Right for the Right Reason: Training Agnostic Networks

Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini

Intelligent Systems Laboratory, University of Bristol, Bristol BS8 1UB, UK,
{sen.jia, thomas.lansdall-welfare, nello.cristianini}@bris.ac.uk

Abstract. We consider the problem of a neural network being requested to classify images (or other inputs) without making implicit use of a "protected concept", that is a concept that should not play any role in the decision of the network. Typically these concepts include information such as gender or race, or other contextual information such as image backgrounds that might be implicitly reflected in unknown correlations with other variables, making it insufficient to simply remove them from the input features. In other words, making accurate predictions is not good enough if those predictions rely on information that should not be used: predictive performance is not the only important metric for learning systems. We apply a method developed in the context of domain adaptation to address this problem of "being right for the right reason", where we request a classifier to make a decision in a way that is entirely 'agnostic' to a given protected concept (*e.g.* gender, race, background etc.), even if this could be implicitly reflected in other attributes via unknown correlations. After defining the concept of an 'agnostic model', we demonstrate how the Domain-Adversarial Neural Network can remove unwanted information from a model using a gradient reversal layer.

Keywords: Agnostic models, Explainable AI, Fairness in AI, Trust

1 Introduction

Data-driven Artificial Intelligence (AI) is behind the new generation of success stories in the field, and is predicated not just on a few technological breakthroughs, but on a cultural shift amongst its practitioners: namely the belief that predictions are more important than explanations, and that correlations count more than causations [4, 8]. Powerful black-box algorithms have been developed to sift through data and detect any possible correlation between inputs and intended outputs, exploiting anything that can increase predictive performance. Computer vision (CV) is one of the fields that has benefited the most from this choice, and therefore can serve as a test bed for more general ideas in AI.

This paper targets the important problem of ensuring trust in AI systems. Consider a case as simple as object classification. It is true that exploiting contextual clues can be beneficial in CV and generally in AI tasks. After all, if an

algorithm thinks it is seeing an elephant (the object) in a telephone box (the context), or Mickey Mouse driving a Ferrari, it is probably wrong. This illustrates that even though your classifier might have an opinion about the objects in an image, the context around it can be used to improve your performance (*e.g.* telling you that it is unlikely to be an elephant inside a telephone box), as shown in many recent works [3, 13, 14].

However, making predictions based on context can also lead to problems and creates various concerns, one of which is the use of classifiers in "out of domain" situations, a problem that leads to research questions in domain adaptation [6, 18]. Other concerns are also created around issues of bias, *e.g.* classifiers incorporating biases that are present in the data and are not intended to be used [2], which run the risk of reinforcing or amplifying cultural (and other) biases [20]. Therefore, both predictive accuracy and fairness are heavily influenced by the choices made when developing black-box machine-learning models.

Since the limiting factor in training models is often sourcing labelled data, a common choice is to resort to reusing existing data for a new purpose, such as using web queries to generate training data, and employing various strategies to annotate labels, *i.e.* using proxy signals that are expected to be somewhat correlated to the intended target concept [5, 11]. These methods come with no guarantees of being unbiased, or even to reflect the deployment conditions necessarily, with any data collected "in the wild" [8, 10] carrying with it the biases that come from the wild.

To address these issues, a shift in thinking is needed, from the aforementioned belief that predictions are more important than explanations, to ideally developing models that make predictions that are right for the right reason, and consider other metrics, such as fairness, transparency and trustworthiness, as equally important as predictive performance. This means that we want to ensure that certain protected concepts are not used as part of making critical decisions (*e.g.* decisions about jobs should not be based on gender or race) for example, or that similarly, predictions about objects in an image should not be based on contextual information (gender of a subject in an image should not be based on the background).

In this direction, we demonstrate how the Domain-Adversarial Neural Network (DANN) developed in the context of domain adaptation [6] can be modified to generate 'agnostic' feature representations that do not incorporate any implicit contextual (correlated) information that we do not want, and is therefore unbiased and fair. We note that this is a far stronger requirement than simply removing protected features from the input that might otherwise implicitly remain in the model due to unforeseen correlations with other features.

We present a series of experiments, showing how the relevant pixels used to make a decision move from the contextual information to the relevant parts of the image. This addresses the problem of relying on contextual information, exemplified by the Husky/Wolf problem in [15], but more importantly shows a way to de-bias classifiers in the feature engineering step, allowing it to be applied

generally for different models, whether that is word embeddings, support vector machines, or deep networks etc.

Ultimately, this ties into the current debate about how to build trust in these tools, whether this is about their predictive performance, their being right for the right reason, their being fair, or their decisions being explainable.

2 Agnostic Models

Methods have previously been proposed to remove biases, based on various principles, one of which is distribution matching [20]: ensuring that the ratio between protected attributes is the same in the training instances and in the testing instances. However, this does not avoid using the wrong reasons in assessing an input but simply enforces a post-hoc rescaling of scores, to ensure that the outcome matches the desired statistical requirements of fairness.

In our case, we do not want to have an output distribution that only looks as if it has been done without using protected concepts. We actually want a model that cannot even represent them within its internal representations, where we call such a model *agnostic*. This is a model that does not represent a protected concept internally, and therefore cannot use it even indirectly. Of course this kind of constraint is likely to lead to lower accuracy. However, we should keep in mind that this reduction in accuracy is a direct result of no longer using contextual clues and correlations that we explicitly wish to prevent.

In this direction, we consider classification tasks where X is the input space and $Y = \{0, 1, \ldots, L-1\}$ is the set of L possible labels. An agnostic model (or feature representation) $G_f : X \to \mathbb{R}^D$, parameterized by θ_f, maps a data example $(\mathbf{x}_i, \mathbf{y}_i)$ into a new D-dimensional feature representation $\mathbf{z} \in \mathbb{R}^D$ such that for a given label $p \in Y$, there does not exist an algorithm $G_y : \mathbb{R}^D \to [0,1]^L$ which can predict p with better than random performance.

3 Domain-Adversarial Neural Networks

One possible way to learn an agnostic model is to use a DANN [6], recently proposed for domain adaptation, which explicitly implements the idea raised in [1] of learning a representation that is unable to distinguish between training and test domains. In our case, we wish for the model to be able to learn a representation that is agnostic to a protected concept.

DANNs are a type of Convolutional Neural Network (CNN) that can achieve an agnostic representation using three components. A feature extractor $G_f(\cdot;\theta_f)$, a label prediction output layer $G_y(\cdot;\theta_y)$ and an additional protected concept prediction layer $G_p : \mathbb{R}^D \to [0,1]$, parameterized by θ_p. During training, two different losses are then computed: a target prediction loss for the i-th data instance $\mathcal{L}_y^i(\theta_f, \theta_y) = \mathcal{L}_y(G_y(G_f(\mathbf{x}_i;\theta_f);\theta_y), \mathbf{y}_i)$ and a protected concept loss $\mathcal{L}_p^i(\theta_f, \theta_p) = \mathcal{L}_p(G_p(G_f(\mathbf{x}_i;\theta_f);\theta_p), p_i)$, where $\mathcal{L}_y$ and $\mathcal{L}_p$ are both given by the cross-entropy loss and p_i is the label denoting the protected concept we wish to be unable to distinguish using the learnt representation.

Training the network then attempts to optimise

$$E(\theta_f, \theta_y, \theta_p) = (1-\alpha)\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_y^i(\theta_f, \theta_y) - \alpha\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_p^i(\theta_f, \theta_p) + \frac{1}{n'}\sum_{i=n+1}^{N}\mathcal{L}_p^i(\theta_f, \theta_p)\right), \tag{1}$$

where $n' = N - n$ and α is the hyper-parameter for the trade-off between the two losses, finding the saddle point $\hat{\theta}_f, \hat{\theta}_y, \hat{\theta}_p$ such that

$$(\hat{\theta}_f, \hat{\theta}_y) = \underset{\theta_f, \theta_y}{\operatorname{argmin}} E(\theta_f, \theta_y, \hat{\theta}_p), \tag{2}$$

$$\hat{\theta}_p = \underset{\theta_p}{\operatorname{argmax}} E(\hat{\theta}_f, \hat{\theta}_y, \theta_p). \tag{3}$$

As further detailed in [6], introducing a *gradient reversal layer* (GRL) between the feature extractor G_f and the protected concept classifier G_p allows (1) to be framed as a standard stochastic gradient descent (SGD) procedure as commonly implemented in most deep learning libraries.

The network can therefore be learnt using a simple stochastic gradient procedure, where updates to θ_f are made in the opposite direction of the gradient for the maximizing parameters, and in the direction of the gradient for the minimizing parameters. Stochastic estimates of the gradient are made, both for the target concept and for the protected concept, using the training set. We can see this as the two parts of the neural network (target classifier G_y and protected concept classifier G_p) are competing with each other for the control of the internal representation. DANN will attempt to learn a model G_f that maps an example into a representation allowing the target classifier to accurately classify instances, but crippling the ability of the protected concept classifier to discriminate inputs by their label for the protected concept.

4 Experiments

To test the use of DANNs for learning representations that can be used to make predictions for the right reasons, we ran two different experiments. In Experiment 1, we first demonstrate the issue of using contextual information to make predictions in a cross-domain classification task, before using a DANN in Experiment 2, showing that the network can learn an agnostic representation that allows us to make predictions on a target concept without using information from a correlated contextual concept (the protected concept in this case), such as the image background.

4.1 Data Description

In this work, we combine two datasets, making use of the 'Jaguar' and 'Killer whale' categories from the ImageNet dataset [16], as well as the 'Forest path' and 'Coast' categories from the Places dataset [21].

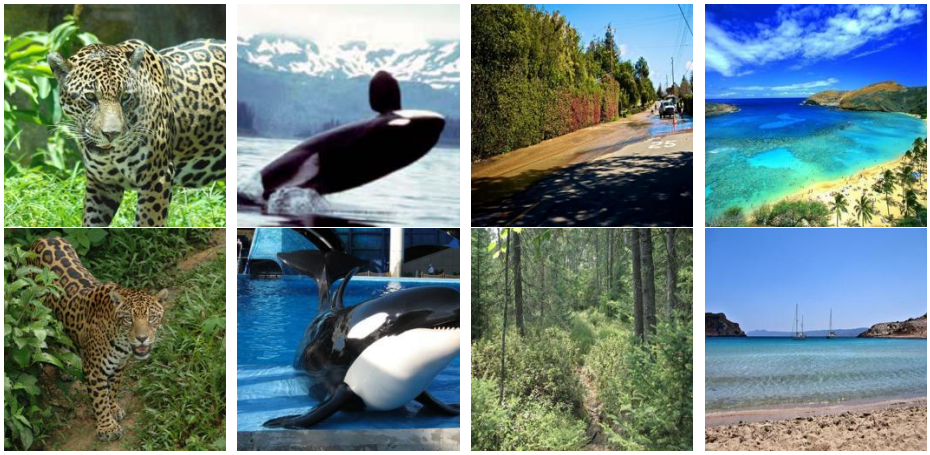


Fig. 1. Example images taken from the 'Jaguar', 'Killer whale', 'Forest path' and 'Coast' categories of the ImageNet and Places datasets respectively (left-right).

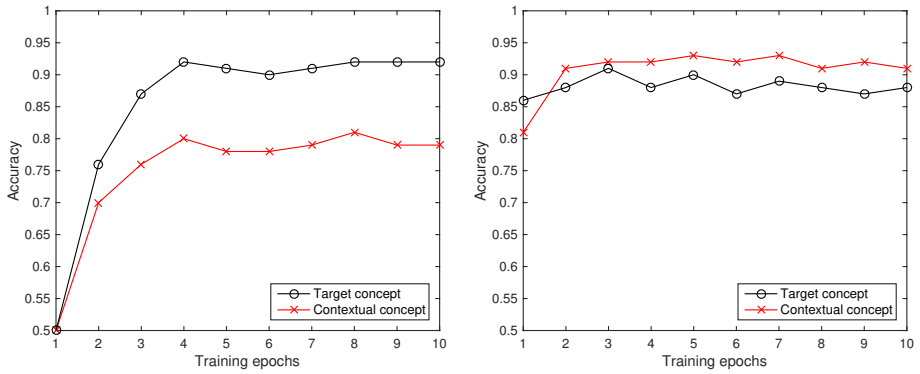
A two-part training set was constructed containing 2,524 images from the 'Jaguar' category, and the same number for the 'Killer whale' category from ImageNet (the target concept training set). This was further supplemented with 5,000 images from each of the two categories ('Forest path' and 'Coast') from the Places dataset (the contextual concept training set), for a total of 15,048 images in the combined training set. Two separate hold-out sets were also created, one for the target concept containing 50 hold-out images from each of the 'Jaguar' and 'Killer whale' categories, and one for the contextual concept containing 50 hold-out images from each of the 'Forest path' and 'Coast' categories.

Data augmentation was performed on the training set to increase the number of instances by creating new images that are multi-crops of 224×224 pixels and horizontally flipping copies of the training set images. All images in our experiments were also pre-processed to be 256×256 pixels by a process of multi-cropping where each image is resized before cropping the final size from the centre region, as in [9, 12]. Example images from the training set used for the experiments can be seen in Fig. 1.

4.2 Network structure

The network structure used for our experiments in this paper are based upon a simplified version of the VGG-net CNN used in [17], where the feature extraction layers G_f consist of five convolutional layers: conv3-64[1], conv3-128, conv3-256, conv3-512 and conv3-512, with ReLU activation and max-pooling layers inserted after each convolutional layer. The output prediction classifiers G_y and G_p are each composed of four fully connected layers, fc-1024, with ReLU and dropout layers with a dropout of 0.5 after each fully connected layer.

[1] conv$a$-b denotes a convolutional layer consisting of b filters of size $a \times a$.



(a) CNN trained on the target concept training set (animals)

(b) CNN trained on the contextual concept training set (backgrounds)

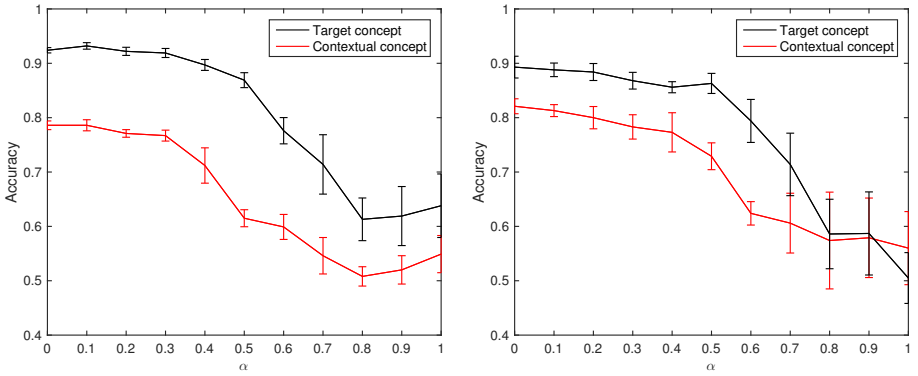
Fig. 2. Results from Experiment 1, showing that a standard CNN model trained on the target concept will also learn how to classify in the contextual concept and vice versa.

4.3 Experiment 1: Cross-domain classification

In this first experiment, we motivate our approach by demonstrating the problem we wish to address, namely that contextual information can be used to make classification decisions about our target concept that is not related to the target that we actually wish to learn.

We began by training from scratch two independent CNNs with the same network architecture, one on the target concept training set and one on contextual concept training set. The layers of the network are described in Sec 4.2 with a single output prediction classifer G_y per model, *i.e.* each CNN is composed of five convolution layers, followed by four fully connected layers with no shared features across the models. Each model was trained for 10 epochs using the following model parameters: a batch size of 32, a starting learning rate of $\eta = 0.01$ that decays every three epochs by a factor of 10, a momentum of 0.5 and a weight decay of $5e^{-4}$.

The accuracy of each model was measured on both the target and contextual test sets after each epoch as shown in Fig. 2. As one might expect, we can see that the model trained on the target concept achieves an accuracy of 92% on the target test set, while the contextual concept model achieves an accuracy of 91% on the contextual test set. More problematically, we can see that the target concept model, trained only on images of animals, also has good performance at classifying images of forest paths and coastlines from the contextual test set, with an accuracy of 79%. Similarly, the contextual concept model, trained only on images of forest paths and coastlines can correctly identify animals with an accuracy of 88%.



(a) Performance of G_y in the DANN trained for the target concept (animals)

(b) Performance of G_p in the DANN trained for the contextual concept (backgrounds)

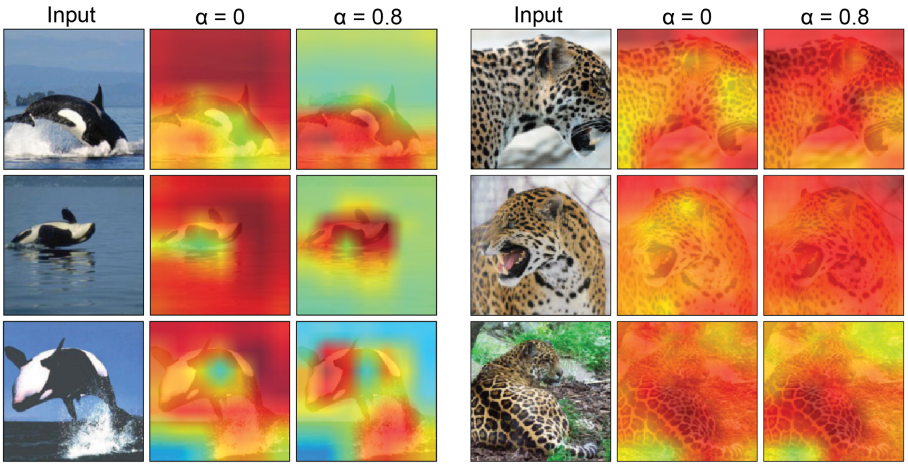
Fig. 3. Accuracy of the two independent classifiers in the DANN using the shared feature space on the test sets for different values of α.

4.4 Experiment 2: Learning with Domain-adversarial neural networks

In this next experiment, we show that with our proposed use of DANNs maximises its performance on the target concept whilst following the constraint that it should not learn useful features for the protected contextual concept. We further examine the most informative pixels (*e.g.* those pixels which have the strongest response in the feature map) used for classification [7, 19], showing that the most informative pixels are no longer found in the image background.

Keeping all the model parameters, apart from a new learning rate ($\eta = 0.001$), the same as in Experiment 1, we trained a single DANN model on the combined training set, with the network layers outlined in Sec 4.2, with the target prediction output layers G_y predicting the target concept, and the protected concept prediction layers predicting the contextual concept. By doing so, we force the model to learn a shared data representation (feature space) that maximises performance on the target while incorporating no knowledge of features which are useful for classifying the contextual concept images. This process was repeated for different gradient trade-offs in the range $\alpha = [0, 0.1, \ldots, 1]$ using a grid-search procedure, where $\alpha = 0$ represents simply training the shared feature space on the target concept, and $\alpha = 1$ represents training the shared feature space to maximise the loss for the contextual concept. We repeated this process 10 times, reporting the average accuracy for each run, along with the standard deviation.

In Fig. 3, we can see the accuracy of the DANN for varying gradient trade-off values on the target and contextual concept test sets. Our results show that as α increases and is further constrained in its use of information from the contextual concept, the performance on the target concept decreases, suggesting that the performance on the intended target concept was indeed being helped



(a) Activation maps for the 'Killer whale' category

(b) Activation maps for the 'Jaguar' category

Fig. 4. Activation maps based on the strongest response of the shared feature representation. Examples selected are those with the least correlation between the activation maps for $\alpha = 0$ and $\alpha = 0.8$ as shown in the images.

by the contextual background information. Our results show that once we have removed features which are useful for predicting the contextual concept, our target classifier achieves an accuracy of 64%, while the contextual classifier can only maintain an accuracy close to random guessing.

We further investigated whether after applying the minimax procedure of the DANN that the most informative pixels for prediction corresponded with the location of the target concept in the image, or whether they were focused on the background scene of the image. Fig. 4 shows activation maps for the feature representation shared between the independent classifiers on a set of three images for each target concept category. Examples were selected as those with the least correlation between the activation maps for the contrasting α values of 0 and 0.8 shown, where α values were chosen as the two extremes in the classification accuracy.

We can observe that for the 'Killer whale' category, the most informative pixels for $\alpha = 0$ are indeed found in the background of the image, while for $\alpha = 0.8$ the activation maps show that the network is focusing on the actual body of the animal instead, as desired. For the 'Jaguar' category, analysis of the most informative pixels is less clear, with activation generally being spread widely across the image. However, we do see some evidence of a stronger activation response to parts of the jaguar's body overall.

5 Discussion

In our experiments, we found that as the model learns the shared features with increasingly less contextual information, accuracy of the target classifier decreases. This is exactly what we expect and directly addresses our main argument, that previously the classifier was relying on the protected contextual background information that should not be used to make its predictions.

At one extreme, where $\alpha = 1$, the network is using no information from the target concept in its data representation, instead trying to maximise its loss on the protected concept in the shared feature space G_f, while minimising its loss in the protected classifier G_p. This tension between the two parts of the network leads to a minimax scenario where if there is any information which can be exploited to correctly predict in the protected concept, it is subsequently removed from the data representation.

We note that ideally α should be set to 1 for similar experiments, given that for any other setting the learning system would still be exploiting forbidden information from the protected concept, and would not be satisfying the original requirements of the task: to learn to predict without the contextual information. However, since in this scenario the shared feature space would not rely on the target domain at all, α needs to be slowly increased as training progresses until reaching its maximum. In this way, the features will be guided by the target domain as well, forming a saddle point in the exploration of the feature space as required.

Results from investigating the most informative pixels for classification at differing levels of α revealed that the constraint of the contextual concept appears to have been more successful for the 'Killer whale' and 'Coast' images than for the 'Jaguar' and 'Forest path' pairing. This can perhaps be best explained by how closely the contextual concept training images represent the contextual concept found in the target concept training images, *i.e.* the whales are always pictured next to or in the ocean, whereas jaguars will sometimes be found outside of the jungle with different backgrounds, and therefore the 'Forest path' category does not match 'Jaguar' backgrounds as closely as 'Coast' does for the 'Killer whale' category.

Further theoretical and experimental analysis of additional minimax architectures is needed to explain the phenomena of the target classifier accuracy increasing on both target and contextual test sets for values of $\alpha \geq 0.8$.

6 Conclusions

The creation of a new generation of AI systems that can be trusted to make fair and unbiased decisions is an urgent task for researchers. As AI rapidly conquers technical challenges related to predictive performance, we are discovering a new dimension to the design of such systems that must be addressed: the fairness and trust in the system's decisions.

In this paper, we address this critical issue of trust in AI by not only proposing a new high standard for models to meet, being agnostic to a protected concept,

but also proposing a method to achieve such models. We define a model to be agnostic with respect to a set of concepts if we can show that it makes its decisions without ever using these concepts. This is a much stronger requirement than in distributional matching or other definitions of fairness. We focus on the case where a small set of protected concepts should not be used in decisions, and can be exemplified by samples of data. We have demonstrated how ideas developed in the context of domain adaptation can deliver agnostic representations that are important to ensure fairness and therefore trust.

Our experiments demonstrate that the DANN successfully removes unwanted contextual information, and makes decisions for the right reasons. While demonstrated here by ignoring the physical background context of an object in an image, the same approach can be used to ensure that other protected information does not make its way into black-box classifiers deployed to make decisions about people in other domains and classification tasks.

Acknowledgements

SJ, TLW and NC are support by the FP7 Ideas: European Research Council Grant 339365 - ThinkBIG.

References

1. Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.
2. Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
3. Wenqing Chu and Deng Cai. Deep feature based contextual model for object detection. *Neurocomputing*, 275:1035–1042, 2018.
4. Nello Cristianini. On the current paradigm in artificial intelligence. *AI Communications*, 27(1):37–43, 2014.
5. 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.
6. Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
7. Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.
8. Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.
9. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

10. Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.
11. Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini. Gender classification by deep learning on millions of weakly labelled images. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 462–467. IEEE, 2016.
12. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
13. Jianan Li, Yunchao Wei, Xiaodan Liang, Jian Dong, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Attentive contexts for object detection. *IEEE Transactions on Multimedia*, 19(5):944–954, 2017.
14. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
15. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.
16. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
17. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Eprint Arxiv*, 2014.
18. Markus Wulfmeier, Alex Bewley, and Ingmar Posner. Addressing appearance change in outdoor robotics with adversarial domain adaptation. *arXiv preprint arXiv:1703.01461*, 2017.
19. Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.
20. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.
21. Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.