Machine Learning for Facial Analysis

Ting-Yang Hung [a] Nicole Lee [b] Sudiksha Sarvepalli [c]
[a] University of California, San Diego
[b] University of California, San Diego
[c] University of California, San Diego

Abstract

Due to the burgeoning of machine learning and artificial intelligence technology, it may feel as though there are eyes perpetually watching us. It is undeniable that, whether it is through surveillance cameras, phones, or desktops, we are always exposed to being analyzed by merely living our everyday lives. The most frightening part about this phenomenon is that most people are unaware of what is actually being seen and how. As our society begins to yield more responsibility and credibility to image analysis and other machine learning software, it is important to educate the public about them. Our interactive web application conducts facial analysis and utilizes explainable artificial intelligence (XAI) to aid in communicating the inner-workings of the machine learning "black box." In addition, we discuss the importance of model fairness, role of XAI in ensuring fairness, and potential discriminatory practices that stem from the imprudent use of machine learning.

Keywords: Deep Learning, Convolutional Neural Network, Grad-CAM, Integrated-Gradient

1 Introduction

The problems that we are investigating focus around how we can detect model bias in a race classification model. It can be difficult for users to interpret a complex model's results especially when it seems to be generating biased or unethical results. Therefore, we want to explore how we can apply explainable AI techniques to generate visual explanations to interpret the performance of a race classification model. We will need to research how to use neural network models and activation map algorithms to produce these visual explanations.

This problem is particularly interesting because it addresses challenging concepts related to image classification through techniques cited in the explainable AI field for detecting model bias. We want to be able to generate interpetable and detailed visual explanations that explain what specific features a race classification model focuses on when making its

predictions. It is especially challenging to localize salient facial features doing so requires distinguishing small details to capture from an input facial image. Therefore, we will use class activation map algorithms to show the salient features models highlight when making their predictions and how it is able to generalize to new faces. It would also be really useful to implement visual question answering to be able to ask the model specific questions about characteristics such as the race/ethnicity, gender, and age of the person to observe if the model's answers are appropriate and accurate.

Our eventual goal is to have a better understanding of why the model is making its predictions, in hopes that it will establish better trust between users and the model and also show how the model might need to be improved. This can be accomplished by developing a tool that takes in an input image and generates activation heatmaps and can also generate confidence scores for its answers to visual question answering tasks to measure the success of the model's performance.

2 Dataset

Our project uses the FairFace dataset to perform classification and analysis. FairFace supplies 108501 images of faces from an equally distributed pool of seven race categories, two genders, and nine age groups. Figure 1 displays some samples from the FairFace dataset and Figure 2 shows the distribution of race for FairFace dataset. In addition to being uniquely comprehensive and applicable to our project, the size of this dataset allows us to create subsets of the data in order to display biased training sets. The biased dataset was generated based on the actual US population as recorded in the 2019 US Census dataset: White: 60%, Black: 13%, Latino Hispanics: 18.5%, East Asian: 2.2%, Southeast Asian: 2.2%, Indian: 1.2%, Middle Eastern: 2.4%.

Figure 1: image sampels from the FairFace Dataset

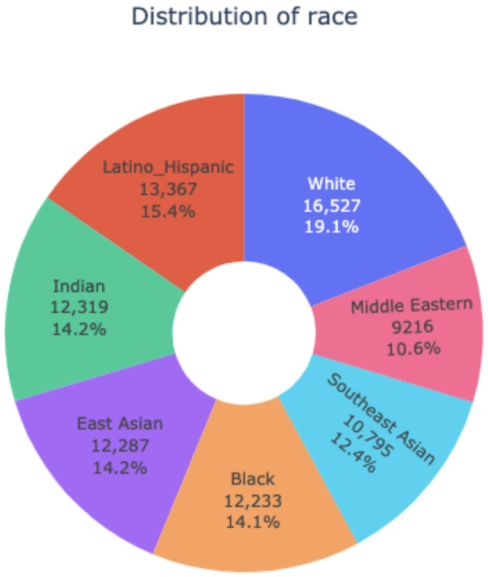

Figure 2: The distribution of race for FairFace Dataset

3 Methods

3.1 *Model*

We used Convolutional Neural Network (CNN) as our model. We applied transfer learning with resnet50 by taking the first 14 layers and fixing their weights. Then, we concatenated them with our self-defined layers. We preprocessed the training images by resizing it to 224 x 224 x 3 and applied the resnet50 preprocessing function from Keras and adding random rotation, horizontal flip, and vertical flip.

The parameters for training are listed as follows: batch size = 128, learning rate =

0.008, optimizer = Nadam, loss = categorical cross entropy. The learning rate is halved if the validation loss does not decrease for 10 consecutive epochs and early stopping would be triggered if the validation loss does not decrease for 30 consecutive epochs.

We trained a total of 4 models for different classes: age, race, and gender and one biased model for race. All the models are trained with the same model architecture defined above except the last output layer is being adjusted by the number of categories each class has. We achieved 66% accuracy on race, 55% on age, 91% on gender, and 38% accuracy on biased race models. The training accuracy and loss curves are displayed in the Appendix(Figure 11-18).

3.2 *Explainable AI (XAI)*

XAI techniques can be implemented to help with model interpretability by visually showing what parts of our input face images our custom trained model is focusing on when making its predictions/classifications. We used the Grad-CAM and Integrated Gradients algorithms to generate heatmaps that are class discriminative but have coarse localization. We used an implementation of Grad-CAM in Keras to take in our custom trained models and to generate heat maps given an input facial image from the FairFace dataset. This implementation is compatible with Tensorflow and Keras 2.0, and this architecture is applicable to any CNN model architecture.

3.2.1 *Grad-CAM*

The Grad-CAM algorithm focuses on the feature maps from the final convolutional layer in the neural network since this last convolutional layer would store the most detailed spatial and semantic information about the features in the input image. Then, the feature maps produced from this layer are fed into the fully connected layers which add weights to the features to then get the probabilities for each class. The class with the highest probability is chosen as the classification/prediction y for the input image. The calculation steps for the Grad-CAM is the following:

1. Compute the gradient of the prediction y (raw score) with respect to the feature maps generated from the final convolutional layer.

2. The feature maps from the final convolutional layer are weighted using "alpha values" which are calculated by averaging the gradients using Global Average Pooling. These weights represent the importance weight of a feature map k to the target class c.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

3. Calculate the Grad-CAM heatmap by calculating the weighted combination of the feature maps with its weights. The ReLU function is applied to put more importance on the positive values and replace the negative values with 0.

$$L^c_{\text{Grad-CAM}} = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

Then, the heatmap resulting from this Grad-CAM procedure needs to be resized to match the dimensions of the input image so that it can overlay on top of it to return the final visualization. Figure 3 shows the general process of applying Grad-CAM to a CNN architecture. Figure 4 shows how the Grad-CAM algorithm can be applied to a that is trained specifically for racial classification.

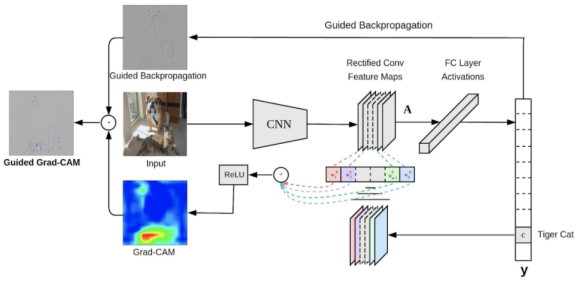

Figure 3: The general steps of performing Grad-CAM

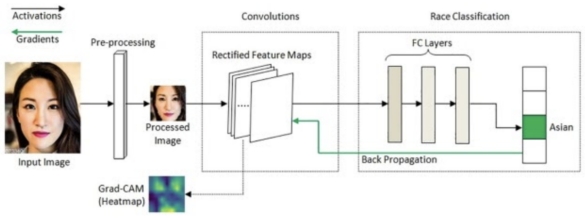

Figure 4: The illustration of how Grad-CAM can be applied to racial classification. The network chooses the best classification label from 4 different races but in our specific application we have 7 different racial classifications

3.2.2 *Integrated-Gradient*

Integrated-Gradient (IG) explains the relationship between the predictions and the learned features. Unlike Grad-CAM that only looks at the final convolutional layer from the CNN model, IG takes the entire model into account. IG requires a baseline, such as black or white background, and a set of interpolated images from a given input image. The gradient maps for each interpolated image are being calculated. For example, if an input image has size 224x224x3, the gradient map will also have the same size of the image because gradient is calculated by taking the derivative of a particular output channel w.r.t a pixel. We do this for every pixel, hence produce the gradient map that has the same size as the input image. Then, we take the average of the gradient maps for each interpolated image and multiply by a scaling factor to produce the heatmap. The value of the heatmap is simply the gradient of each pixel. This heatmap displays a decent face localization resolution for the input image. Figure 5 illustrates the steps of performing IG thoroughly.

This equation summarizes integrated gradient:

$$IntegratedGrad_i^{approx}(x) = (x_i - x_i^{'}) \sum_{k=1}^{m} \frac{\partial F(x^{'} + \frac{k}{m} \times (x - x^{'}))}{\partial x_i}$$

where x_i = input image, $x_i^{'}$ = baseline, m = total number of interpolated images, F = output channel

The general procedure of IG is as follows:

1. Determine m. The common value of m is ≥ 20 in practice
2. Generate Interpolated images = $x^{'} + \frac{k}{m} \times (x - x^{'})$
3. Compute gradient between model F output predictions w.r.t features= $\frac{\partial interpolatedpathinput}{\partial x_i}$

4. Integral approximation through averaging gradients= $\sum_{k=1}^{m} gradients \times \frac{1}{m}$

5. Scale integrated gradients w.r.t input image= $(x_i - x^{'}) \times IntegratedGradients$. The reason this step is necessary is to make sure that the attribution values accumulated across multiple interpolated images are in the same units and faithfully represent the pixel importance on the input image

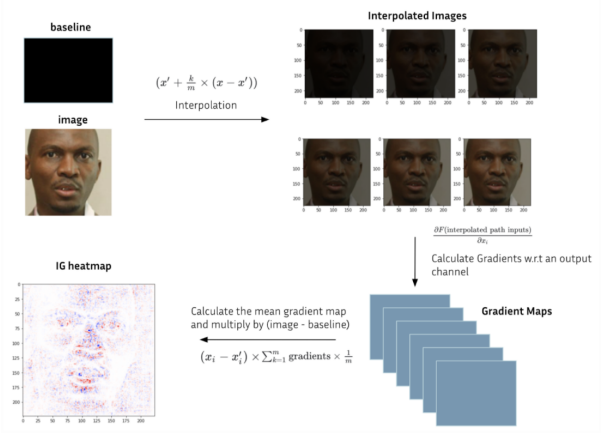

Figure 5: Diagram to show the steps of IG

4 Results

We displayed the heatmaps generated with Grad-CAM and IG for both the fair race model and the biased race model. We selected four samples that include Indian, White, East Asian, and Black people. Figures 6-9 show the results of our works. The first row contains the Grad-CAM results, the second row includes the Guided-Grad-CAM results, and the third row contains The IG results. We will not discuss Guided-Grad-CAM because it is just an alternative visualization of Grad-CAM.

Figure 6 is an image of a young Indian girl. The fair model predicted the race correctly as Indian, but the biased model predicted Latino Hispanic. In this example, the Grad-CAM results for the fair model show a strong focus on the eye region, and the biased model covers a similar region, but the activation is not as strong. The IG result for the fair model shows a robust face localization, and the biased model does not show apparent features captured

by the model. The Indian race was underrepresented in the biased dataset, which can be depicted by the biased and unbiased models' performance after applying Grad-CAM and IG.

Figure 7 is an image is of a White lady. The fair model predicted White and shows that the model made its prediction by focusing on the region around the eye. The biased model that also predicted White shows a slight amount of activation in the same eye region. These results depict that the fair model was stronger in this case since it had more robust activation for the highlighted features than the biased model, which seems to be the weaker model. The biased model appears to have weaker activation since in the biased dataset White is the over-represented race and therefore could be assumed as a default prediction which led to the model not picking out specific salient features to make its classification. The same reasoning applies to IG. The heatmap generated by the fair model shows stronger activation depicted by the pixel's intensity, but the heatmap generated by the biased model shows clearer face localization, and the face shape is easier to identify. This example shows even when two models make the same prediction, users can use Grad-CAM and IG to distinguish the stronger and weaker model.

Figure 8 is an image of an East Asian lady. The Grad-CAM shows that the fair model focuses on the eye's inner region, whereas there are not many activations for any facial features for the biased model. This could explain why the biased model's accuracy for this race is relatively low since it did not learn any specific features to classify this image. The IG heatmaps generated by both models show decent face localization. But the heatmap generated by the fair model has more activation for the face features by showing a more apparent face pattern.

Figure 9 is an image of a Black gentleman. The fair and biased model both predicted Black, but in this example, you can see from the Grad-CAM results that the biased model seems to show better results since it shows more robust activation on the face than the fair model. This example shows that the fair model seems to not perform as well for classifying the men in the Black race even though it makes the correct prediction and the biased model had an easier time picking out features that represented the Black race. We noticed that on Black women the fair model was able to pick out facial features to focus on so this shows that maybe in the biased dataset the gender is not distributed evenly for the Black race.

On the other hand, the biased model successfully picked out features that represented the Black race. However, the IG heatmaps generated by both models display a robust face localization. This example shows that we can utilize different XAI techniques to evaluate the model. Perhaps in the fair model, the model does not show a robust object detection, but by looking at IG heatmaps, we know that the model does localize the face well. If you want to see more results, please visit our GitHub and follow the instructions to play around with our code!

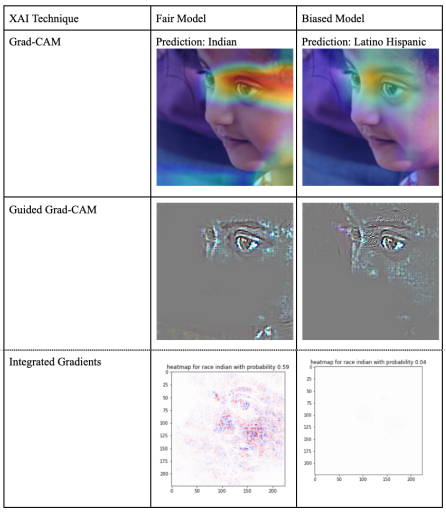

Figure 6: Indian girl

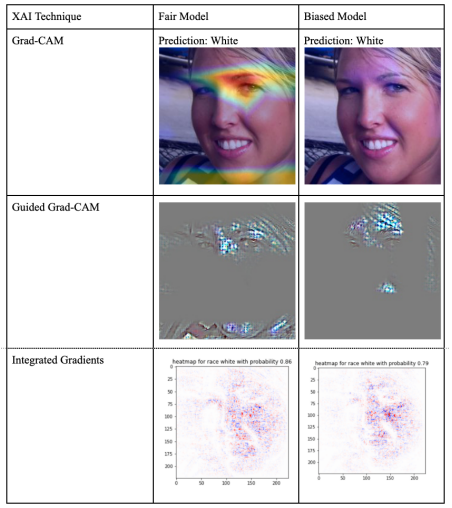

Figure 7: White lady

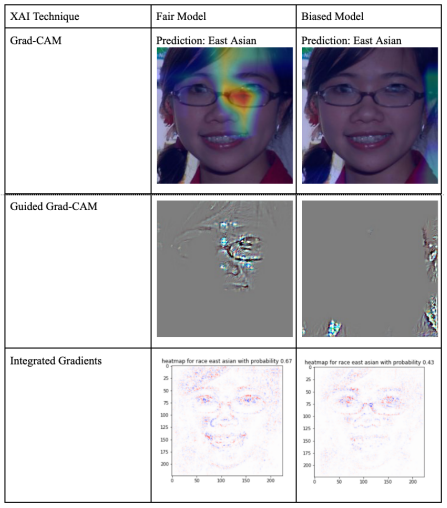

Figure 8: East Asian lady

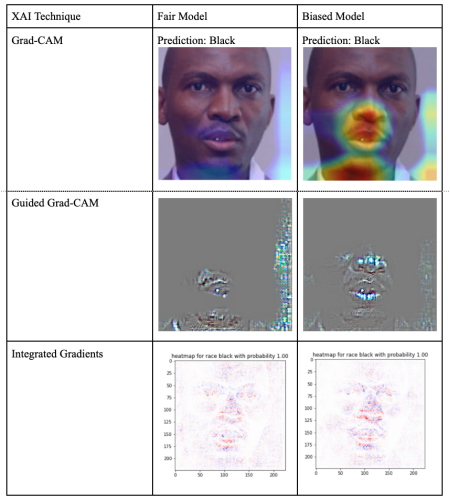

Figure 9: Black gentleman

5 Discussion

We compared the heatmaps generated with the fair race model and the biased race. The Grad-CAM visualized the important features well and the Integrated gradient visualized the face localization well. Overall, the fair race model has better heatmap representations. We are aware that we only show the four heatmaps samples. We did not find a solution to calculate the aggregate heatmap with respect to each class(e.g. the heatmap for White people). The reason is that the face from each image is located at a different place. Therefore, calculating the aggregate heatmap would not make sense. However, it is worth investigating a solution to align face from each image to the same location so that the aggregate heatmap can be calculated. Another issue that we'd encountered is the training data quality. In the FairFace dataset, some images have faces facing sideways and some images have very poor resolution. We displayed some poor images from the dataset in Figure 9. The first two rows include images with face facing sideways and the last row includes images that either have poor resolution or have multiple faces. Our models are susceptible to make wrong predictions for those images and this indirectly influences the quality of the heatmaps generated by XAI techniques. We should definitely clean the dataset so that the models are trained with high quality data and can yield higher accuracy for model predictions and better heatmaps visualization.

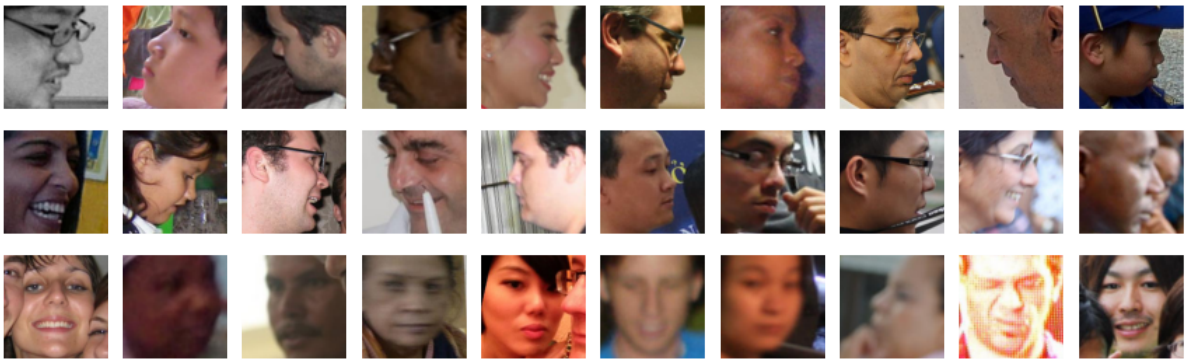

Figure 10: Bad examples from Fairface Dataset. The first two rows include images of faces that are facing sideways. The last row includes poor resolution images

6 Conclusion

Improving the model's explainability is a crucial step to understand AI. We trained CNN models and visualized the heatmaps for input images using Grad-CAM and Integrated-Gradient algorithms. We compared the heatmaps between the fair race model and the biased race model and show that the fair race model is able to capture more salient features through the heatmaps visualization. This demonstrates the importance of having a fair dataset beforehand and possible unintentional caveat to develop a biased model if the dataset is not ideal.

7 Appendix

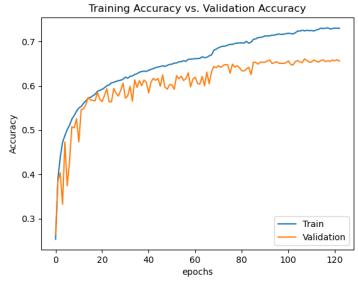

Figure 11: Race model accuracy curve

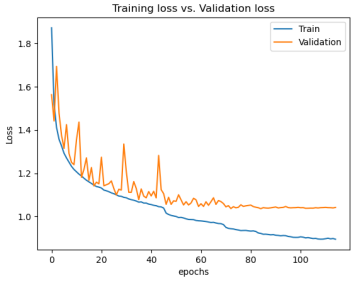

Figure 12: Race model loss curve

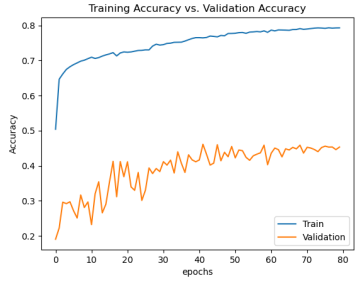

Figure 13: Biased race model accuracy curve

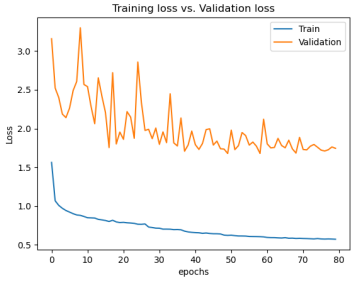

Figure 14: Biased race model loss curve

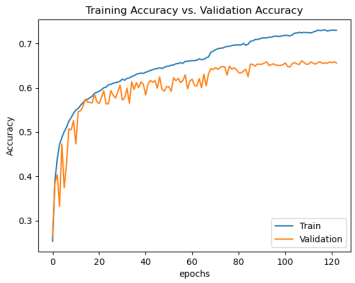

Figure 15: Age model accuracy curve

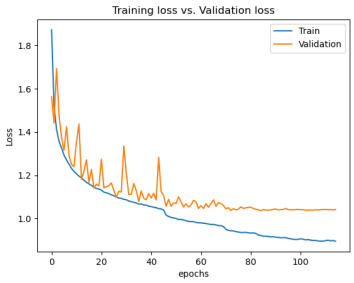

Figure 16: Age model loss curve

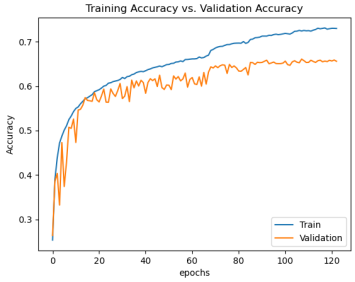

Figure 17: Gender model accuracy curve

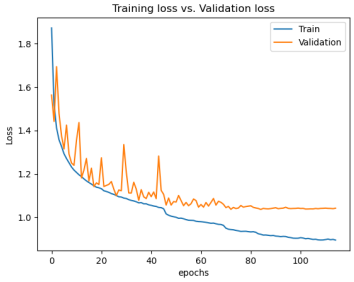

Figure 18: Gender model accuracy curve

References

[1]Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.
[2]Grad-CAM implementation in Keras[Source code]. https://github.com/jacobgil/keras-grad-cam.
[3]Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International Conference on Machine Learning. PMLR, 2017.
[4]Integrated Gradients[Source code]. https://github.com/hiranumn/IntegratedGradients.
[5]@inproceedingskarkkainenfairface, title=FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation, author=Karkkainen, Kimmo and Joo, Jungseock, booktitle=Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, year=2021, pages=1548–1558
[6] FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age[Source code]. https://github.com/dchen236/FairFace.
[7] Draelos, Rachel. "Grad-CAM: Visual Explanations from Deep Networks." Glass Box, 29 May 2020. https://glassboxmedicine.com/2020/05/29/grad-cam-visual-explanations-from-deep-networks/: :text=Grad