Bias in Artifical Intelligence

fsdn71
Department of Computer Science
University of Durham
Durham, United Kingdom
fsdn71@durham.ac.uk

I. Research Commentary

In *Learning Adversarially Fair and Transferable Representations*, Madras et al. present a novel application of adversarial learning to the problem of removing disparate treatment and impact from data. Many other fairness methods describe algorithms applied to training data (pre-processing methods), independent from the type and properties of machine learning models which learn from it. However, this has several weaknesses- it does not guarantee that the model outputs will be free from bias - only that the training data is unbiased (at best). Furthermore, by excluding any modelling considerations from the fairness method, the utility of the model trained on the resulting data can be significantly reduced. The paper achieves a more reliable method of ensuring fairness by incorporating the naturally adversarial interests of fairness and utility into the training process.

The specific implementation of these interests is remarkably close to real-life pressures on machine learning models: a model is trained to maximise its predictive power, and questioned when its outputs are themselves predictive of a sensitive attribute. The proposed solution is close enough to the semantics of the actual objective that it seems inarguably appropriate on most machine learning models. The results of the paper present a solution that is extremely broad in its possible uses, as well as showing significant promise for development: questions on how to improve the balance between the adversary and the other models, or how to specialise the encoder-decoder models for specific datasets invite future research.

This paper is highly beneficial to the field of AI bias: it applies the novel topic of adversarial learning, behind many highly effective innovations such as in image processing [1] [2], therefore enabling future research in that area to benefit AI fairness research. Furthermore, the paper shows promising results on all definitions of fairness, adding to the research scene and opening new research applications in adversarial learning and fairness. The findings have influenced a number of other works involving AI bias and adversarial learning (for example, *FairGAN* [3]).

II. Future Approaches to Bias

Many of the fields in or adjacent to machine learning present significant opportunities for bias research in the future - adversarial learning one of them. Generative approaches are not just relevant to the field of bias, and therefore bias research will likely benefit from improvements in accuracy and the ability of models to generate convincing data under set constraints (for example, fairness measures). It is likely that performance in both utility and fairness can be improved by increasing the joint accuracy of fairness of pre-, in- and post-processing methods. In future, it might be expected that research will attempt a more unified solution to algorithmic bias that combines all three of these approaches - the division of AI bias counter-measures into these three categories is largely due to the different implementations required, rather than any fundamental difference in their purposes.

However, the question of what developments in the field of AI bias will occur in the future strongly depends on the evolution of the shared understanding of bias and discrimination among the AI bias research community. Prominently featured in the introductions of many of the most influential publications are various opinionated assertions over which measures of bias are correct and which are antithetical to the idea of fairness itself. For example, some introduce new approaches for reducing disparities in model error [4], whereas others focus on reducing disparate impact [5].

However, I find the notion of equality of opportunity within algorithms to be extremely flawed. There are only two justifications for preferring equal treatment by an algorithm. The first, because this target enables an algorithm to be profitable, is not morally persuasive. The second is unfortunately embedding deeply in the field of AI bias, and is itself a bias towards particular biases - *status quo bias*. This is the notion that of the biases present in the outputs of an algorithm, biases which also existed in the inputs (i.e. in the status quo) are less important. I find a utilitarian approach to fairness to be more logically sound: an algorithm is *in net* fair if it reduces the aggregate bias present in all of society, regardless of whether the algorithm itself shows bias. A preference for equal treatment is only justified by deliberately choosing to ignore an algorithm's interaction with the environment it is applied in, for which I do not believe there is any rationale.

The current methods for improving fairness under different methods are promising, but not sufficient. This is partially because of the hindrance to progress caused by differing definitions of fairness - not all of which correlate with each other [6]. Bias in AI has several challenges to overcome - some in implementation, others more fundamental - and the persistent demand for greater algorithmic fairness as automation increases in society restates the value it will generate.

References

[1] R. Huang, S. Zhang, T. Li, and R. He, “Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis,” *CoRR*, vol. abs/1704.04086, 2017. [Online]. Available: http://arxiv.org/abs/1704.04086

[2] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *CoRR*, vol. abs/1703.10593, 2017. [Online]. Available: http://arxiv.org/abs/1703.10593

[3] D. Xu, S. Yuan, L. Zhang, and X. Wu, “Fairgan: Fairness-aware generative adversarial networks,” *CoRR*, vol. abs/1805.11202, 2018. [Online]. Available: http://arxiv.org/abs/1805.11202

[4] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” 2016.

[5] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and removing disparate impact,” 2015.

[6] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” 2018.