Now You See Me: Detecting Bias Through Interpretable Deep Learning

Caroline Cocca
Whiting School of Engineering
Johns Hopkins University
Baltimore, MD, USA
ccocca2@jhu.edu

***Abstract*—When deep neural networks are deployed, their black-box nature results in many models being used in the field without an understanding of how they are making classification decisions. This lack of insight can lead to many performance issues including undetected bias during the training and evaluation phase of the machine learning pipeline. Failure to identify and correct such biases can cause significant harm to underrepresented groups that the model is biased against. Recent advancements in explainable deep learning models, such as ProtoPNet [1], help us to unravel the black-box nature of neural networks and gain a better understanding of the patterns learned during training. My experiment in training a model with the ProtoPNet architecture on a gender classification dataset explores the visualization of the predictive process on a biased dataset and whether such visualization can assist in detecting bias.**

Keywords—deep learning, explainable machine learning, convolutional neural network, bias detection.

I. INTRODUCTION

In the past decade, deep learning has been applied to a wide variety of fields including legal [2], medical [3], and real estate [4], to name only a few of many. The global deep learning market was US$12.3 billion in 2020, and is expected to grow to a staggering US$60.5 billion by 2025 [5]. At least in the projected short-term, it is clear that deep learning is only going to see an increase in its breadth and depth of applications. As we progress towards a future where deep learning is heavily utilized in our society, we must also consider current weaknesses and risks that need to be addressed and mitigated before deployment occurs.

Although deep learning brings with it powerful data analysis and significant time saved from human labor, these models do not offer the perspective of the unbiased machine that many perceive to be the case. On the contrary, a model will learn the bias that is present in its training data [6], and as a result it is just as capable of human bias as we are. Along with the increase in usage of deep learning models, there have been a number of scandals on high-profile use cases that contained significant bias. Examples include delivery apps [7], Amazon hiring practices [8], and Google Translate [9]. Biased models such as these stand to deal significant harm to underrepresented groups, such as costing a candidate a job opportunity that they would have otherwise secured, or a defendant a harsher sentence than another defendant of a different racial background would have received. Clearly, investigating methods of bias detection and mitigation is one of the most pressing areas of deep learning research today.

Alongside recent increases in usage of deep learning models, there has also been advancement in developing explainable models. For example, ProtoPNet is a convolutional neural network that classifies images of birds by species in a qualitatively similar manner to how a human would do so [1]. The model is forced to compare input images to 'prototype' images of species from its training set, which is done so through weighting sections of the input image by similarity with sections of the prototype image. For example, one species may receive a large positive weight due to the beak section of the input image being evaluated as similar to the beak section of that species' prototype image [1]. This prototype layer transforms the convolutional neural network from a black box to blindly trust into a classifier with an architecture that is explainable even to a layman with little knowledge of deep learning.

While we advance machine learning explainability, it is crucial that we must ask how this explainability can drive improvement in other deep learning research areas. There are areas of research that have not yet been thoroughly explored in the application of explainable models, such as developing methods for deep neural network debugging, ensuring model trustworthiness, and evaluating explainability methods beyond accuracy [10]. In this study, I explore these areas through investigating the potential for detecting bias through the prototype layer of the ProtoPNet model.

The logic behind this investigation is driven by the hypothesis that if a deep learning model contains explainable pattern recognition, then its biases will also be explainable. In a traditional convolutional neural network, it is impossible to determine its specifically learned biases through output observations alone. What features has it learned useful relationships with the target variable for males that it has not successfully learned for females? Which features common to specific racial backgrounds is it weighting inappropriately? In the prototype layer of ProtoPNet, it is possible to observe the specific sections of the input images that it is giving positive or negative weight towards sections of prototype images. In this experiment, I explore the ability to observe bias in this layer after training ProtoPNet on a biased gender classification dataset where dark-skinned individuals are underrepresented.

II. RELATED WORK

In a similar study, a researcher proposed a method named score-based resampling (SBR) in order to identify underrepresented samples of a training dataset based on CNN model output and augment those samples in order to reduce

bias [11]. This method involved observing the scores in the final layer of a typical CNN architecture for each class as a measure of distance of the given training sample from the classes. Data augmentation was then performed on poor-scoring observations and added back into the training dataset. Utilizing the augmented dataset to re-train the CNN resulted in improved accuracy for underrepresented samples in the dataset. However, the scores in the final layer of a CNN are not guaranteed to be accurate indicators of the true confidence of the network towards that prediction, and often significant modifications are required to calibrate the scoring outputs of the last layer so that they are better indicators of model confidence on a given sample [12]. ProtoPNet does not rely on the scoring outputs of the final layer to judge similarity, and thus will not suffer the same weakness when utilizing the similarity weightings of its prototype layer to detect underrepresented samples.

Researchers in another recent study investigated the potential for orthogonal feature projection to diagnose bias in black-box models [13]. In this experiment, an iterative process is performed where one feature is made orthogonal to all other features in the training dataset to 'remove' this feature. Each resulting dataset is used to train a black-box model, and the following set of models are tested for variability of performance. Through this process, an attribute ranking is obtained that shows the model's predictive dependence on each feature. This can be used to diagnose bias by observing if a model is inappropriately relying heavily on a specific feature, such as income, for its predictions. This method can be used with tabular data, but in the case of unstructured data such as images, this attribute ranking cannot be obtained, and thus further investigation must be performed towards detecting bias in imagery datasets and trained CNN models.

III. Dataset

This investigation was performed with a gender classification dataset hosted on Kaggle [14], which contains 47,009 training images and 11,649 test images. Each observation is an RGB image of a roughly centered face which is labeled either as male or female, samples of which are shown in figure 1. Due to hardware limitations, the training and test sets used to train the ProtoPNet model were sampled down to 25% of their original size, bringing the train set size to 11,752 images and the test set size to 2,912 images. This size reduction was made in order to reduce exceedingly long training and testing times. Furthermore, the images were compressed to 64x64 pixels to standardize their size and convert the rectangular images to square images in order to obtain the correct input format for the model.

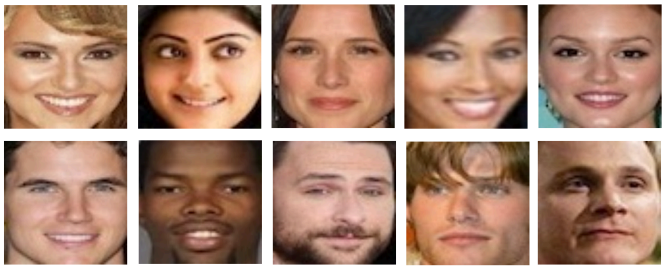

Fig. 1: Female (top) and male (bottom) samples from the gender classification dataset.

The selection of this dataset was motivated by [11] in which Derman utilizes SBR to determine images that contain underrepresented samples. In the paper, Derman discovers a discrepancy in the accuracy between genders, where performance tended to be worse on classification of female images. There were also slight discrepancies between performance on classification of light-skinned individuals vs. dark-skinned individuals. This points towards gender bias, as well as potential racial bias, being present in the dataset. Because the goal of this experiment is to detect bias in datasets during the training and evaluation phase of model development, this experiment required a biased dataset, which is demonstrated to be true for the selected gender classification dataset in [11]. Furthermore, using this dataset allows for a direct comparison in bias detection between SBR and the prototype layer of ProtoPNet.

IV. Methodology and Discussion

The ProtoPNet architecture was trained on the gender classification dataset subset described above, with 5 epochs completed before prototype generation. During prototype generation, prototype images are selected for each class by determining which images result in the smallest change between previously correct predictions and the consequent predictions after the model is forced to make said prototype comparisons [1]. After this stage, 20 more training epochs were performed, resulting in a final train accuracy of 94.11% and test accuracy of 93.51%. Compared with the standard CNN performance in [11] of 95.23% validation accuracy, these forced prototype comparisons sacrifice only a small amount of accuracy for increased model explainability. These results are similar to the performance of ProtoPNet on the CUB-200-2011 dataset compared with standard CNN models [1].

Images from the validation set were then randomly sampled and fed forward through the trained network in order to record the prototype layer output for the image subset. Surprisingly, as can be seen in figure 2, the highlighted sections of the validation images that ranked the most highly when compared with the prototype images were almost always the top right and left corners of the image for the female class, i.e. the forehead or the start of the hairline. Similarly, as in figure 3, almost all highly ranked sections for male images were either the lower right corner of the image containing the side of the chin, or one of the top corners. Although focusing on the chin may often make sense for images of males due to seeking out the presence of facial hair as a sign of gender, this does not fully explain the utter lack of importance placed on sections other than the image corners. As such, the eyes, nose, and mouth are not compared significantly with the prototype images for either class.

Although we are able to see the model's predictions in a way that is qualitatively similar to how a human would classify an image (i.e. placing importance on comparisons between different sections of the image and a known ground truth), because these comparisons are largely being made based on an exploited pattern that is not a true indicator of gender, this visualization is not helpful for spotting the underrepresentation of female and/or dark-skinned individuals in this dataset.

V. Conclusion

Although the prototype comparison outputs of the trained ProtoPNet model did not contribute significantly to visualizing

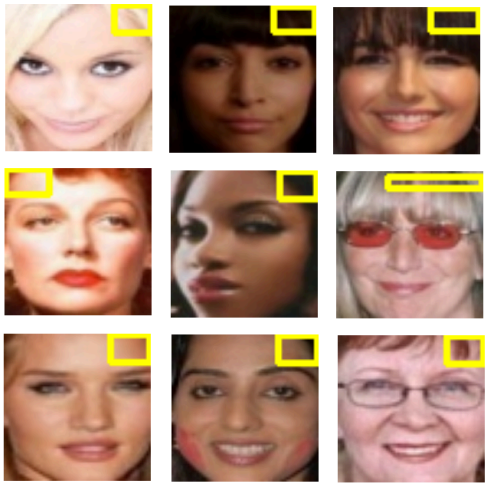

Fig. 2: Most highly scored prototype sections for light-skinned female samples (left), dark-skinned female samples (middle), and examples of the female prototype images (right).

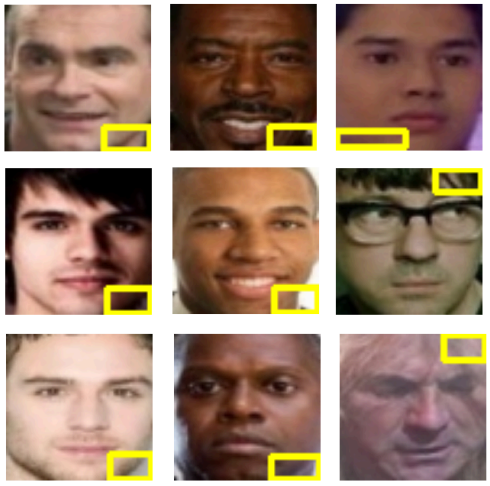

Fig. 3: Most highly scored prototype sections for light-skinned male samples (left), dark-skinned male samples (middle), and examples of the male prototype images (right).

model bias against female or dark-skinned individuals, these outputs did reveal that the model was not making useful comparisons for a *majority* of images, regardless of gender or skin tone. Just as with undetected bias, it can be very dangerous to deploy a model that seems highly confident and accurate in its predictions, but in reality is exploiting a pattern that does not contain useful, generalizable information on the target variable. Thus, there is value alone in the prototype visualizations revealing such exploitation.

Furthermore, this discovery prompts the question of the accuracy of [11] in identifying underrepresented samples in this dataset. As stated previously, the scores of the final layer of a CNN are not absolute indicators of true network confidence towards a given prediction. It is possible that the distance measures used to define underrepresented samples in [11] may have merely been an indication of images that did not fit into an exploited pattern that the model may have learned. These results serve as a reminder that when researching methods of bias detection, one must keep in mind that the black-box nature of neural networks may produce deceptive results— and having a tool such as ProtoPNet can be invaluable in visualizing precisely what the model has learned in order to determine whether that learned knowledge truly contains bias.

References

1. C. Chen, O. Li, C. Tao, A. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," in Proc. 33rd Conf. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 8930–8941.
2. B. Marr, "How AI and machine learning are transforming law firms and the legal sector," Forbes, 2018, https://www.forbes.com/sites/bernardmarr/2018/05/23/how-ai-and-machine-learning-are-transforming-law-firms-and-the-legal-sector/
3. J. Bresnick, "What is deep learning and how will it change healthcare?" HealthITAnalytics, Xtelligent Healthcare Media, 2018, https://healthitanalytics.com/features/what-is-deep-learning-and-how-will-it-change-healthcare
4. D. Olick, "Artificial intelligence is taking over real estate– here's what that means for homebuyers," CNBC, 2021, https://www.cnbc.com/2021/09/17/what-artificial-intelligence-means-for-homebuyers-real-estate-market.html
5. A. Dutta, "The global deep learning market is projected to reach US$60.5 billion by 2025," Analytics Insight, 2021, https://www.analyticsinsight.net/the-global-deep-learning-market-is-projected-to-reach-us60-5-billion-by-2025/
6. G. Ras, M. van Gerven, and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," pages 19–36. Springer International Publishing, Cham, 2018. ISBN 978-3-319-98131-4. doi:10.1007/978-3-319-98131-4_2.
7. N. Lomas, "Italian court rules against 'discriminatory' Deliveroo rider-ranking algorithm," TechCrunch, 2021, https://techcrunch.com/2021/01/04/italian-court-rules-against-discriminatory-deliveroo-rider-ranking-algorithm/
8. J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
9. N. Kayser-Bril, "Female historians and male nurses do not exist, Google Translate tells its European users," Algorithm Watch, 2020, https://algorithmwatch.org/en/google-translate-gender-bias/
10. N. Xie, G. Ras, M. van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," arXiv preprint arXiv:2004.14545, 2020
11. E. Derman, "Dataset bias mitigation through analysis of CNN training scores," arXiv preprint arXiv:2106.14829, 2021
12. A. Mandelbaum and D. Weinshall, "Distance-based confidence score for neural network classifiers," arXiv preprint arXiv:1709.09844, 2017
13. J. Adebayo and L. Kagal, "Iterative orthogonal feature projection for diagnosing bias in black-box models," arXiv preprint arXiv:1611.04967, 2016
14. A. Chauhan. "Gender Classification Dataset," Kaggle, 2020, https://www.kaggle.com/datasets/cashutosh/gender-classification-dataset