Two-step adversarial debiasing with partial learning - medical image case-studies

Ramon Correa[1], **Jiwoong Jason Jeong**[1], **Bhavik Patel**[2,1], **Hari Trivedi**[3], **Judy W. Gichoya**[3], **Imon Banerjee**[2,1]
[1]SCAI, Arizona State University, USA
[2]Department of Radiology, Mayo Clinic, Arizona, USA
[3]Department of Radiology, Emory University, Atlanta, USA
imon.banerjee@asu.edu

Abstract

The use of artificial intelligence (AI) in healthcare has become a very active research area in the last few years. While significant progress has been made in image classification tasks, only a few AI methods are actually being deployed in hospitals. A major hurdle in actively using clinical AI models currently is the trustworthiness of these models. More often than not, these complex models are black boxes in which promising results are generated. However, when scrutinized, these models begin to reveal implicit biases during the decision making, such as detecting race and having bias towards ethnic groups and subpopulations. In our ongoing study, we develop a two-step adversarial debiasing approach with partial learning that can reduce the racial disparity while preserving the performance of the targeted task. The methodology has been evaluated on two independent medical image case-studies - chest X-ray and mammograms, and showed promises in bias reduction while preserving the targeted performance.

Introduction

Artificial Intelligence (AI) models have demonstrated expert-level performance in image-based diagnostic tasks, resulting in increased clinical adoption and FDA approvals. The new challenge in AI is to understand the limitations of models from the perspective of demographic bias in order to reduce potential harm. The unknown disparities based on demographic factors could worsen currently existing inequalities worsening patient care for some groups.

AI bias can be defined as models with outputs providing outcomes that negatively affects one sub-group of the study population more than others. Examples include differing allocation of healthcare resources based on patient demographics (Obermeyer et al. 2019; Benjamin 2019), bias in language models developed on clinical notes (Zhang et al. 2020), and melanoma detection models developed primarily on images of light-colored skin (Adamson and Smith 2018). In the clinical domain, unintended bias in AI systems affecting individuals unfairly based on race, gender, and other clinical characteristics has been highlighted in multiple studies (Parikh, Teeple, and Navathe ; Whittaker et al.).

AI models applied in other applications have also presented similar biases such as face detection models failing to correctly identify individuals of minority groups (Buolamwini and Gebru). Given such examples, biased AI systems can result in a variety of fairness-related harms, particularly in healthcare.

A core challenge for reducing AI bias is that the model final performance and reasons resulting in unfairness of AI models are not mutually exclusive and can often exacerbate one another. Recently, (Banerjee et al. 2021) showed that AI models trained for diagnosis can learn unintended racial information from different imaging modalities. Thus, AI models may use learned demographic information for detecting a diagnosis even when such attribute is not associated with the diagnosis. There are examples of race-ethnicity and gender influencing clinician decision-making, and given that AI is trained on real-world data, it is reasonable to expect that models would learn similar biases.

Common technique adopted within the community to reduce biases is curating a training dataset with greater number of positive cases across demographics (Larrazabal et al.). Other popular approaches to remove biases such as building demographic-specific models often suffer from a lack of demographic representation. We observed that techniques attempting to decouple demographic information and task predictions are not able to match baseline model performances (Seyyed-Kalantari et al.). Our goal is to develop an efficient methodology for model debiasing without the need for demographically balanced datasets and simultaneously match the baseline model performance.

The core contributions of the current article are -

1. Present AI model bias in terms of patients' race for two prevailing use-cases - chest X-ray and mammogram image interpretation.

2. Reducing racial bias by implementing a novel adversarial debiasing technique with partial model tuning while preserving the baseline model performance.

3. Compare the performance of full and partial debiasing techniques for both chest X-ray and mammogram image interpretation.

Methodology

In Fig. 1, we present a simplistic visual of the proposed architecture which contains two parallel branches after core CNN backbone network - (1) **predictor** - train to predict targeted variable y given input X by minimizing cost $L_{predictor}(y, \hat{y})$. $\hat{y}$ is the model prediction given input X which can be modeled as a $f(X)$; (2) **adversarial** - predict the protected variable Z given input X and reverse the gradient for penalizing learning of protected variable. Hypothetically, often $f(X)$ is highly dependent on protected variable Z and penalizing the learning of protected often significantly hamper the prediction performance of the target data. Demographic factors, including race, can be modeled as a protected variable.

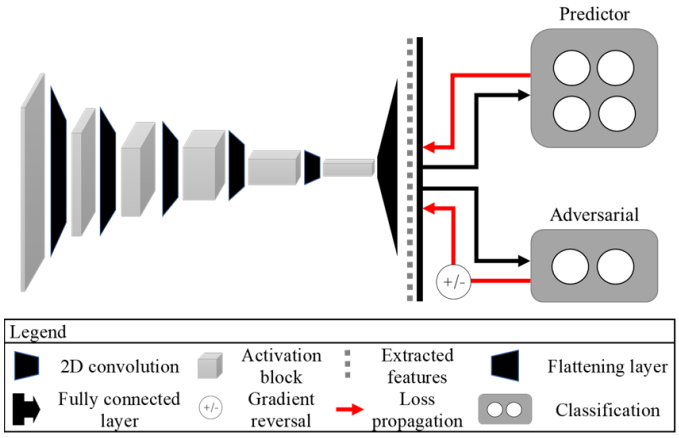

Figure 1: Architecture of the proposed architecture - adversarial and predictor branch

Training of the proposed architecture involves two backward passes. In the first pass, the model minimize the loss for both predictor and adversarial branches as $L = L_{predictor} + \lambda * L_{adversarial}$ where $L_{adversarial}$ is loss for learning the protected variable Z and λ is the regularization factor which can be in the range of $[0, 1]$. In the second pass, we only backpropagate the sign flipped gradient corresponding to the adversarial branch for modeling the penalization as $L = -\lambda * L_{adversarial}$. First pass helps the model to learn simultaneously the targeted and protected variables while the second pass intend to suppress learning of rotected variable Z. We used the same $\lambda = 0.53$ for both phases.

Partial fine-tuning: We explore the potential of improving model performance even after de-biasing by fine-tuning a subset of convolution layers of a pre-trained model instead of the complete network. The layers used for finetuning are identified via an *ablation study* which measures the performance difference of the targeted and protected variable prediction after dropping 10% of the top similar filters relative to the total number of filters from the targeted convolution layers (Meyes et al. 2019). It should be noted that due to the increasing sizes of the different convolutional layers, the same proportion may correspond to a different number of ablated filters. Layers where the protected variable prediction experienced higher performance degradation, was identified for finetuning alongside its downstream layers with the proposed two-step adversarial learning.

Table 1: Dataset description - chest X-ray and mammography images. Tissue density 1 = fatty, 2 = fibrogranular density, 3 = heterogeneously dense, 4 = extremely dense.

Chest X-ray Dataset					
Race		Gender		Age	
Black/ African American	38,024	Male	34,857	0-19	1,566
				20-39	13,237
White/ Caucasian	35,348	Female	38,515	40-59	21,227
				60-79	28,374
				80+	8,968
Total Patients		73,372			
Total Images		137,985			
Mammography Dataset					
Race		Tissue Density		Age	
Asian	1,305	1	1,853	<45	2,941
Black/ African American	8,164	2	7,408	45-59	7,279
		3	7,205	60-79	6,705
White/ Caucasian	7,924	4	873	80+	468
Total Patients		17,393			
Total Images		34,134			

Results

Datasets

In this study, we validated the performance of the proposed architecture for the following two independent use-cases - (1) *Diagnosis from chest X-ray images* - We received the de-identified dataset of 137,985 chest x-ray images of 73,372 unique patients from Emory University hospital. The demographic factors are described in Table 1. The targeted task is to identify four common radiology findings - i) atelectasis, 2) edema, 3) pneumothorax, and 4) normal cases, and patient race is considered as protected variable.
(2) *Infer tissue density from the mammogram images* - We received the de-identified dataset of 34,134 mammogram images of 17,393 unique patients from Emory University hospital (see Table 1). The targeted task is to classify the images based on breast tissue density inferred manually by the expert radiologist, and similar to the chest x-ray, patient race is considered as protected variable.

Quantitative Performance

We have compared the performance three models - (1) *Baseline* - a CNN backbone but with no de-biasing, (2) *Full debias* - 2-step adversarial training of the same CNN backbone, (3) *Partial debias* - 2-step adversarial training of the same CNN backbone with ablation for indentifying optimal

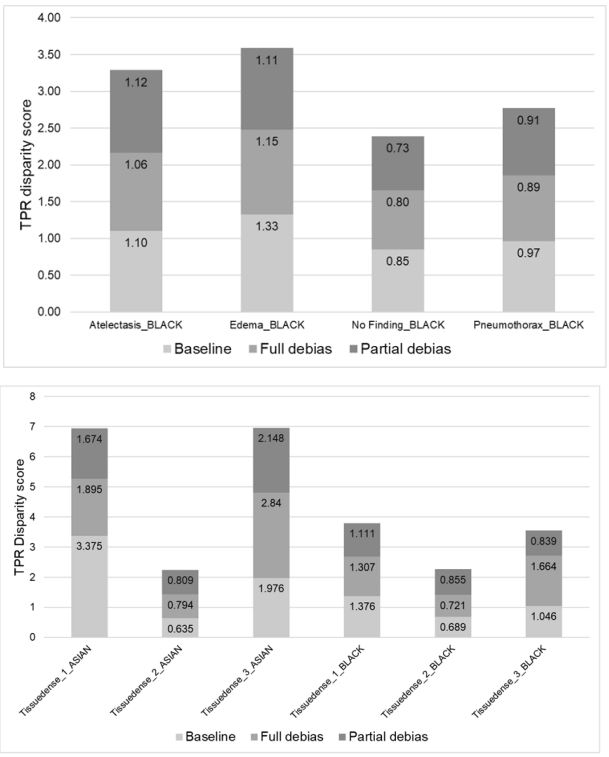

Figure 2: Class-wise TPR disparity plots for baseline, partial and full debias model: top plot presents chest x-ray and bottom plot presents mammogram use case. Caucasian(white) is represented as ref group in both disparity measures.

layers for learning. In order to evaluate the models, we generated a patient-wise separation of the $train : validation : test$ splits for both case-studies as $60 : 20 : 20$ and presented the results of all the models on the test set. Table 2 presents class-wise performance of all the three models in terms of AUCROC, Precision and recall. Bias of the models is evaluated in terms of True Positive Disparity (TPR) score where the true positive rate of the target patient subgroup is compared with the reference group as $\frac{TPR_{target}}{TPR_{ref}}$. any disparity measure between 0.8 and 1.25 can be deemed fair as per the 80 percent rule for determining disparate impact (Corbett-Davies and Goel 2018).

As seen from Table 2, with partial and full debiasing, we achieved comparable overall target task performance on the mammogram images with the baseline (no debiasing). Interestingly, for chest X-ray case-study, we even observed slight performance improvement with partial debiasing over the baseline - no finding AUC improved from 0.86 to 0.89, atelectasis improved from 0.86 to 0.87.

Chest X-ray cases study is a classic case of low bias AI model where the TPR disparity of the chest X-ray case studies was within the acceptable range for the baseline model, except for the Edema class (see Fig. 2). With partial debiasing, we managed to reduce the TPR disparity of Edema from 1.33 to the acceptable range of 1.11 with no significant change in performance. For the mammography use-

Table 2: Performance analysis for the targeted task - chest X-ray diagnosis and breast density classification is represented within the same table.

Diagnosis from Chest X-Ray				
Disease	Metric	Model comparison		
		Baseline	Partial	Full
Atelectasis	AUC	0.865	0.870	0.873
	Precision	0.889	0.891	0.893
	Recall	0.925	0.924	0.926
Edema	AUC	0.898	0.883	0.884
	Precision	0.503	0.457	0.405
	Recall	0.511	0.525	0.484
Pneumothorax	AUC	0.829	0.837	0.857
	Precision	0.558	0.586	0.591
	Recall	0.460	0.505	0.512
No Finding	AUC	0.866	0.889	0.846
	Precision	0.346	0.336	0.349
	Recall	0.188	0.313	0.313
Mammography Debiasing Results				
Infer breast tissue densities	Metric	Model comparison		
		Baseline	Partial	Full
1	AUC	0.965	0.960	0.942
	Precision	0.637	0.709	0.637
	Recall	0.858	0.677	0.682
2	AUC	0.899	0.896	0.879
	Precision	0.781	0.769	0.763
	Recall	0.765	0.758	0.736
3	AUC	0.923	0.895	0.917
	Precision	0.879	0.825	0.831
	Recall	0.739	0.727	0.821
4	AUC	0.979	0.972	0.957
	Precision	0.413	0.324	0.482
	Recall	0.867	0.883	0.625

case, the TPR disparity was prominent in baseline model for both African American and Asian patients for all the tissue density classes. Given the known correlation between patient race and breast tissue density, full debias model performance degraded from the baseline while disparity didn't improve. With the new partial de-biasing technique, we preserve the baseline performance while reducing the TPR disparity for African American patients. TPR disparities for Asian patients on the low tissue density classes were improved from 3.37 to 1.89 with full debiasing and to 1.67 with partial debiasing. However, no significant change was observed in disparity score for heterogeneously dense tissue class for Asian patients. This could be due to the fact that most of the Asian women have denser breasts on mammography thus reducing disparity for denser tissue class is extremely challenging (Bae and Kim 2016).

Conclusion

We proposed a two step adversarial debiasing method with partial learning and evaluated the approach on two distinct medical image datasets. The proposed architecture can successfully preserved the targeted task performance while reducing the TPR disparity. This describes our initial experiments with the proposed methodology. In future, we plan to apply this technique for predictive analytics for healthcare problems while reducing the scocio-economical bias. The adversarial training approach described can be applied regardless of predictor's model architecture, as long as the model is trained using a gradient-based method.

References

[Adamson and Smith 2018] Adamson, A. S., and Smith, A. 2018. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology* 154(11):1247–1248.

[Bae and Kim 2016] Bae, J.-M., and Kim, E. H. 2016. Breast density and risk of breast cancer in asian women: a meta-analysis of observational studies. *Journal of Preventive Medicine and Public Health* 49(6):367.

[Banerjee et al. 2021] Banerjee, I.; Bhimireddy, A. R.; Burns, J. L.; Celi, L. A.; Chen, L.-C.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.-C.; Kuo, P.-C.; Lungren, M. P.; Palmer, L.; Price, B. J.; Purkayastha, S.; Pyrros, A.; Oakden-Rayner, L.; Okechukwu, C.; Seyyed-Kalantari, L.; Trivedi, H.; Wang, R.; Zaiman, Z.; Zhang, H.; and Gichoya, J. W. 2021. Reading race: Ai recognises patient's racial identity in medical images.

[Benjamin 2019] Benjamin, R. 2019. Assessing risk, automating racism. *Science* 366(6464):421–422.

[Buolamwini and Gebru] Buolamwini, J., and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. 15.

[Corbett-Davies and Goel 2018] Corbett-Davies, S., and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

[Larrazabal et al.] Larrazabal, A. J.; Nieto, N.; Peterson, V.; Milone, D. H.; and Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. 117(23):12592–12594. Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/117/23/12592.full.pdf.

[Meyes et al. 2019] Meyes, R.; Lu, M.; de Puiseau, C. W.; and Meisen, T. 2019. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*.

[Obermeyer et al. 2019] Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453.

[Parikh, Teeple, and Navathe] Parikh, R. B.; Teeple, S.; and Navathe, A. S. Addressing bias in artificial intelligence in health care. 322(24):2377.

[Seyyed-Kalantari et al.] Seyyed-Kalantari, L.; Liu, G.; McDermott, M.; Chen, I. Y.; and Ghassemi, M. CheXclusion: Fairness gaps in deep chest x-ray classifiers.

[Whittaker et al.] Whittaker, M.; Alper, M.; College, O.; Kaziunas, L.; and Morris, M. R. Disability, bias, and AI. 32.

[Zhang et al. 2020] Zhang, H.; Lu, A. X.; Abdalla, M.; McDermott, M.; and Ghassemi, M. 2020. Hurtful words: Quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, 110–120. New York, NY, USA: Association for Computing Machinery.