Fairness in Machine Learning, Project Proposal

Andrew Choi
205348339
asjchoi@ucla.edu

Zeyu Zhang
505030513
zeyuzhang@ucla.edu

Steven Gong
804846708
nikepupu@ucla.edu

Gaohong Liu
705352121
cheimu@ucla.edu

Abstract

As machine learning systems continue to vastly integrate themselves within society, it is important that decisions produced by such systems are fair. Such requirements are even more crucial when dealing with decisions that have lasting social, legal, and/or financial implications towards individuals. Unfortunately, it has been shown that bias is often encoded into machine learning systems due to preexisting dataset bias. Additionally, often times, it has been shown that many learning algorithms will produce predictors that amplify this bias effectively worsening the problem. Two datasets which have been found to contain a large amount of gender bias are the imSitu vSRL and MS-COCO datasets. In this work, we propose a framework for reducing such bias amplification by incorporating a set of auxiliary tasks to aid the learning algorithm. We will then compare the results to a naive approach as well as preexisting techniques to show that the bias reduction is comparable for certain combinations of auxiliary tasks.

1 Introduction

In recent years, deep learning systems have demonstrated tremendous potential in their ability to perform complex cognitive tasks that were previously thought to be only achievable by humans. These systems can recognize objects [7], locate objects[9], reason about relations[5] and play complex video games[19]. However, recent studies indicate that systems trained on selected databases might suffer from training set bias and produce unfair predictions. This issue is especially evident when we consider the task where minority groups are involved. For example, there existed an instance where an image classifier once labeled an Africa American as a gorilla which led to social outrage [1]. Occurrences such as this have led to the emerging field of ensuring machine learning fairness. The term fairness is oftentimes subjective with numerous possible interpretations [10]. In this project, we try to develop a general learning framework that attempts to ensure fairness by mitigating gender-bias originating from the training date set. The intuitive idea is that we can leverage auxiliary tasks, which could be dynamically selected according to the information provided, in the objective function. Not only could the auxiliary tasks alleviate biases, but it also could give an explanation what biases are discerned.

2 Related Work

Machine learning techniques is wildly adopted in many arenas of our life, which sometimes may exert negative influence or discrimination due to implicit biases in training data [12]. In word embeddings, verbs such as "cooking" were seen to be heavily biased towards females when compared to males. Therefore, it is important to take these issues into consideration when designing learning algorithms. To mitigate the above mentioned issues, researchers use a variety of tools to reduce dataset bias. For example, Zhao et al.[22] proposed to use corpus-level constraints to limit the output of the inference results. Zhao et al. [23] also proposed to learn a gender neural word embeddings. Yang et al. [20] proposed to use causality to reduce dataset bias.

Preprint. Work in progress.

Mutli-task learning[6] through auxiliary tasks [4] has demonstrated having a more robust and consistent representation. Researchers use this approach to achieve SOTA results on different learning tasks. For example, Sun et. al. [16] used auxiliary tasks to generate programs from video demonstrations. Li et al. [8] proposed to generate explanations while perform VQA tasks. This approach not only produced more accurate results, it also generated an explanation to users which increased users' trust in the system as a side effect. However, these systems treat the explanation process in the same manner as image captioning which are also heavily affected by dataset bias. Therefore, to overcome this issue we propose that a good learning system should not only provide explanations, but also know what to explain.

Neural Module Network [3] proposed by Andreas et al. can dynamically assemble neural networks to solve a complex VQA problem. Each neural module perform its own functionality. A reinforcement learning system will assemble different modules together to solve the problem according to the provided question and image. The idea of having a master policy to determine a sub policy has been extensively studied in[2, 11, 15, 17]. Hierarchical methods provide a indirect supervision through the structure of the hierarchy, and has been empirically shown to be more data efficient and more robust to noise.

Two distinct approaches are developed in debiasing, one manipulates training data, the other adjusts algorithm itself. In data manipulation, [23] proposed a data augmentation method to balance a disproportionate class by creating an augmented data set which is identical to the original data set and offsets the minority class. [18] proposed a tagging techniques to handle the gendered training data set. [13] introduced the bias fine-tuning approach which incorporates transfer learning from an unbiased data set to minimize bias before fine-tuning a model on a biased data set. Some approaches focused on adjusting algorithms to debias a biased training data set. [22] proposed Reducing Bias Amplification based on a constrained conditional model [14]. [21] mitigates bias by utilizing generative adversarial network to protect gendered data set.

References

[1] Google apologises for photos app's racist blunder, Jul 2015.

[2] Jacob Andreas, Dan Klein, and Sergey Levine. Modular multitask reinforcement learning with policy sketches, 2016.

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[4] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997.

[5] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*, 2020.

[6] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks, 2016.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions, 2018.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[10] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2019.

[11] Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning, 2018.

[12] United States. Executive Office of the President and John Podesta. *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President, 2014.

[13] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

[14] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. Technical report, ILLINOIS UNIV AT URBANA-CHAMPAIGN DEPT OF COMPUTER SCIENCE, 2004.

[15] Tianmin Shu, Caiming Xiong, and Richard Socher. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning, 2017.

[16] Shao-Hua Sun, Hyeonwoo Noh, Sriram Somasundaram, and Joseph Lim. Neural program synthesis from diverse demonstration videos. In *International Conference on Machine Learning*, pages 4790–4799, 2018.

[17] Shao-Hua Sun, Te-Lin Wu, and Joseph J. Lim. Program guided agent. In *International Conference on Learning Representations*, 2020.

[18] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*, 2019.

[19] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[20] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020.

[21] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[22] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

[23] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018.