36	19.42	12.29	8.31	41.17	48.09	38.80
Acc. $\times$ 8	21.32	10.37	5.06	2.55	29.53	32.92	29.52
Acc. $\times$ 64	21.58	11.11	4.97	2.35	30.39	35.10	29.07

## 6.7 3D-DWI Extension

Using the 3D DWI volumes directly removes the need for time-consuming tasks such as segmentation, slice selection and out-of-distribution data discarding that was performed when creating the 2D axial-slice dataset. We therefore performed a study to evaluate the accelerated latent space report generation framework when trained directly on 3D volumes. Additionally,

we performed an ablation study on the auxiliary tasks of classification and image reconstruction without accelerated acquisition to determine the optimal balance between the two loss functions. The encoder was tuned through a set of experiments with a validation set on the report generation task. Finally, the encoder model is tuned with accelerated acquisitions for DWI report generation without an intermediate reconstruction phase.

**Ablation study** One goal of our study was to ascertain the balance between our auxiliary tasks for the latent space learning from 3D volumes and thus optimise the parameter  $\gamma$  in Equation 6.1. We assess the quality of the latent space by training and then sampling from the report generation model, and evaluating the predicted reports against the true reports using the BLEU metric, averaged across samples. The results are shown in Table 6.3. After evaluating on the validation set, we found that the classification only model performed best on BLEU-1, however, when  $\gamma = 1e10$ , the model performs better on higher n-gram BLEU metrics. Higher BLEU metrics on longer n-grams indicates that a more contextual report is learned (i.e. greater overlap of 2, 3, and 4 sequential words). This is consistent with our hypothesis that the auxiliary task of reconstruction improves the semantic-preserving ability of the latent space.

Table 6.3: Results of ablation study

Model	Acc.	Precision	Recall	B-1	B-2	B-3	B-4
Classification Only	0.79	0.85	0.67	<b>21.10</b>	8.73	4.14	0.47
$\gamma = 1e8$	0.54	0.50	0.53	12.28	3.02	2.10	0.00
$\gamma = 1e9$	0.62	0.58	0.60	18.49	9.02	1.85	0.70
$\gamma = 1e10$	0.78	0.82	0.65	20.83	<b>11.82</b>	<b>8.62</b>	<b>7.59</b>
$\gamma = 1e11$	0.77	0.76	0.73	18.68	9.88	2.65	1.32

**Accelerated DWI report generation** With the optimal hyperparameters chosen for the auxiliary learning task, the ‘accelerated’ encoder was trained to produce the same embeddings of ‘fully-sampled’ encoder via an L2 loss. The result was that the semantic embeddings were produced from extremely accelerated acquisitions of pathological brain volumes. We found that even highly accelerated acquisitions were able to be encoded to representations very close to that of fully-sampled acquisitions. These embeddings were then used to produce the associated accelerated radiological text report. The BLEU scores evaluated on the test dataset for each



acceleration rate is shown in Figure 6.5. As expected, higher acceleration rates lead to worse BLEU scores but it is important to note that even at x8 acceleration, the reports are still of good quality as shown in the samples in Table 6.4.

Table 6.4: Sample ground truth and generated reports from fully sampled and undersampled 3D brain DWI. Correctly identified concepts are highlighted.

True:	no, acute, ischaemic, lesion, intracranial, haemorrhage
No Acc.:	no, acute, infarct, intra, extraaxial, haemorrhage, demonstrated
Acc. x8:	per, mri, study, performed, earlier, today, no, acute, intracranial, abnormality, evident
True:	multiple, small, acute, infarcts, scattered, throughout, left, superior, temporal, inferior, frontal, superior, parietal, lobe
No Acc.:	acute, cortical, left, mca, territory, infarct, within, left, parietal, lobe
Acc. x8:	appear, small, acute, left, left, superior
True:	restricted, diffusion, involving, left, posterior, temporal, lobe, external, capsule, posteriorly, extending, left, parietal, lobe, appearances, keeping, acute, left, mca, infarct
No Acc.:	several, small, foci, restricted, diffusion, within, left, parietal, lobe, keeping, acute, right, mca, territory
Acc. x8:	minor, microangiopathic, ischaemic, changes, involving, left, occipital, lobe, extending, posterior, internal, capsule
True:	no, acute, infarction, intracranial, haemorrhage
No Acc.:	no, acute, infarct, haemorrhage, demonstrated
Acc. x8:	no, acute, infarct, evidence, recent, haemorrhage, demonstrated
True:	acute, infarcts, seen, left, frontal, corona
No Acc.:	acute, infarct, left, corona, radiata, involving, june, posterior, limb, left, internal, capsule
Acc. x8:	acute, infarct, left, corona, radiata
True:	acute, infarction, right, mca, territory, involving, caudate, nucleus, anterior, limb, internal, capsule, entire, lentiform, nucleus
No Acc.:	complete, right, aca, mca, territory, infarcts
Acc. x8:	note, made, extensive, right, mca, territory, subacute, infarct, involving, right, corpus, striatum, corona, radiata, external, capsule, insular, right, frontoparietal, cortices, confluent, large, infarct

## 6.8 Conclusion

We demonstrate how a latent space capturing pathological and spatial information can be learned from accelerated brain DWI images and subsequently used to train a diagnostic report generation network with promising results. In future works, we wish to explore radial under-sampling trajectories for DWI brain imaging which are expected to provide improved diagnostic

embeddings. We also present a streamlined pipeline that directly transforms a 3D accelerated DWI acquisition into a semantically-rich embedding space, from which radiological text reports can be learned. Another aim of this preliminary study was to ascertain the use of balanced reconstruction and classification auxiliary tasks for the generation of image embeddings in the context of accelerated radiological report generation. Future progress from this preliminary study includes investigations into different acceleration schemes and more appropriate language models.

## 6.9 Related Publications

Gasimova, A., Seegoolam, G., Chen, L., Bentley, P., Rueckert, D. (2020, October). Spatial Semantic-Preserving Latent Space Learning for Accelerated DWI Diagnostic Report Generation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 333-342). Springer, Cham.

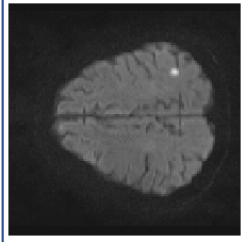
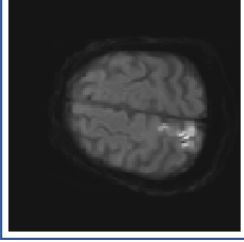
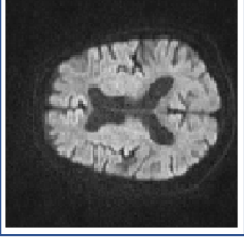
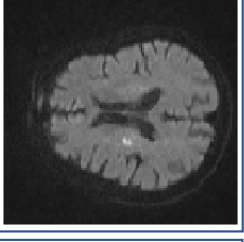
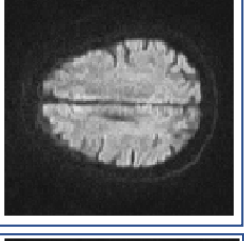
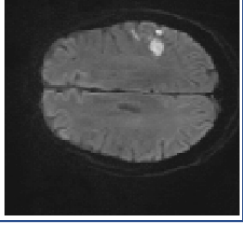
	<p><b>Acute:</b> <b>Y True report:</b> restricted diffusion right posterior insula several additional foci within parietal lobe keeping multiple small right mca infarcts</p> <p><b>Acc x1:</b> tiny foci restricted diffusion within right parietal lobe right</p> <p><b>Acc x2:</b> acute embolic looking infarcts within right parietal lobe</p> <p><b>Acc x4:</b> acute infarcts within right mca territory bilaterally</p> <p><b>Acc x8:</b> tiny acute cortical infarcts right mca territory involving right frontal parietal</p> <p><b>Acc x64:</b> several cortical **unknown** infarcts within right parietal lobe</p>
	<p><b>Acute:</b> <b>Y True report:</b> cortical restricted diffusion centred left parasagittal frontal parietal region involving **unknown** lobule superior</p> <p><b>Acc x1:</b> cortical restricted diffusion centred left parasagittal parietal region involving posterior</p> <p><b>Acc x2:</b> multiple cortical subcortical acute infarcts centred left corona radiata</p> <p><b>Acc x4:</b> cortical subcortical acute ischaemic changes involving left parietal region</p> <p><b>Acc x8:</b> acute cortical infarct centred left parietal region</p> <p><b>Acc x64:</b> several acute infarction within left mca territory</p>
	<p><b>Acute:</b> <b>N True report:</b> no acute infarcts demonstrated</p> <p><b>Acc x1:</b> no acute intracranial abnormality identified intracranial haemorrhage</p> <p><b>Acc x2:</b> no acute intracranial abnormality demonstrated particular no acute infarct intra extraaxial haemorrhage</p> <p><b>Acc x4:</b> no acute ischaemic changes</p> <p><b>Acc x8:</b> no acute ischaemic lesion intracranial haemorrhage</p> <p><b>Acc x64:</b> no acute infarction intracranial haemorrhage</p>
	<p><b>Acute:</b> <b>Y True report:</b> small acute white matter infarct left corona radiata</p> <p><b>Acc x1:</b> small area acute infarct left corona radiata</p> <p><b>Acc x2:</b> small area restricted diffusion within left mca territory infarct</p> <p><b>Acc x4:</b> focal area signal within left corona radiata</p> <p><b>Acc x8:</b> multiple small foci acute ischaemia left gyrus</p> <p><b>Acc x64:</b> area restricted diffusion accompanying flair within left corona radiata suggest **unknown**</p>
	<p><b>Acute:</b> <b>N True report:</b> no acute infarction</p> <p><b>Acc x1:</b> no acute ischaemic lesion intracranial haemorrhage</p> <p><b>Acc x2:</b> no acute infarct</p> <p><b>Acc x4:</b> no acute ischaemic lesion</p> <p><b>Acc x8:</b> small acute infarct centred left parietal region</p> <p><b>Acc x64:</b> no acute ischaemic lesion</p>
	<p><b>Acute:</b> <b>N True report:</b> modest volume acute right middle cerebral artery territory ischaemia noted no evidence haemorrhagic transformation</p> <p><b>Acc x1:</b> no evidence acute infarct</p> <p><b>Acc x2:</b> no acute infarct intra extraaxial haemorrhage</p> <p><b>Acc x4:</b> no acute intracranial haemorrhage demonstrated</p> <p><b>Acc x8:</b> acute infarcts within right mca territory areas days</p> <p><b>Acc x64:</b> focal subcortical restricted diffusion within left parietal lobe keeping acute infarct</p>

Figure 6.4: Sample brain slices and associated reports generated from non-accelerated and increasingly accelerated image embeddings. Correctly identified pathology (acute/non-acute) and spatial contexts are highlighted in blue.

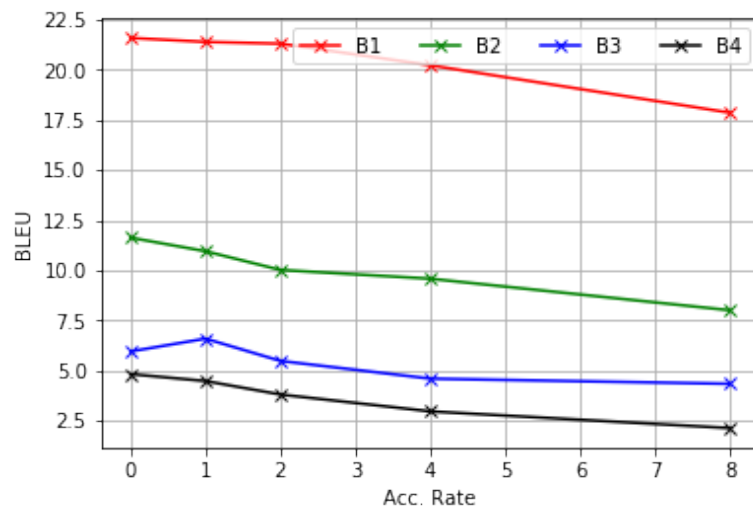


Figure 6.5: Average BLEU-n scores of accelerated brain volumes.

# Chapter 7

## Conclusion

### 7.1 Summary of Thesis Achievements

In this thesis, the problem of learning to auto-generate diagnostics from radiological images by training machine learning algorithms on radiological reports is approached in three ways. By taking inspiration from natural image captioning, the first approach was to train a language generation model on the reports, and explore ways in which to condition the generated words on static vector radiological image features. It was demonstrated that the best performance according to BLEU, ROUGE and DAPS metrics was the encoder-decoder model SERepGen-merge, in which the LSTM encoder is trained to encode the textual reports, a pre-trained CNN encoder is used to extract static image features, and a dense neural network acts as a decoder of the two encoders. The better performance is attributed to the fact that LSTMs were designed to model sequential dependence, and therefore more suitable at modelling purely linguistic features, as opposed to the combination of artificial image-word sequences, which are not sequentially dependent. Although it was difficult to evaluate the generated reports against the true ones using BLEU and ROUGE, the DAPS averaged F1 scores gave an indication that the predicted diseases were better than random.

The static encoder-decoder model achieved poorer results when trained on more complex datasets, including the chest X-rays dataset consisting of multiple views (posterior-anterior

and lateral) and more complex reports. Reports were made more complex by including the previously filtered out MeSH annotations to make reports consist of references to multiple disease. The disease F1 scores were lower when training on the multi-view and multi-disease IU-CX dataset. One reason may be that a static image embedding model, especially one that is extracted using a network pre-trained on natural images as opposed to chest X-rays, cannot capture the full range of disease features, especially more subtle ones that even radiologists have trouble distinguishing. Additionally, a static image feature does not encode location-specific information, meaning that generating a separate location per identified disease is technically impossible (though there may be some implicit location information being encoded into the feature vector, it is still impossible to separate out this information for diseases in multiple locations). By taking lower convolutional layer outputs, prior to max-pooling (and therefore the loss of location-specific features), and using attention mechanism of Mnih et al. [105], the quality of the generated reports when training on multi-view image and free-text reports. (according to BLEU and ROUGE) was moderately improved.

Although BLEU and ROUGE had improved, it was still difficult to assess the predicted reports for diagnostic content. Therefore, the next two chapters explored ways of extracting diagnostic information from free-text radiological reports. The first approach was a combination of using MetaMap to extract and tag medical concepts, followed by representing the words and phrases using word vectors, and grouping them together by meaning using k-means. The resulting clusters were evaluated on their ability to be predicted from images by training an image classifier and evaluating average per-class and per-instance precision and recall. Results indicated that, on a per-class performance basis, all of the methods failed at providing distinguishable labels for the purpose of image classification. This may be due to a number of assumptions that needed to be made in creating these labels. For instance, extracting different MetaMap tag combinations made a large difference in the number of identified disease concepts and the proportion of exams identified as being ‘normal’. The inherent ambiguity of natural language makes it difficult for a pattern-matching algorithm to determine the intended meaning of a phrase, and whether to tag it as a ‘disease’, a ‘finding’ or ‘abnormality’. Even if all disease phrases are correctly extracted, grouping them together under a shared disease label

also requires a number of assumptions. In order to group phrases such as ‘pleural effusion’ and ‘pleural effusion bilateral’, it was assumed that their vector representations (made either through tf-idf or word2vec) would be close together, and therefore could be clustered using distance measures. However, this assumption did not hold true for words and phrases that appeared very rarely within the reports, which resulted in their representations being clustered together. Lastly, the main assumption that a chest X-ray image can be labeled with a single, independent disease label was also challenged by the fact that the image classifier achieved poor results even when trained on manual disease label annotations taken from MeSH.

An alternative to single disease-label extraction from free-text reports was developed and evaluated in the form of abstractive text summarisation. In contrast to MetaMap’s extractive approach, the abstractive encoder-decoder frameworks were better able to match the manual MeSH annotations of the IU-CX free-text reports. The best performing encoder-decoder according the BLEU and ROUGE metrics between the generated summaries and original MeSH was the TextCNN2Seq. The free-text encoder was inspired by successful application of CNNs to document encoding and classification, and a CNN architecture was designed that was able to capture local word-level and phrase-level features by successively applying convolutional layers over word embeddings. An LSTM decoder with attention over the CNN output features was then trained to generate the MeSH annotations from free-text reports. This technique was successful in mapping long, free-text reports with a large and diverse vocabulary, into short, vocab-controlled MeSH phrases consisting of pathology, severity and anatomical location concepts. There are several potential applications of this technique, including training a radiological report generation model on the more concise summaries instead of the original free-text reports, or using the annotations for image retrieval. Having vocab-controlled annotations is important to both of these tasks as it greatly reduces the variety of ways of referring to a disease, making annotations more consistent across images.

Learning to auto-generate radiological reports from images requires an image encoder capable of capturing the semantic information described in the report. These are typically a combination of presence of disease, disease severity and anatomical location, therefore an encoder must capture both pathological and anatomical information. One method of encoding both pathological

information and anatomical structure is through autoencoders trained for disease classification and image reconstruction, which is explored as a method of encoding ischemic brain DWIs. This method had several advantages over the use of pre-trained CNN networks as image encoders. Firstly, the networks could be trained on 3D images, which allowed for the use of full 3D brain volumes and removed the need for preprocessing the volumes into 2D using segmentation and slice selection. Secondly, training the autoencoder to perform accelerated image reconstruction mean the report generation model could also generate reports for accelerated images. This is especially useful for magnetic resonance imaging examinations where the treatment is time-sensitive, such as in the case of ischemic strokes. Lastly, the reports generated by the model trained using the autoencoder latent space outperformed those trained using ResNet50 pre-trained networks, according to both BLEU and ROUGE metrics, meaning the learned latent space was better able to capture the spatial and pathological information from the images.

## 7.2 Limitations and Future Work

The main limitation that all of the described methods have in common is in assessing the evaluation metrics, namely BLEU and ROUGE. N-gram matching approaches to evaluating NLP tasks have long been critiqued in literature. Several studies questioned the validity of the claim that BLEU correlates well with human judgement [158, 159, 160] and found low to no correlation with human judgement for both meaning and fluency. The proposed alternative method of categorising words under tags and evaluating sudo-precision and recall metrics, introduced as DAPS in this thesis, does provide a better metric for the presence/absence of specific words that are relevant in evaluating the reports for clinical value, but are not ideal as it requires a corpus-specific set of categorised terms, which is very difficult to create in an automated way, especially from free-text reports.

Additionally, BLEU, ROUGE and DAPS are all discrete measures of text similarity, which is inherently problematic due to the presence of synonyms and word modifiers. A more suitable metric for measuring text similarity is by representing the words in continuous space,



where words with similar meaning appear near each other within this space. These measures are model-specific, i.e. they require that a model be trained on a corpus in order to build this ‘meaning’ space. One such example is BERTscore [161] which computes cosine similarity between word embeddings using the Bidirectional Encoder Representations from Transformer (BERT) language model [162]. The BERT model could be trained on the corpus of radiological free-text reports, and could potentially result in all the different variations of a disease concept to be mapped to the same vector space, allowing for generated reports to be evaluated on their meaning irrespective of the exact wording.

Another aspect of natural language that limits the validity of the metrics (even in the case of continuous space measures) is the use of uncertainty and negation, such ‘there may be <disease >’ or ‘there is no presence of <disease >’. The problem was largely avoided by using NegEx [118] to tag and remove negated phrases, but this simple pattern-matching system fails on more complicated sentences, and fails to detect uncertainty in findings altogether. There has been an increasing focus on identifying negation and uncertainty in medical text using machine learning, such as the negation classifiers of Morante et al. [163], and negation and speculation classifier of Diaz et al. [164]. These classifiers can be used in conjunction with the current evaluation metrics in order to identify which phrases have been negated or are speculative to help determine whether the correct diseases have been negated.

Lastly, the main limitation of applying machine learning techniques to radiological images and reports is the lack of large, curated datasets. Even though data is indeed available in large quantities from hospitals, the methods demonstrated in this thesis required manual data curation and would benefit significantly from pre-trained networks. For instance, instead of one-hot encoding words and relying on the report generation training to result in meaningful word encodings, the word embeddings can be initialised using a medical concept specific language model pre-trained on a large corpus of medical text. These contextual word embeddings, such as ELMo [165] and BERT [162], have been shown to improve the performance of many NLP tasks, and have recently been applied to the medical domain through Clinical BERT [166], BioBERT [167] and Med-BERT [168].

# Bibliography

- [1] Open-i: An open access biomedical search engine. URL <https://openi.nlm.nih.gov/>.
- [2] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
- [3] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506, 2016.
- [4] Robin Smithuis and Otto van Delden. The radiology assistant: Basic interpretation. URL <https://radiologyassistant.nl/chest/chest-x-ray/basic-interpretation>.
- [5] Liang Chen, Paul Bentley, and Daniel Rueckert. Fully automatic acute ischemic lesion segmentation in dwi using convolutional neural networks. *NeuroImage: Clinical*, 15: 633–643, 2017.
- [6] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.
- [7] Sanja Fidler, Abhishek Sharma, and Raquel Urtasun. A sentence is worth a thousand pixels. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1995–2002, 2013.

- [8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [9] Xinlei Chen and C Lawrence Zitnick. Mind’s eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2422–2431, 2015.
- [10] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.
- [11] Fred Winsberg, Milton Elkin, Josiah Macy Jr, Victoria Bordaz, and William Weymouth. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*, 89(2):211–215, 1967.
- [12] Samuel G Armato, Maryellen L Giger, Catherine J Moran, James T Blackburn, Kunio Doi, and Heber MacMahon. Computerized detection of pulmonary nodules on ct scans. *Radiographics*, 19(5):1303–1311, 1999.
- [13] Maryellen Lisseak Giger, Zhimin Huo, Matthew A Kupinski, and Carl J Vyborny. Computer-aided diagnosis in mammography. *Handbook of medical imaging*, 2:915–1004, 2000.
- [14] Kenji Suzuki, Samuel G Armato III, Feng Li, Shusuke Sone, and Kunio Doi. Massive training artificial neural network (mtann) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. *Medical physics*, 30(7):1602–1617, 2003.
- [15] Temesguen Messay, Russell C Hardie, and Steven K Rogers. A new computationally efficient cad system for pulmonary nodule detection in ct imagery. *Medical image analysis*, 14(3):390–406, 2010.

- [16] Sumit K Shah, Michael F McNitt-Gray, Sarah R Rogers, Jonathan G Goldin, Robert D Suh, James W Sayre, Iva Petkovska, Hyun J Kim, and Denise R Aberle. Computer aided characterization of the solitary pulmonary nodule using volumetric and contrast enhancement features<sup>1</sup>. *Academic radiology*, 12(10):1310–1319, 2005.
- [17] Ted W Way, Berkman Sahiner, Heang-Ping Chan, Lubomir Hadjiiski, Philip N Cascade, Aamer Chughtai, Naama Bogot, and Ella Kazerooni. Computer-aided diagnosis of pulmonary nodules on ct scans: improvement of classification performance with nodule surface features. *Medical physics*, 36(7):3086–3098, 2009.
- [18] Kenji Suzuki, Feng Li, Shusuke Sone, and Kunio Doi. Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose ct by use of massive training artificial neural network. *IEEE transactions on medical imaging*, 24(9):1138–1150, 2005.
- [19] Masahiro Endo, Takeshi Aramaki, Koiku Asakura, Michihisa Moriguchi, Masahiro Akimaru, Akira Osawa, Ryuji Hisanaga, Yoshiyuki Moriya, Kazuo Shimura, Hiroyoshi Furukawa, et al. Content-based image-retrieval system in chest computed tomography for a solitary pulmonary nodule: method and preliminary experiments. *International journal of computer assisted radiology and surgery*, 7(2):331–338, 2012.
- [20] Macedo Firmino, Giovani Angelo, Higor Morais, Marcel R Dantas, and Ricardo Valentin. Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy. *Biomedical engineering online*, 15(1):1–17, 2016.
- [21] K. Doi. Current status and future potential of computer-aided diagnosis in medical imaging. *British Journal of Radiology*, 78(SPEC. ISS.):21–30, 2005. ISSN 00071285. doi: 10.1259/bjr/82933343.
- [22] Jun-Ichiro Toriwaki, Yasuhito Suenaga, Toshio Negoro, and Teruo Fukumura. Pattern recognition of chest x-ray images. *Computer Graphics and Image Processing*, 2(3-4):252–271, 1973.

- [23] Rafael Wiemker, Patrick Rogalla, Andre Zwartkruis, and Thomas Blaffert. Computer-aided lung nodule detection on high-resolution ct data. In *Medical Imaging 2002: Image Processing*, volume 4684, pages 677–688. International Society for Optics and Photonics, 2002.
- [24] Rafael Wiemker, Thomas Bülow, and Thomas Blaffert. Unsupervised extraction of the pulmonary interlobar fissures from high resolution thoracic ct data. In *International Congress Series*, volume 1281, pages 1121–1126. Elsevier, 2005.
- [25] Michael F McNitt-Gray, Eric M Hart, Nathaniel Wyckoff, James W Sayre, Jonathan G Goldin, and Denise R Aberle. A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution ct: preliminary results. *Medical physics*, 26(6):880–888, 1999.
- [26] Eva M Van Rikxoort, Bram Van Ginneken, Mark Klik, and Mathias Prokop. Supervised enhancement filters: Application to fissure detection in chest ct scans. *IEEE transactions on medical imaging*, 27(1):1–10, 2007.
- [27] Kenji Suzuki. Machine learning in computer-aided diagnosis of the thorax and colon in ct: a survey. *IEICE transactions on information and systems*, 96(4):772–783, 2013.
- [28] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale convolutional neural networks for lung nodule classification. In *International conference on information processing in medical imaging*, pages 588–599. Springer, 2015.
- [29] Francesco Ciompi, Kaman Chung, Sarah J Van Riel, Arnaud Arindra Adiyoso Setio, Paul K Gerke, Colin Jacobs, Ernst Th Scholten, Cornelia Schaefer-Prokop, Mathilde MW Wille, Alfonso Marchiano, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific reports*, 7(1):1–11, 2017.
- [30] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

- [31] Z Huo, ML Giger, CJ Vyborny, and CE Metz. Effectiveness of cad in the diagnosis of breast cancer: An observer study on an independent database of mammograms. *Radiology*, 224:560–568, 2002.
- [32] Lubomir Hadjiiski, Heang-Ping Chan, Berkman Sahiner, Mark A Helvie, Marilyn A Roubidoux, Caroline Blane, Chintana Paramagul, Nicholas Petrick, Janet Bailey, Katherine Klein, et al. Improvement in radiologists characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: An roc study 1. *Radiology*, 233(1):255–265, 2004.
- [33] Shingo Kakeda, Junji Moriya, Hiromi Sato, Takatoshi Aoki, Hideyuki Watanabe, Hajime Nakata, Nobuhiro Oda, Shigehiko Katsuragawa, Keiji Yamamoto, and Kunio Doi. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. *American Journal of Roentgenology*, 182(2):505–510, 2004.
- [34] Kazuo Awai, Kohei Murao, Akio Ozawa, Masanori Komi, Haruo Hayakawa, Shinichi Hori, and Yasumasa Nishimura. Pulmonary nodules at chest ct: Effect of computer-aided diagnosis on radiologists detection performance 1. *Radiology*, 230(2):347–352, 2004.
- [35] Bruce I Reiner. The challenges, opportunities, and imperative of structured reporting in medical imaging. *Journal of digital imaging*, 22(6):562–568, 2009.
- [36] Daniel K Powell and James E Silberzweig. State of structured reporting in radiology, a survey. *Academic radiology*, 22(2):226–233, 2015.
- [37] European Society of Radiology (ESR) communications@ myesr. org. Esr paper on structured reporting in radiology. *Insights into imaging*, 9:1–7, 2018.
- [38] Jan Bosmans, Lieve Peremans, Maurizio Menni, A Schepper, philippe duyck, and Paul Parizel. Structured reporting: If, why, when, how-and at what expense? results of a focus group meeting of radiology professionals from eight countries. *Insights into imaging*, 3: 295–302, 06 2012. doi: 10.1007/s13244-012-0148-1.
- [39] ROGERS FB. Medical subject headings. *Bulletin of the Medical Library Association*, 51: 114–116, 1963.

- [40] Curtis P Langlotz. Radlex: a new method for indexing online educational materials, 2006.
- [41] Medline indexing online training. <http://www.nlm.nih.gov/bsd/indexing/training/INT030.html> 2015 – 01 – 11.
- [42] Open-i: Open-edit radiology resource, compiled by radiologists and other health professionals from across the globe. URL <https://radiopaedia.org/>.
- [43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- [44] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [45] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
- [46] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.
- [47] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [48] Marvin Minsky and Seymour Papert. Perceptrons. 1969.
- [49] Paul Werbos. Beyond regression: New tools for prediction and analysis in the behavioral sciences. 1974.
- [50] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.

- [51] Alex Krizhevsky, Ilya Sutskever, and Hinton Geoffrey E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS2012)*, pages 1–9, 2012. ISSN 10495258. doi: 10.1109/5.726791. URL <https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-n>
- [52] B Boser Le Cun, JS Denker, D Henderson, RE Howard, W Hubbard, and LD Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, 1990.
- [53] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [54] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [55] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
- [56] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.
- [57] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [58] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [59] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.
- [60] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to



- attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.
- [61] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.
- [62] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.
- [63] The unreasonable effectiveness of recurrent neural networks. <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>. Accessed: 2021-10-25.
- [64] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *ICML*, 2011.
- [65] Ross Turner, Somayajulu Sripada, Ehud Reiter, and Ian P Davy. Generating spatio-temporal descriptions in pollen forecasts. In *Demonstrations*, 2006.
- [66] Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169, 2005.
- [67] Mary Dee Harris. Building a large-scale commercial nlg system for an emr. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 157–160, 2008.
- [68] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.
- [69] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

- [70] Xinlei Chen and C Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014.
- [71] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.
- [72] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.
- [73] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [74] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [75] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
- [76] NIH. Unified medical language system, 2011. URL <https://www.nlm.nih.gov/research/umls/quickstart.html>.
- [77] L Smith, Thomas Rindflesch, W John Wilbur, et al. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321, 2004.
- [78] Susanne M Humphrey, Willie J Rogers, Halil Kilicoglu, Dina Demner-Fushman, and Thomas C Rindflesch. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the Association for Information Science and Technology*, 57(1):96–113, 2006.
- [79] Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

- [80] Krishna Prasad Chodey and Gongzhu Hu. Clinical Text Analysis using Machine Learning Methods. *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*, 2016. doi: 10.1109/ICIS.2016.7550908.
- [81] Annotation of chest radiology reports for indexing and retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9059:99–111, 2015. ISSN 16113349.
- [82] Patricia Flatley Brennan and Alan R Aronson. Towards linking patients and clinical information: detecting umls concepts in e-mail. *Journal of biomedical informatics*, 36(4): 334–341, 2003.
- [83] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Dont count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, 2014.
- [84] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.
- [85] Dana H Ballard. Modular learning in neural networks. In *AAAI*, volume 647, pages 279–284, 1987.
- [86] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294, 1988.
- [87] Hoo-Chang Shin, Matthew R Orton, David J Collins, Simon J Doran, and Martin O Leach. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1930–1943, 2012.
- [88] Yulian Zhu, Li Wang, Mingxia Liu, Chunjun Qian, Ambereen Yousuf, Aytakin Oto, and Dinggang Shen. Mri-based prostate cancer detection with high-level representation and hierarchical classification. *Medical physics*, 44(3):1028–1039, 2017.

- [89] Le Hou, Vu Nguyen, Ariel B Kanevsky, Dimitris Samaras, Tahsin M Kurc, Tianhao Zhao, Rajarsi R Gupta, Yi Gao, Wenjin Chen, David Foran, et al. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. *Pattern recognition*, 86:188–200, 2019.
- [90] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)*, pages 241–246. IEEE, 2016.
- [91] Gavin Seegoolam, Jo Schlemper, Chen Qin, Anthony Price, Jo Hajnal, and Daniel Rueckert. Exploiting motion for deep learning reconstruction of extremely-undersampled dynamic mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 704–712. Springer, 2019.
- [92] Chen Qin, Jo Schlemper, Jose Caballero, Anthony N Price, Joseph V Hajnal, and Daniel Rueckert. Convolutional recurrent neural networks for dynamic mr image reconstruction. *IEEE transactions on medical imaging*, 38(1):280–290, 2018.
- [93] Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger. Sense: sensitivity encoding for fast mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(5):952–962, 1999.
- [94] Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (grappa). *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002.
- [95] Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, 2018.
- [96] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

- [97] Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2017.
- [98] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.
- [99] Dahua Lin, Chen Kong, Sanja Fidler, and Raquel Urtasun. Generating multi-sentence lingual descriptions of indoor scenes. *arXiv preprint arXiv:1503.00064*, 2015.
- [100] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [101] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *International conference on machine learning*, pages 595–603, 2014.
- [102] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [103] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010.
- [104] Tomas Mikolov and Geoffrey Zweig. Context dependent recurrent neural network language model. *SLT*, 12:234–239, 2012.
- [105] Recurrent Models of Visual Attention. *Nips-2014*, pages 1–9, 2014. ISSN 0157244X. doi: ng. URL <http://arxiv.org/abs/1406.6247>.
- [106] Ke Yan, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, and Ronald M Summers. Holistic and comprehensive annotation of clinically significant findings on diverse ct images: Learning from radiology reports and label ontology. In *Proceedings*

- of the *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8523–8532, 2019.
- [107] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- [108] Robert Leaman, Ritu Khare, and Zhiyong Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57: 28–37, 2015.
- [109] Thomas Schlegl, Sebastian M. Waldstein, Wolf Dieter Vogl, Ursula Schmidt-Erfurth, and Georg Langs. Predicting semantic descriptions from medical images with convolutional neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9123:437–448, 2015. ISSN 16113349.
- [110] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers. Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation. *Journal of Machine Learning Research*, 17(1-31):2, 2016.
- [111] Xiaosong Wang, Le Lu, Hoo-Chang Shin, Lauren Kim, Mohammadhadi Bagheri, Isabella Nogues, Jianhua Yao, and Ronald M Summers. Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 998–1007. IEEE, 2017.
- [112] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6428–6436, 2017.

- [113] Zizhao Zhang, Pingjun Chen, Manish Sapkota, and Lin Yang. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 320–328. Springer, 2017.
- [114] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.
- [115] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer, 2019.
- [116] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, 2018.
- [117] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [118] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.
- [119] Marc Tanti, Albert Gatt, and Kenneth P Camilleri. What is the role of recurrent neural networks (rnns) in an image caption generator? *arXiv preprint arXiv:1708.02043*, 2017.
- [120] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [121] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

- [122] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- [123] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*, 2019.
- [124] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*, 2019.
- [125] William Hersh, Mark Maillhot, Catherine Arnott-Smith, and Henry Lowe. Selective automated indexing of findings and diagnoses in radiology reports. *Journal of biomedical informatics*, 34(4):262–273, 2001.
- [126] Burke W Mamlin, Daniel T Heinze, and Clement J McDonald. Automated extraction and normalization of findings from cancer-related free-text radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2003, page 420. American Medical Informatics Association, 2003.
- [127] Axel Gerstmaier, Philipp Daumke, Kai Simon, Mathias Langer, and Elmar Kotter. Intelligent image retrieval based on radiology reports. *European radiology*, 22(12):2750–2758, 2012.
- [128] Marcelo Fiszman, Wendy W Chapman, Dominik Aronsky, R Scott Evans, and Peter J Haug. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *Journal of the American Medical Informatics Association*, 7(6):593–604, 2000.
- [129] Merlijn Sevenster, Rob Van Ommering, and Yuechen Qian. Automatically correlating clinical findings and body locations in radiology reports using medlee. *Journal of digital imaging*, 25(2):240–249, 2012.
- [130] Bao H Do, Andrew S Wu, Joan Maley, and Sandip Biswal. Automatic retrieval of bone



- fracture knowledge using natural language processing. *Journal of digital imaging*, 26(4):709–713, 2013.
- [131] Richárd Farkas and György Szarvas. Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, 9(3):S10, 2008.
- [132] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [133] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.
- [134] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [135] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.
- [136] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018.
- [137] Allen Institute for Brain Science allen human brain ontology. <http://human.brain-map.org/ontology.html>. Accessed: 2018-04-09.
- [138] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [139] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [140] Robert E Schapire and Yoram Singer. Boostexter: A boosting-based system for text categorization. *Machine learning*, 39(2):135–168, 2000.
- [141] Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of*

- the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, 2016.
- [142] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. Neural document summarization by jointly learning to score and select sentences. *arXiv preprint arXiv:1807.02305*, 2018.
- [143] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
- [144] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.
- [145] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*, 2016.
- [146] Jianpeng Cheng and Mirella Lapata. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*, 2016.
- [147] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [148] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.
- [149] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

- [150] Jo Schlemper, Ozan Oktay, Wenjia Bai, Daniel C Castro, Jinming Duan, Chen Qin, Jo V Hajnal, and Daniel Rueckert. Cardiac mr segmentation from undersampled k-space using deep latent representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 259–267. Springer, 2018.
- [151] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–466. Springer, 2018.
- [152] Andreas Merrem, Sabine Hofer, Dirk Voit, K-Dietmar Merboldt, Jakob Klosowski, Markus Untenberger, Julius Fleischhammer, Jens Frahm, et al. Rapid diffusion-weighted magnetic resonance imaging of the brain without susceptibility artifacts: Single-shot steam with radial undersampling and iterative reconstruction. *Investigative radiology*, 52(7):428–433, 2017.
- [153] Wenchuan Wu and Karla L Miller. Image formation in diffusion mri: a review of recent technical developments. *Journal of Magnetic Resonance Imaging*, 46(3):646–662, 2017.
- [154] Jakob Weiss, Petros Martirosian, Jana Taron, Ahmed E Othman, Thomas Kuestner, Michael Erb, Jens Bedke, Fabian Bamberg, Konstantin Nikolaou, and Mike Notohamiprodjo. Feasibility of accelerated simultaneous multislice diffusion-weighted mri of the prostate. *Journal of Magnetic Resonance Imaging*, 46(5):1507–1515, 2017.
- [155] Alexander Ciritsis, Cristina Rossi, Magda Marcon, Valerie Doan Phi Van, and Andreas Boss. Accelerated diffusion-weighted imaging for lymph node assessment in the pelvis applying simultaneous multislice acquisition: a healthy volunteer study. *Medicine*, 97(32), 2018.
- [156] Aydan Gasimova. Automated enriched medical concept generation for chest x-ray images. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 83–92. Springer, 2019.

- [157] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157, 2003.
- [158] Ehud Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401, 2018.
- [159] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.
- [160] Elior Sulem, Omri Abend, and Ari Rappoport. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*, 2018.
- [161] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [162] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [163] Roser Morante and Walter Daelemans. A metalearning approach to processing the scope of negation. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 21–29, 2009.
- [164] Noa P Cruz Díaz, Manuel J Mana López, Jacinto Mata Vázquez, and Victoria Pachón Álvarez. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the American society for information science and technology*, 63(7):1398–1410, 2012.
- [165] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

- [166] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.
- [167] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [168] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13, 2021.

## .1 Patterns and diagnoses groupings of MeSH terms

## .2 MetaMap extracted terms

Table 1: Chest X-ray patterns and diagnoses manually extracted from MeSH annotations of the IU-CX images [1]. Definition of patterns and diagnosis taken from Smithuis and van Delden [4].

(a)	
Patterns	
mass	
degenerative bone	
calcinosis	
opacity	
mastectomy	
interstitial	
tortuous	
catheters indwelling	
cardiomegaly	
foreign bodies	
consolidation	
atelectasis	
hypoinflation	
bronchial thickening	
dislocations	
congestion	
scoliosis	
osteophyte	
hernia	
hyperinflation	
hyperdistention	
atherosclerosis	
lucency	
deformity	
fractures bone	
kyphosis	
spinal fusion	
diaphragmatic eventration	
lung hyperlucent	
pneumoperitoneum	
colonic interposition	
epicardial fat	
diaphragm elevation	
pneumonectomy	
nipple shadow	
adipose tissue	
bullae	
contrast media	
funnel chest	
hypovolemia	
bronchiectasis	
aorta dilated	
pectus carinatum	
	(b)
	Diagnoses
	normal
	airspace disease
	pleural effusion
	bone diseases metabolic
	calcified granuloma
	cicatrix
	pulmonary congestion
	pneumothorax
	granulomatous disease
	pulmonary emphysema
	spondylosis
	emphysema
	pulmonary disease chronic obstructive
	pulmonary edema
	lung diseases interstitial
	pneumonia
	arthritis
	cysts
	hydropneumothorax
	sclerosis
	pulmonary fibrosis
	granuloma
	hyperostosis diffuse idiopathic skeletal
	volume loss
	bronchitis
	cystic fibrosis
	heart atria
	aortic aneurysm
	bullous emphysema
	hemopneumothorax
	hemothorax
	osteoporosis

Table 2: Metamap extracted ‘Disease or Syndrome’, ‘Finding’ and ‘Pathologic Function’ terms that appear in at least 30 reports, and their degree of overlap with other terms.

Disease/Finding/Pathology Term	Total appearances	Appearances w/ and within other terms	Overlap %
clear	3350	2651	79
normal	2512	2848	113
heart size	1812	2434	134
intact	711	664	93
normal heart size	703	506	72
atelectasis	630	696	110
cardiomegaly	532	494	93
opacities	479	477	100
heart size normal	397	359	90
unchanged	342	319	93
thoracic spine degeneration	336	306	91
opacity	327	387	118
disease	319	678	212
degenerative spine	259	230	89
pleural effusion	244	343	140
crowding	165	159	96
normal breast	163	159	98
emphysema	157	169	107
tortuous aorta	157	153	97
tortuous	157	288	183
granulomatous disease	154	170	110
vascular	150	222	148
limited	120	114	95
identified	116	114	98
pneumothorax	107	146	136
thoracic spondylosis	105	101	96
pneumonia	104	189	181
negative	98	86	88
pleural effusions bilateral	93	89	96
infection nos	84	80	95
followup	83	89	107
copd	72	74	102
probable	72	72	100
nodular opacity	71	69	97
congestion	68	71	104
calcifications	65	67	103
consolidation	64	80	125
hiatal hernia	63	61	97
arthritic changes	59	57	97
osteophytes	58	73	126
pleural thickening	57	56	98
scar	56	65	116
pulmonary oedema	56	56	100
right chest	47	47	100
sequela	46	69	150
nodular density	45	43	96
atelectasis focal	43	43	100
thickening	42	94	224