A Compliance Checking Framework for DNN Models

Sunny Verma[1,2], **Chen Wang**[2,*], **Liming Zhu**[2] and **Wei Liu**[1]

[1]Advanced Analytics Institute, School of Computer Science, University of Technology Sydney, Australia
[2]Commonwealth Scientific and Industrial Research Organisation, CSIRO, Data61, Sydney, Australia
Sunny.Verma@student.uts.edu.au, Wei.Liu@uts.edu.au, {Chen.Wang, Liming.Zhu}@data61.csiro.au

Abstract

Growing awareness towards ethical use of machine learning (ML) models has created a surge for the development of fair models. Existing work in this regard assumes the presence of sensitive attributes in the data and hence can build classifiers whose decisions remain agnostic to such attributes. However, in the real world settings, the end-user of the ML model is unaware of the training data; besides, building custom models is not always feasible. Moreover, utilizing a pre-trained model with high accuracy on certain dataset can not be assumed to be fair. Unknown biases in the training data are the true culprit for unfair models (i.e., disparate performance for groups in the dataset). In this preliminary research, we propose a different lens for building fair models by enabling the user with tools to discover blind spots and biases in a pre-trained model and augment them with corrective measures.

1 Introduction

Deep learning has significantly improved classification accuracy for supervised image recognition tasks. These tremendous performances are reported on test-sets which are drawn from distribution identical to their training datasets. Therefore, evaluating the performance of classifiers with error rate is necessary but not sufficient, since the error rate is agnostic to real-world complexities like open-set and dataset-bias. Hence multiple evaluation criteria are required to robustly evaluate classifiers' performance especially when social deployment of such models is required [Kleinberg *et al.*, 2019].

Recently several solutions have been proposed which proposes corrections while training models to reduce such disparate performances [Kleinberg *et al.*, 2019]. However, the proposed work assumed prior knowledge regarding sensitive attributes in the dataset and hence their approach is limited to the creation of fair (or unbiased) models only when one has training data. This assumption is unrealistic in the real-world due to three reasons 1) the training data is not always available to model users; 2) unknown dataset biases persist; and 3) limited to structured text datasets and not images. Defining sensitive attributes related to an image (or group of images) is impossible [Torralba and Efros, 2011] and varies on the user's perspective. Therefore, defining constraints for image classification models is like searching needle in a haystack.

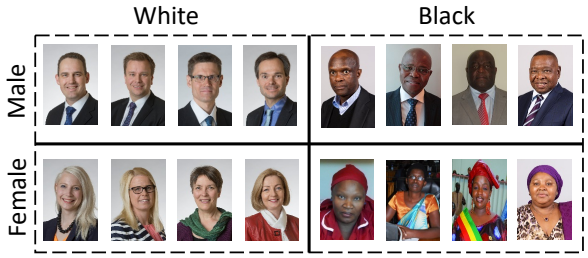

Figure 1: Images from PPB dataset [Buolamwini and Gebru, 2018].

A promising empirical evaluation for the above discussion is presented in [Buolamwini and Gebru, 2018], where the authors created a test-bed named PPB for gender recognition. PPB was created with demographic parity based on Fitzpatrick skin types (example images shown in Fig. 1). The authors demonstrated that despite of low error rates achieved by commercial classifier, their misclassification on PPB test-set is biased towards darker skin individuals. Moreover, the bias in these classifiers is remarkably significant for darker skin females. In this research, we ask whether there is a systematic way to identify all-types of biases.

2 Proposed Methodology

We define feature vector of an image I by $x \in \mathbb{R}$ and an ideal function $f(.)$ which classifies x into it's respective category $y \in Y$. Since f is not known, we approximate it with another function $g(.)$ by learning it in a supervised manner on dataset X. Also, assume the feature utilized for classification is $h(x)$, where $h(.)$ represents a deep neural network. Since $g(.)$ learns the representation h from X, this h might be biased due to presence of unkown (or known) dataset biases [Torralba and Efros, 2011]. In other words, the decision made by $g(h(x))$ might be biased on certain attributes like skin-color, whereas the decision from the ideal function i.e. $f(h(x))$ is unbiased to any such attribute.

We begin from searching blind spots in our model $g(h(x))$ by defining policies $P_i, i = [1, n]$ which the model must comply. Each policy states the restriction of using certain features in decision making, for example surveillance models for criminal suspicion must be uninfluenced by skin color or de-

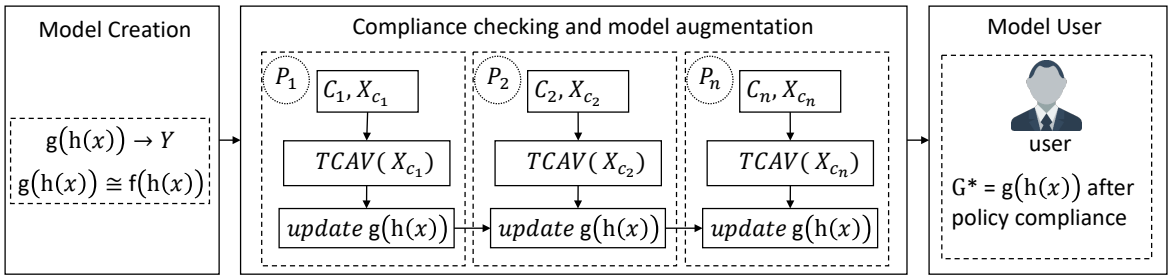

Figure 2: Compliance checking for a pre-trained model.

InceptionV3	Male	Female
White	100%	91%
Black	90%	74%

Table 1: F1-score of different gender groups from PPB dataset.

mography of face images. Since defining sensitive attributes with individual image is very difficult and time consuming, we define concepts c related to each policy. Collecting images given concept can be easily obtained by visual inspection for example concept can be females (males) with bangs.

These concept images are utilized to determine the sensitivity of $g(h(x|c))$. Prior work for determining important features in an image's prediction is achieved by saliency analysis. However saliency analysis has two drawbacks 1) it is unreliable [Kim *et al.*, 2018] and more importantly 2) we do not know how the concept is represented in the hidden representation of image. Hence, we utilize testing with concept activation vector (TCAV) [Kim *et al.*, 2018] to get the representation of high-level features, or concepts to discover blind-spots and biases in pre-trained model.

The user can utilize images collected for each concept and determine whether the model is biased or actually contains a blind-spot when faced with these two situations a) disparate TCAV-score between categories and 2) low TCAV-score for all the categories. The utilization of TCAV framework by defining policy as 'gender recognition must not depend on skin color' is described in the next section. The proposed model compliance framework is shown in Fig. 2 and the accuracy for each gender based on this policy is shown in Tab. 1.

3 Results and Discussions

We utilized publicly available InceptionV3[1] model and TCAV for compliance checking by defining policy 'gender recognition should remain unaffected by skin color.' We collected images related to concept skin color (black and white) from Google, and TCAV F1-scores on PPB dataset for various layers in InceptionV3 are plotted as bar-graph in Fig. 3.

It is clearly visible that layers $7b$ and $7c$ in Fig. 3 show that skin color plays a significant role in gender prediction of females than males. The disparate TCAV F1-scores indicates the presence of bias in the pre-trained model. Once discovered, these biases could be removed by model correction procedures which are currently under development for this work.

[1]https://github.com/dpressel/rude-carnie

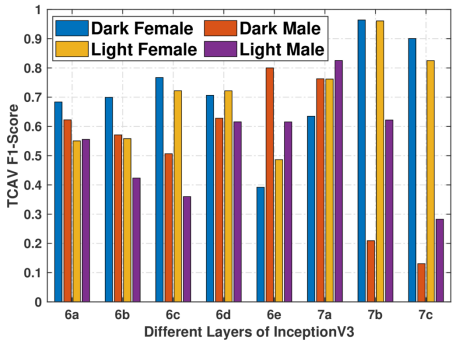

Figure 3: TCAV F1-Score for Females and Males

4 Conclusions and Future Work

In this preliminary research, we have presented a framework on how to discover biases and blind-spots in pre-trained models. In this regard, certain policies need to be defined which helps in determining compliance of the model. Model correction procedure for alleviating discovered biases is currently under development for this work. Moreover, the TCAV framework is subtle to the concept images, and creating a robust alternative is planned as future direction of our work.

References

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 2018.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2673–2682, 2018.

[Kleinberg *et al.*, 2019] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. Technical report, National Bureau of Economic Research, 2019.

[Torralba and Efros, 2011] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition, IEEE Conference on*, 2011.