Interpretation and Adversarial Attacks in CNN Gender Classification

Yuqi Sun, Guo Cheng

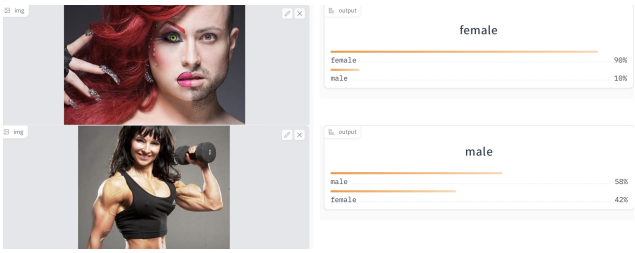

Figure 1. Biased CNN gender classifier with wrong results

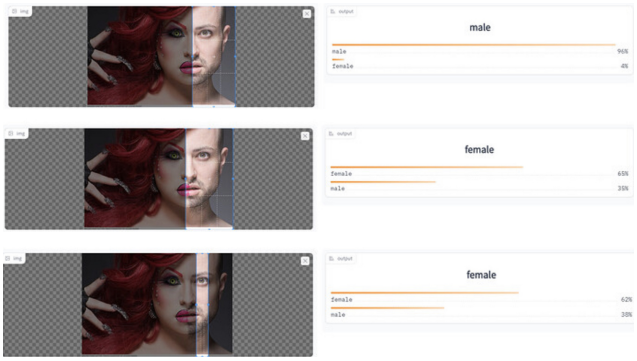

Figure 2. Different results of CNN gender classifier when the same image is cropped.
The upper and middle results show clearly how small changes in ROI could lead to largely varied prediction.

Abstract

In this project, we first used transfer training to train a convolution neural network for gender classification task. Based on its interesting prediction results, we were motivated to explain the model. We then introduced two model interpretation methods, LIME and SHAP. Both of them apply to basically any type of machine learning model. Based on the important features marked by these interpreters, we conducted adversarial attacks to fool the gender classifier.

1. Introduction

Our project starts with a very simple gender classifier using ResNet18 [1] backbone. As Fig. 1 shows, the prediction result of this network is biased, unable to correctly classify some images such as the ones of a muscular woman or a man with makeup.

We made the hypothesis that this might be caused by the potentially biased dataset we acquired by searching results on DuckDuckGo, which could include gender and racial stereotypes. To verify our hypothesis, we explored the deterministic effect of different Region Of Interest (ROI) of the input image, by simply cropping the same image as inputs and comparing the model prediction. As Fig. 2 suggested, the CNN prediction result is not robust, showing strong fluctuation on small changes in ROI.

However, this examination heavily depends on manual operation, thus is time consuming and intrinsically prevents a more thorough examination. As a result, we turned to exploring model interpretation tools to understand the predicting logic lays under our CNN classifier, which was long been considered as a "black box". In addition to our sheer curiosity about this black box, the model interpretation is important because it also helps to discover the errors made by the model and thus to build more reliable machine learning models. In general, linear and tree-based models can be easily interpreted because they make predictions in an intuitive way. On the other hand, ensemble models and deep neural networks are hard to interpreted because of their complexity. Deep neural networks are like black boxes. We don't know how neurons work together to arrive at the final results.

Over the years, researchers have developed many different types of model interpretability techniques, which can be categorized into different types [5]. In our project, we adopted two local, model-agnostic interpretation tools to explain the model. Local means the interpretation method explains an individual prediction rather than the entire model behavior. Model-agnostic means the explainer works for any kind of machine learning model, no matter how complicated the model is.

2. Interpretation Methods

2.1. LIME

LIME stands for Local Interpretable Model-agnostic Explanations, which was proposed by Marco et al. [6] in 2016. The goal of LIME is to identify interpretable models on interpretable representations that are locally faithful to the

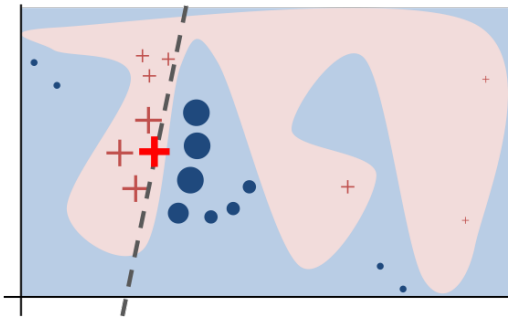

Figure 3. Toy example to present intuition for LIME. [6]

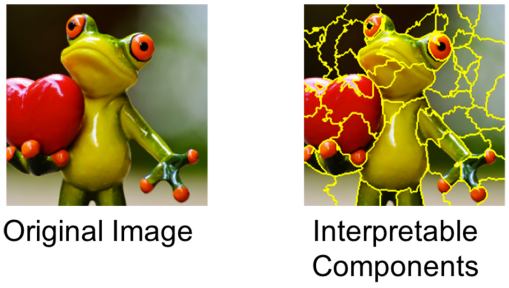

Figure 4. Transforming an image into interpretable components. (Source: Marco Tulio Ribeiro, 2016)

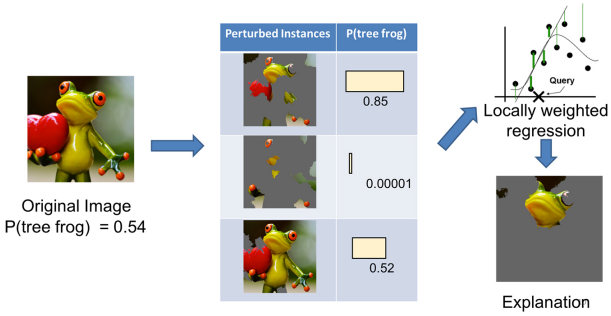

Figure 5. Explaining a prediction with LIME. (Source: Marco Tulio Ribeiro, 2016)

classifier. LIME can provide model-agnostic local explanations for regression and classification problems, and it can be applied to structured datasets and even unstructured datasets such as text and images.

2.1.1 Intuition behind

Intuitively, explanations are local linear approximations of the behavior of the model. The model's decision boundary can be globally complex and squiggly, but in the local region it is simple and linear. So it is easier to approximate the boundary around specific instances. When the model is treated as a black box, the instance to be explained is perturbed, and then a sparse linear model around it is learned as an explanation. Fig. 3 illustrates the intuition of this explanation process. The model's decision boundary is represented by the red and blue region and is clearly non-linear. The bright red cross is the instance being explained. The algorithm randomly generates samples around the instance and weights them. The closer those samples are to the instance, the more important they are. The importance is denoted by magnitude. LIME then learns a linear model (dashed line) that closely approximates the model around the instance, but not necessarily globally.

2.1.2 Explanation for image classifier

Unstructured data like free text and images cannot be classified as easily as structured numerical data. It is easy for LIME to interpret unstructured data using traditional methods of feature importance. However, it is challenging when interpreting complex deep neural networks trained on unstructured data such as images. When used for image classifiers, LIME attempts to highlight superpixels in the image that contribute positively or negatively to model predictions.

Fig. 4 is an example of how LIME works for an image classifier that predicts how likely it is a tree frog is in an image. The first step is to divide the original image into interpretable parts, which are contiguous superpixels.

As illustrated in Fig. 5, a dataset of perturbed instances is then generated by enabling or disabling some of the components. For each perturbed instance, we can get the probability of a tree frog's presence on it. The next step is to learn a linear model weighed locally on this dataset. Finally, the components(superpixels) with the highest positive weights are presented and everything else is grayed out.

2.1.3 Explanations lead to insights

Data collection artifacts often lead to unwanted correlations picked up by classifiers during training. These issues are difficult to identify only by checking the raw data and model predictions, but are likely to be identified by the interpreter. See Fig. 6 for an example of LIME in action for an image classifier distinguishing between wolves and huskies. Rather than examining features like sharp teeth or body shape, the decision is mainly based on the presence or absence of snow in the background. Although this classifier is highly accurate, it is right for the wrong reasons. So the explanation helps us know that this is a biased classifier and we shouldn't trust it.

2.2. SHAP

SHAP is the abbreviation of SHapley Additive exPlanation, proposed by Lundberg et al. [4] in 2017. Like LIME, SHAP is model-agnostic. Its mathematical foundation is

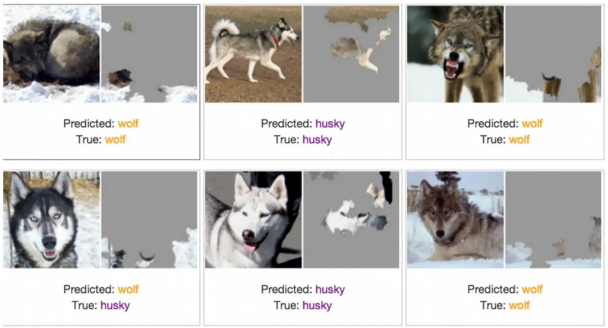

Figure 6. LIME explanation of the results. (Source: Singh, 2016)

the Shapley value from cooperative game theory [7]. The Shapley value is a method for assigning payouts to players depending on their contribution to the total payout. Players cooperate in a coalition and receive a certain profit from this cooperation.

Lundberg et al. [4] introduced the Shapley value to interpretable machine learning. Then what do the concepts of game, gain, and player from game theory represent in interpretable machine learning? The "game" is a prediction task on a single instance of a dataset. The "gain" is the actual prediction for that instance minus the average prediction for all instances. The "players" are values of instance features that work together to make predictions. A player can be an individual feature value, like tabular data. A player can also be a group of feature values. Take image data for example. A player can be superpixels grouped from pixels. The key idea of SHAP is to compute the Shapley value for each feature of a selected instance. Thus, the Shapley value is the average marginal contribution of a feature value across all possible coalitions. More simply, the Shapley value represents the importance of its associated feature for a particular prediction.

The innovation of SHAP is that the explanation model is represented as an additive feature attribution method, which is a linear function of binary variables. This view connects LIME and Shapley values. SHAP specifies the explanation as: [5]

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{1}$$

where g is the explanation model, $z' \in \{0,1\}^M$ is the coalition vector(simplified features), M is the maximum coalition size and $\phi_i \in R$ is the feature attribution for a feature i, the Shapley values. In the coalition vector, an entry of 1 indicates that the corresponding feature value is "present" and 0 that it is "absent". The coalition vector of the instance of interest is a vector of all 1's.

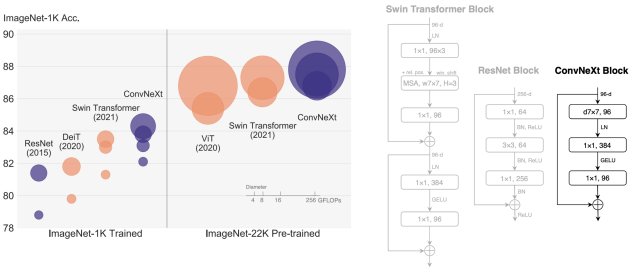

Figure 7. Performance and structure difference between ConvNeXt and ResNet (Source: Liu, 2022)

3. Experiments

3.1. Dataset and Model

We acquire a dataset of size 400 by using the DuckDuckGo image searching API, with a 0.2 valid-test split, and squishing item resizing. It is worth noting the inherent vulnerability and bias of this dataset as it is highly dependent on the search results of www.duckduckgo.com.

For better accuracy and more relevant interpretation results, we used the ConvNeXt_tiny_in22k instead of the original ResNet18 as the backbone of our gender classifier CNN model, and fine tuned it on the dataset we acquired by DuckDuckGo for 3 epochs. The final accuracy is 99.19%.

ConvNeXt [3] is a CNN model with even better results than Transformer, it retains the basic architecture of ResNet [1], while borrowing the successful experience of SwinTransformer [2], such as inverted bottleneck, large kernel and other techniques to obtain the top 1 accuracy on ImageNet. Fig. 7 illustrates the performance and main difference between ConvNeXt and ResNet.

3.2. Explanations

We collected several sets of images of different genders with different features as instances to explain our gender classifier. Interpretations can be displayed directly on those image instances. For the LIME interpretation, green indicates that this part of the image increases the predicted probability, and red indicates a decrease. For the SHAP explanation, red regions corresponding to image parts increase the predicted probability when they are included, while blue regions decrease the predicted probability.

Fig. 8 is the explanation of a girl with long hair. The prediction of the image is female with 100% probability. LIME explanations show that the female prediction is mainly based on long hair and facial features. SHAP explanations show that long hair is the most important feature on which the prediction is based.

What if we cut off the girl's hair? Our assumption is that the probability of female will decrease. Fig. 9 is the image of the same girl as Fig. 8. The only difference is the

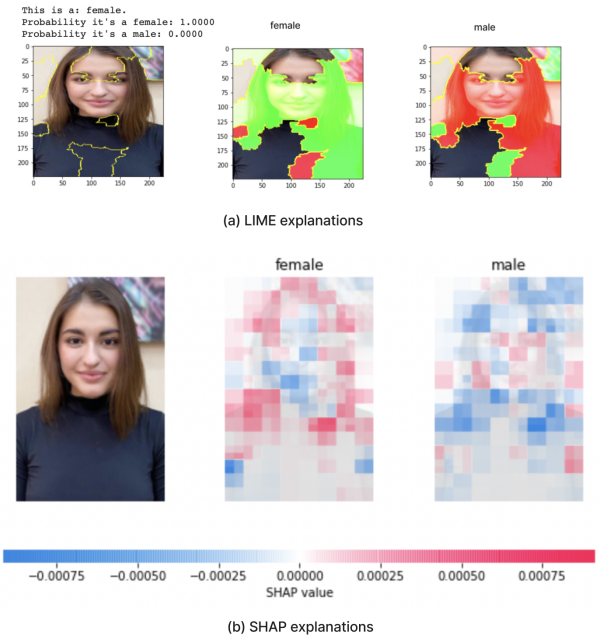

Figure 8. Explanations of a girl with long hair.

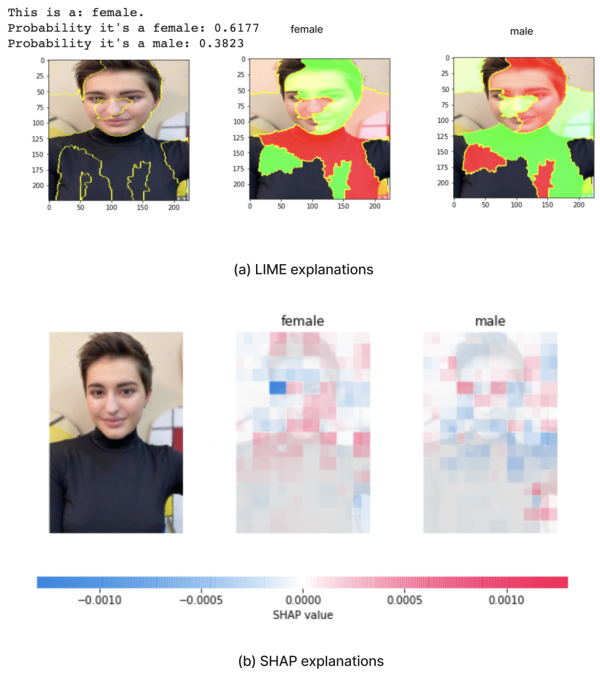

Figure 9. Explanations of a girl with short hair.

length of the hair. The prediction is still female. But the current probability is 62%. Both LIME explanations and SHAP explanations show that in the absence of long hair, the prediction is mainly based on facial features.

Fig. 10 is the explanation of a female with short hair

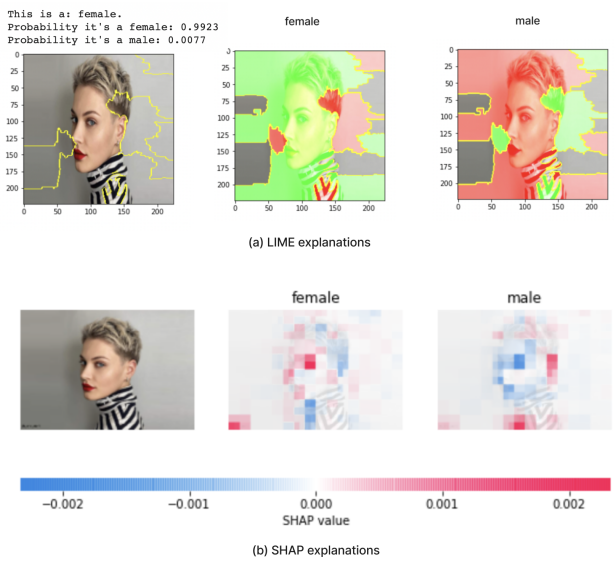

Figure 10. Explanations of a female with short hair and makeup.

and makeup. The prediction is female with a probability of 99%. LIME interpretation may not indicate which exact feature is the most important for prediction. But it is clear in the SHAP interpretation that makeup features contribute the most to the output. Because features like eye shadow and red lips are marked red.

Fig. 11 is the explanation of a male with beard. The prediction is male with a probability of 95%. LIME interpretations show that the prediction is based on beard, facial features and tattoos. SHAP interpretations show that beard has the great impact on the prediction.

Fig. 12 is an interesting example of a muscular female. However, the image is classified as a male with 99% probability. Why the probability of the wrong prediction is so high? Because masculinity occupies a large portion of the image. LIME explanations show that the prediction is based on these big muscles and the dumbbell. And SHAP explanations show that dumbbell is the most important feature to the prediciton.

4. Adversarial attacks

With the interpretation result given by LIME and SHAP, we can now execute adversarial attacks to fool our CNN classifier.

For the muscular female image, SHAP interpretation (Fig. 12) shows that our CNN classifier establishes a high correlation between black dumbbell bases and being a male. By simply adding a white block on the black dumbbell base region, the probability of this figure being a male dramatically dropped from 58% to 16%, as shown in Fig. 13.

The result above reveals how a seemingly well-performed model with high metrics scores can be vulnera-

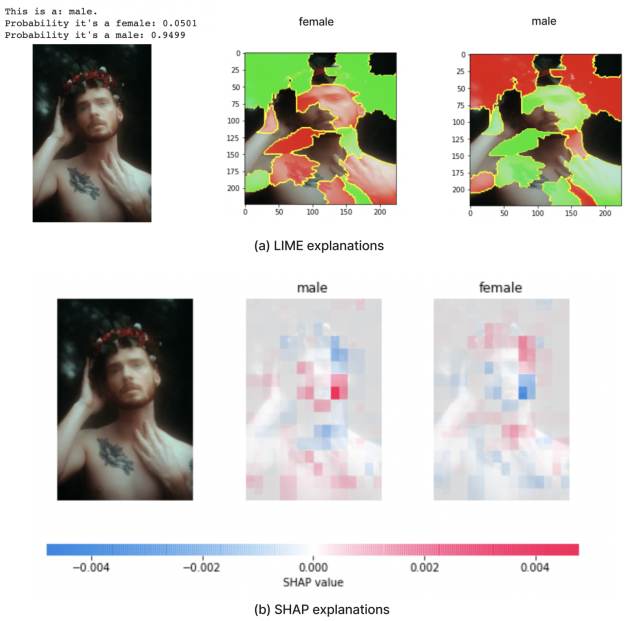

Figure 11. Explanations of a male with beard.

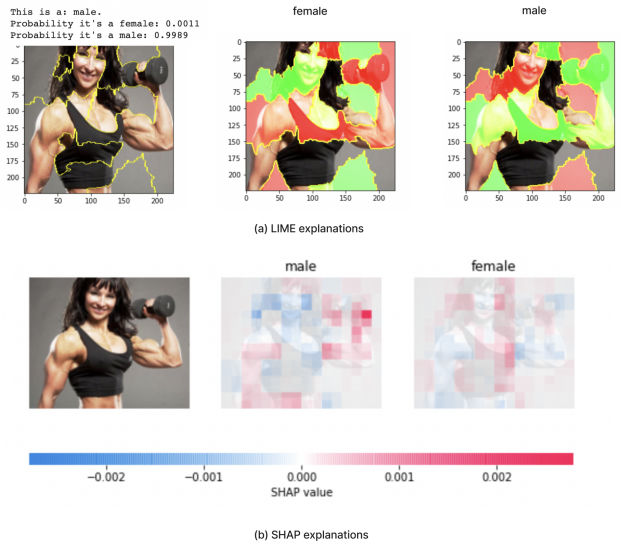

Figure 12. Explanations of a muscular female.

ble to some very simple and straight forward adversarial attacks. It may not be a big issue that a CNN gender classifier were fooled now and then, but considering the growing importance of CNNs and deep learning in general in our lives, the interpretability of "black box" CNN model should be should be taken seriously to prevent potential serious consequences.

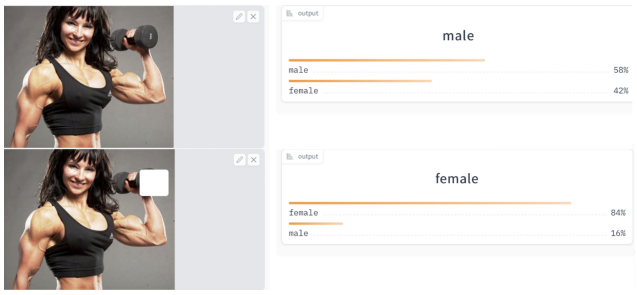

Figure 13. CNN predictions with and without adversarial attacks
Above: Prediction results of the original muscular female image
Below: Prediction results of the original muscular female image with adversarial attack (white block)

5. Conclusion and future work

In this report, we presented some interesting results of biased CNN classifier as our motivation for the interpretation of CNN, then introduced 2 interpretation methods Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanation (SHAP). By rebuilding our model to make it more accurate, we then performed experiments explaining several different instances and knew the features on which the corresponding predictions are based. Lastly, we executed a simple but effective adversarial attack, successfully fooled the high accuracy CNN classifier, revealed the potential dangers of uninterpreted CNN as a black box.

Due to time limits, there are many interesting future work awaits to be explored. For example, by leveraging SHAP and LIME, we can further acquire an adversarial data augmentation, and thus attain a more robust CNN model by fine-tuning on this adversarial augmented dataset.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[3] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 3

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 2, 3

[5] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. 1, 3

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 1, 2

[7] Lloyd S Shapley. A value for n-person games. *Classics in game theory*, 69, 1997. 3