ACE For Bias Detection

Introduction

Explainable AI is critical in data scientists' toolbelts to help mitigate bias in machine learning algorithms, especially facial recognition technologies. However, many explainable AI algorithms may not necessarily be understandable to the layperson. As noted in the paper “Impact of Explainable AI on Reduction of Algorithm Bias in Facial Recognition Technologies,”: “the success of XAI largely depends on whether users can understand explanations provided.” (B et al., 2024). For instance, the feature importance scores or saliency maps generated by a LIME or SHAP (or another feature-based explanation model) might not be understandable to a non-data science audience. Therefore, bias may be hard for the layperson to identify in features or pixels provided. Consequently, the data science community must explore other ways of making artificial intelligence, such as concepts or abstract, interpretable units (e.g., "striped pattern" in images, "tumor size" in medical data).

In “Towards Automatic Concept-based Explanations,” authors Ghorbani, Wexler, Zou, and Kim develop a new algorithm, ACE, to extract visual concepts automatically (Ghorbani et al., 2020). The paper provides systematic experiments to demonstrate that ACE discovers concepts meaningful to humans, coherent, and essential for the neural network’s predictions. ACE aims to explain high-level “concepts” instead of assigning importance to individual features or pixels.

This repository runs a MobileNetV2 image classifier to classify whether an image is of a male or female individual. It then uses the ACE algorithm to create concepts of images to explain why the classifier classified the images the way it did. This repository aims to investigate whether or not ACE can be used to identify bias in a machine learning algorithm.

Concept-Based Explanations on Bias Detection

While TCAV (an algorithm that also uses concepts) needed to have concepts supplied by humans to run its algorithm (Kim et al., 2018), ACE aims to automatically identify higher-level concepts that are meaningful to humans and important to a machine learning model. The original TCAV algorithm required the model user to provide concepts, for example, for a zebra, striped, dotted, or zig-zag, and then provide a score back for the importance of that concept in making a classification decision. Those human-supplied concepts themselves may be biased, making an explanation prejudiced. However, ACE generates concepts (instead of them being created and supplied by humans) by first segmenting images, then grouping similar segments as examples of the same concept, and then returning essential concepts, as scored by TCAV scores (Ghorbani et al., 2020), making it a prime tool for identifying bias in the supplied black box algorithm.

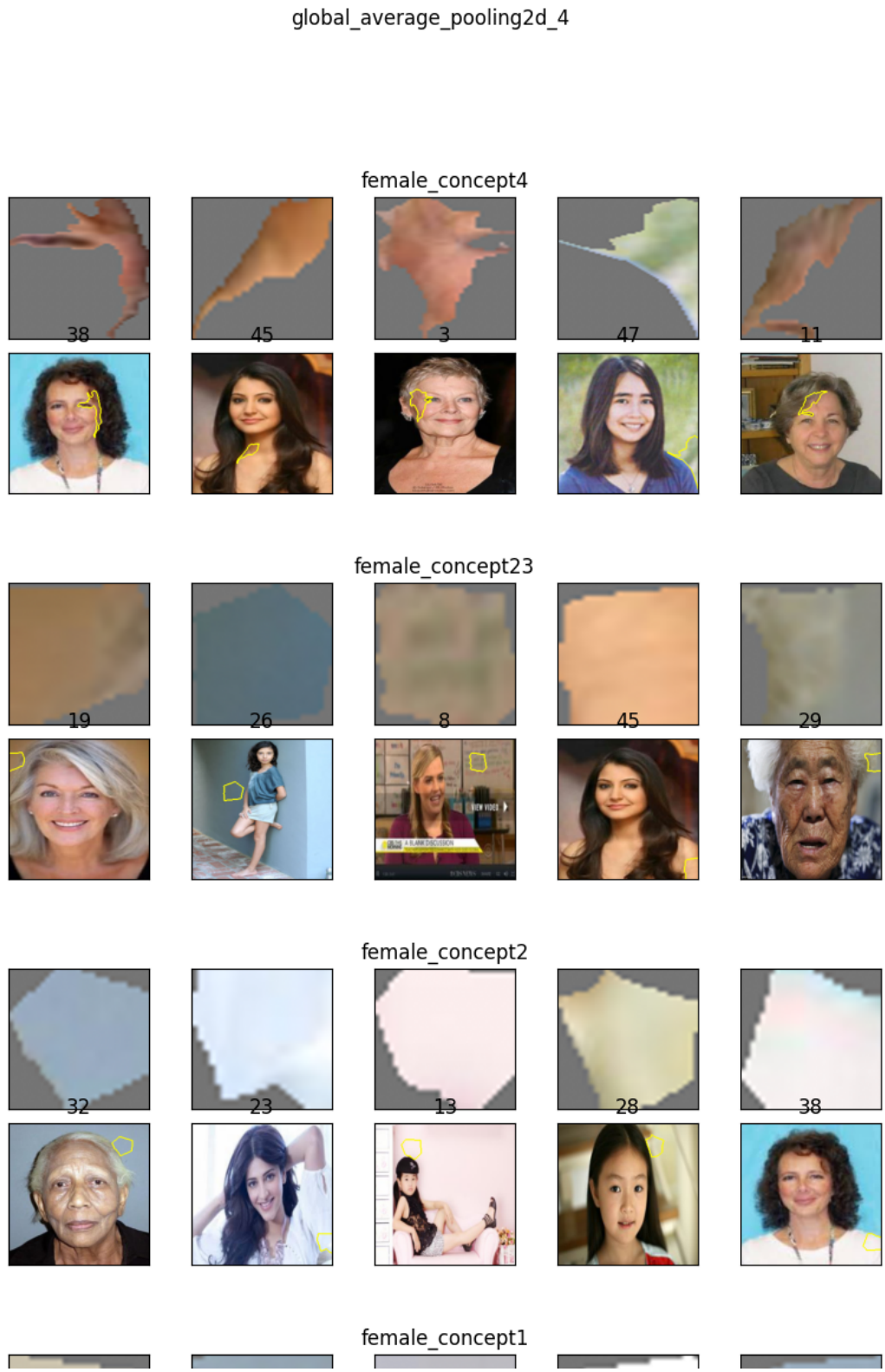
Running ACE on the UTKFaces Dataset

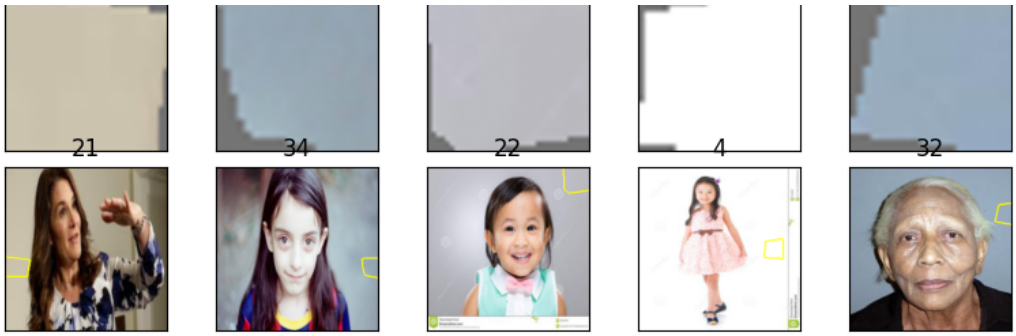
First, this project runs a MobileNetV2 image classifier to classify whether an image from the UTKFaces Part1 Dataset is of a male or female individual in the file Creating_Gender_Classifier.ipynb with a test accuracy of 0.866. Then, it generates concepts for the female and male classes. It uses the ace.py and ace_helper.py files from the original ACE project, though modified to work with TensorFlow2. In addition, it creates a new custom_wrapper.py file with a wrapper modified from the TCAV project since the GoogleNet wrapper did

not seem to work for the MobileNet2 classifier. In addition, the InceptionV3 wrapper did not work while running an InceptionV3 model.

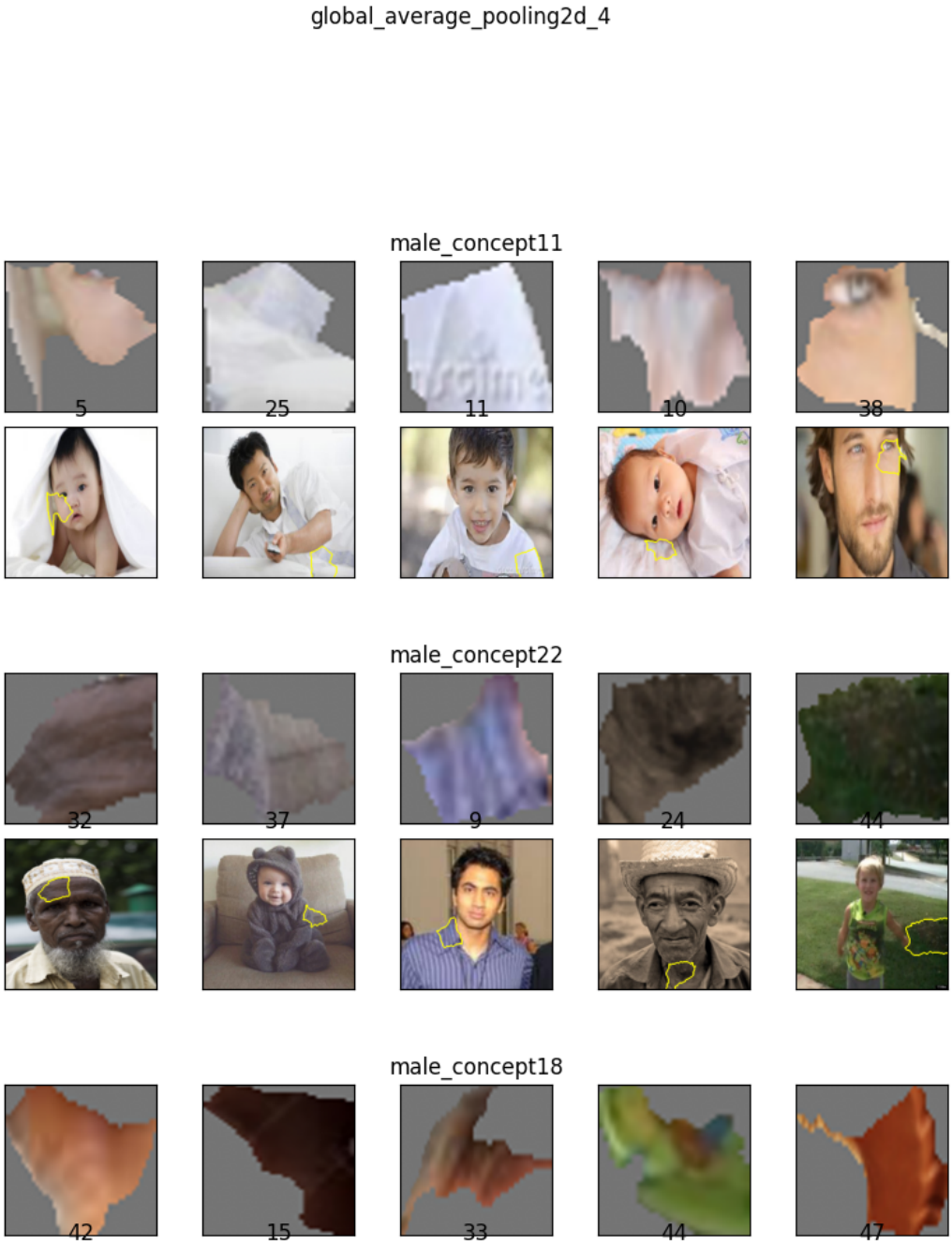
There was also uncertainty about which layer to use as a bottleneck to create activations and concepts. The code in the main branch uses the global_average_pooling2d layer, while testing was also done for the 'dense' layer (code available in a separate branch, with results saved in folders labeled with "11.5" at the end). Results appeared similar, although the global_average_pooling2d layer had higher TCAV scores, which appear below.

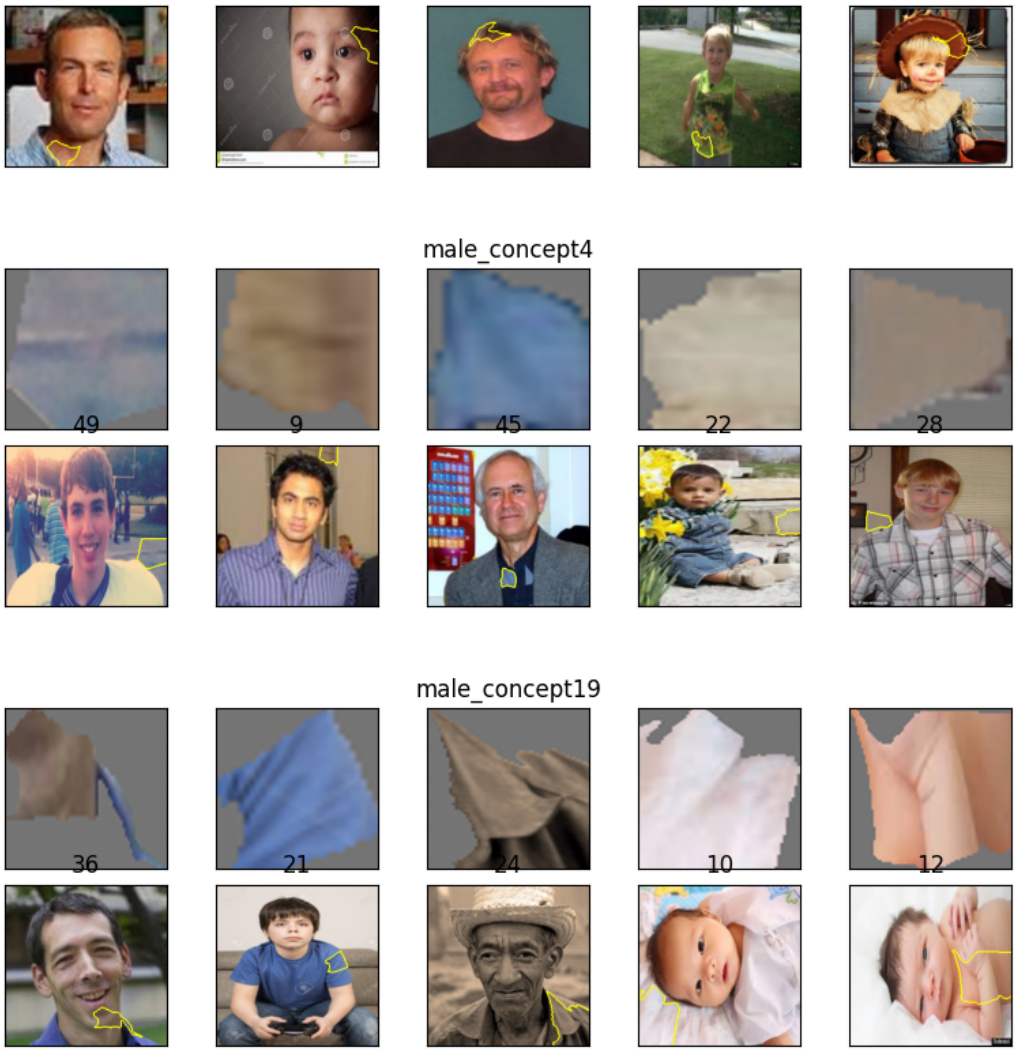
Female Concepts





Male Concepts





Assessing ACE on Bias Detection

Checking for Bias: Homogeneity in Produced Concepts

There are numerous ways in which the ACE algorithm can be used to identify biases in black box facial recognition technologies or the datasets used to train them. One of the ways it may help is to determine if a model is biased by user evaluation of how varied by age, race, and gender the concepts generated by the ACE algorithm are. As noted in Supplementary Figure 1 of “Towards Automatic Concept-based Explanations,” the ACE algorithm is designed to offer a wide range of concepts by using multiresolution segmentation. Given that the algorithm is designed to return a variety of concepts used in identification, one could use this fact to see if the black box model or underlying dataset was unbalanced (featuring mostly homogenous images) or biased.

For instance, in the female results, most women featured in the concepts are light-skinned, although of different races. Notably, any black individuals are missing. A similar issue occurred with the concepts

generated from the male class, although a few images of people of different races are present here.

Checking for Bias: Unuitivity Correlated Concepts

As a result of ACE providing its concepts instead of user-generated ones, ACE often produces ideas that are “unintuitively correlated” to the class one seeks to identify, as noted in Figure 5: Insights. In this figure, the authors provide examples such as “dumbbell” and “hand,” “Jinrikisha” and “Human,” and “Train” and “Pavement” as concepts used to explain how models made their classifications or decisions (Ghorbani et al., 2020). While these correlations referenced in the paper are seemingly harmless, the idea of looking for “unintuitive correlated” concepts could be used to identify potential underlying biases.

While not wholly unintuitive, the female class pulled many pastel-colored concepts from backgrounds and clothing in concepts 1 and 2, which may make one think the model is a bit biased in its pairing of females with pinks, purples, and pastel colors. Interestingly, the male concepts tended to be much more texture-based, but that requires more investigation.

Checking for Bias: Continuous Monitoring

Lastly, one key factor in mitigating bias in facial recognition technologies is to make sure that explanation models continue to serve everyone regardless of their social status: “continuous monitoring is necessary because all systems change over time thus there should always exist a mechanism of frequently scanning the FRT system using XAI methods which will ensure that these mitigation measures remain effective for longer periods without allowing re-emergence of biases again” (B et al., 2024). ACE is a good candidate for an explainable AI algorithm to monitor facial recognition technologies continuously. For example, ACE is relatively easy to run and re-run on various models and can be applied to many convolutional neural networks, deep neural networks, and beyond, including popular frameworks like InceptionV3 and GoogleNet. The ACE algorithm is simple, straightforward, and easily adaptable: a user needs to supply the black box model used to run the initial classification and the data used to train the model. This ease of use and replicability makes ACE’s explanation models easy for continuous monitoring of black box models as time passes, models evolve, and new training data can be added.

While the ACE models here require more fine-tuning to generate more valuable concepts, once that fine-tuning is done, the ACE algorithm is quick and easy to run—each ran in under 5 minutes on my computer and required minimal updating to run again. This makes ACE a prime candidate for continuous monitoring as long as some of its issues are ironed out.

Drawbacks: Meaningless or Incoherent Concepts

ACE, however, is not without its limitations, and those limitations may hinder bias detection and mitigation. One major drawback of ACE “is its susceptibility for returning either meaningless or non-coherent concepts due to the segmentation errors, clustering errors, or errors of the similarity metric.” (Ghorbani et al., 2020). There does not seem to be a lot of documentation on using ACE to detect bias in facial recognition black box models.

Many of the concepts in both the male and female classes are meaningless. While female_concept4 seems relatively coherent, one example pulls from the background image. A similar issue exists with female_concept23. Some concept examples pull from the background, while others pull from hair and skin.

Similar issues exist with the male concepts. Some concept examples, such as male_concept11, male concept22, and male concept18, draw from the skin, while others are from fabric.

Drawbacks: Similar Concepts

Another drawback cited by the authors is “the possibility of returning several concepts that are subjectively duplicates. For example, in Supp. Fig. 2(b), three different ocean surfaces (wavy, calm, and shiny) are discovered separately, and all of them have similarly high TCAV scores.” This results in more concepts being generated with high scores, leaving the user of the explainable model to weed through many concepts. This potentially makes it less explainable and less accessible if a user has to weed through many concepts similar to the human eye.

This did occur within the female concepts generated -- both female_concept1 and female_concept2 seem to mostly be concepts of pastel colors with little texture (and have a matching TCAV score), leading to more work on the human side to look for biases.

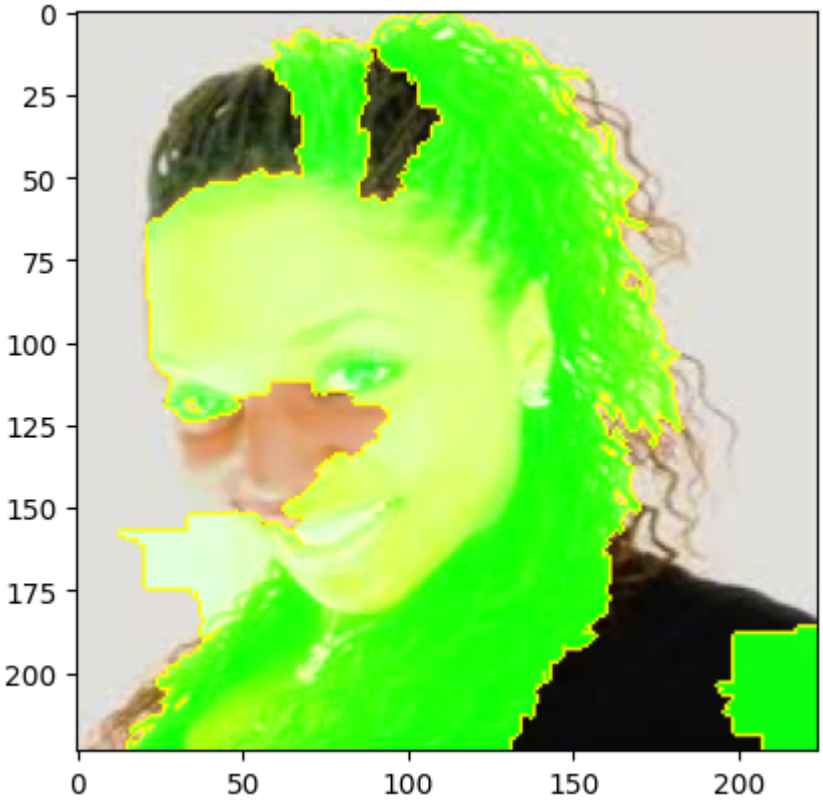
Comparison to Other Methods

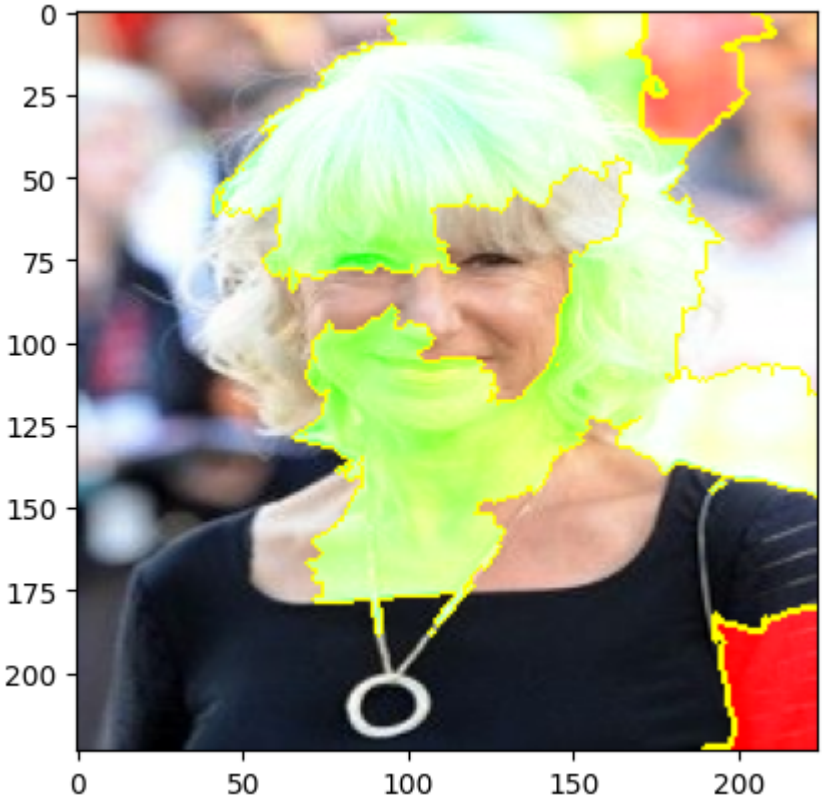
LIME and SHAP

While LIME and SHAP can show the features and pixels of an image that are used in classifying it (see the image below), they must be run on a single image at a time. Therefore, it would require a significant amount of human time to evaluate whether the explanations were biased over various images and a lot of computational usage to run each.

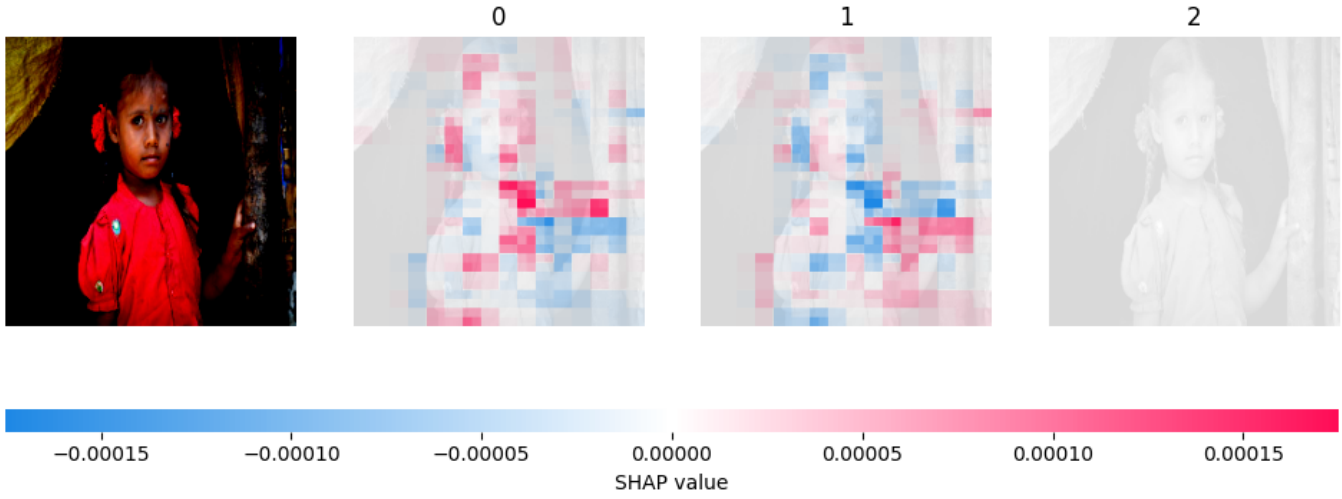
In addition, LIME identifies large swaths of the image, making it hard to tell what features are used. SHAP is difficult to quickly and intuitively identify which features are used based on how it identifies pixels, making them poor candidates for bias detection.

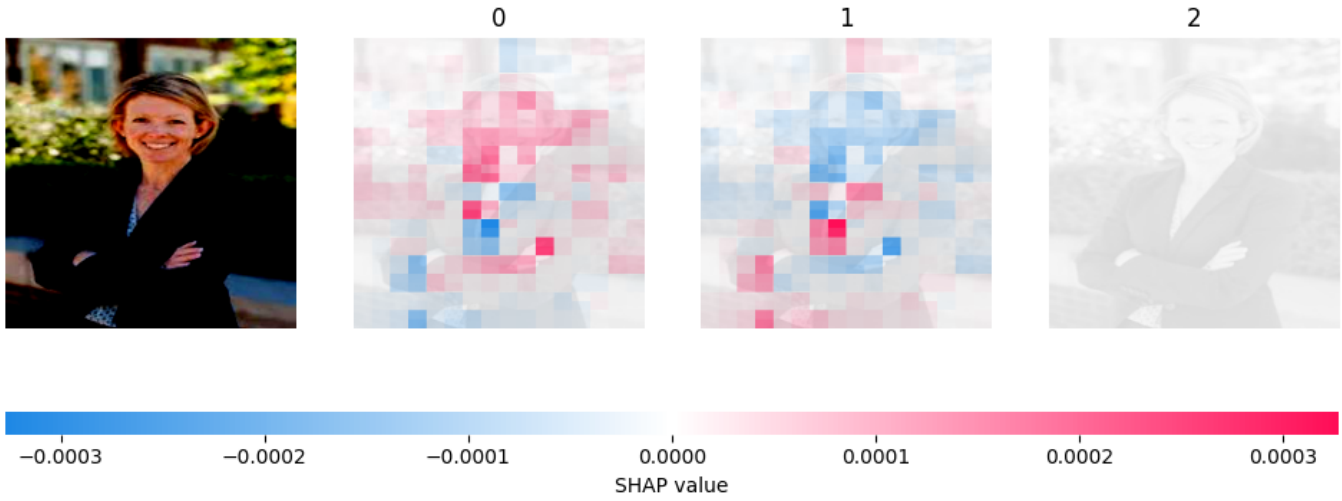
LIME Examples:





SHAP Examples:





Conclusion

In conclusion, ACE has a high potential for being a very effective explainable AI method for identifying and mitigating biases. Although it did not necessarily work well for this image dataset, one can hope that with further tuning and more training data (as well as a more diverse dataset), it might have some success.

Further Experiments

Further experimentation should be done by increasing diversity in these image datasets and re-running to look for bias. In addition, one could try running the dataset on a more cropped, face-based dataset to see if better concepts for bias detection are created.

In addition, to make concepts even more human-friendly, an even easier-to-understand concept creator, like CRAFT, could be explored.