Investigating Interpretability of Gender Classification Models

Ashvin Pai, Martin Strauss

June 2023

Abstract

A Convolutional Neural Networks (CNN) is a type of deep learning algorithm which is often used for image classification tasks. One such image classification task is the classification of a person's gender based upon an image. These gender classification models often pose a bias towards transgender and dark skinned individuals. Such biases are compounded by the fact that large neural networks function as black boxes and lack interpretability; it is difficult to explain why a model classifies an image in a particular way. We investigate Gradient-weighted Class Activation Mapping (Grad-CAM) as a visualization tool which can explain how gender classification models make their decisions.

1 Introduction

An automatic gender recognizer is a model which takes a photo of an individual as an input and then produces a, usually binary, gender classification as an output. Large tech companies such as Microsoft, IBM, and Face++ offer gender recognition as standard on their commercial facial recognition products. Spotify has patented the technology to recognize someone's gender based on voice recordings.

According to a 2021 article by The Verge, automatic gender recognition is used in a wide variety of applications from making digital billboards which display targeted advertisements based on the viewers gender to attempting to create single-gender digital spaces such as the "girls only" social app Giggle (Vincent, 2021). Beyond direct applications, the article explains, "gender identification is used as a filter to produce outcomes that have nothing to do with gender itself." For instance, if a facial recognition algorithm uses gender as a parameter for identification, then any biases the algorithm has with respect to gender will filter into the larger task the facial recognition program is being used for.

It has been shown that these commercial gender classifiers exacerbate existing social biases. A study on the models created by Microsoft, IBM, and Face++ found that they misgender women and darker skinned people at a higher rate than men and lighter skinned people (Boulamwini and Gebru, 2018). It has also been found that when popular deep learning networks such as ResNet, Inception, and VGG are adapted to gender classification they tend to misclassify Black women the most (Krishnan et. al, 2020).

Besides race, these models inherently pose a bias towards transgender and non-binary people. It has been found that most research in the field of AGR views gender as binary, immutable, and an essential concept (Keyes, 2018). As a result most gender classification models only include two categories, man and woman, and are only trained on binary gender data. Indeed, the lack of publicly available face datasets that included labels for non-binary or transgender people was a severe limitation for our own research methodology and forced us to work within a binary gender model. At a deeper level, these models treat gender as something that is to be non-consensualy assigned to people (Keyes, 2018). Such framing goes against transgender people's relationship to their own gender as a personal determination and not something that is determined prescriptively by others. The sum of these issues means that automatic gender recognizers pose a threat of discrimination against transgender and non-binary individuals. With the increasing politicization of LGBTQ+ identities some have speculated that gender classifiers could be used in a more overtly discriminatory way by governments to, for example, limit access to public bathrooms or aid in more extreme projects to repress LGBTQ+ people in countries where such identities are deemed illegal (Vincent, 2021).

Contributing to these issues is the inherent lack of interpretability of such models. Since complex models function as a black box it is not easy to explain how they make their decisions or find where bias could be coming in. To build confidence in such systems, human participants in "trans-technology" studies have expressed that they wished devices such as heatmaps could be applied to AGR systems to show which parts of a person's face contributed to the system's recommendations (Chong et. al, 2021).

In this paper we will explore what insights a method of interpretability, specifically heatmaps, gives into how a CNN gender classification model might contribute to these social issues. We attempt to use Gradient-weighted Class Activation Mapping (Grad-CAM) as a visualization tool which can explain how gender classification models make their decisions. In section two we will describe how such visualizations might benefit transgender people. In section three we recall some technical background and outline the Grad-CAM algorithm. In section four we introduce the data we use and our model training method. In section five we present our results. In the final section we discuss our limitations and potential future results.

This paper assumes an understanding of Convolutional Neural Networks on par with an introductory undergraduate course in Machine Learning. The text *Practical Convolutional Neural Networks* by Sewak, Karim, and Pujari is a good source from which to learn the relevant background material.

2 AGR for Transgender People

Despite the challenges and discrimination that AGR might pose to transgender people, work has been done exploring whether AGR could be used for a "trans technology." Trans technology is defined as "technology that allow[s] trans users the changeability, network separation, and identity realness, along with the queer aspects of multiplicity, fluidity, and ambiguity, needed for gender transition" (Haimson et. al, 2020). A subset of such technologies use machine learning as a tool to aid in identity and gender expression for trans individuals. One such example is a health informatics tool built for transgender voice therapy (Ahmed, 2019). Another example which lent insipiration for this project was a make-up feedback system which aims to help binary transgender individuals know whether their make-up helps them "pass" as their intended gender (Chong et. al, 2021).

In both cases and in surveys of transgender people we see that consensual feedback on gender presentation or ability to freely modify gender presentation, so-called "Body Changing Laboratories", is an area where automatic gender recognition systems can potentially be used to the benefit of trans people (Haimson et. al, 2020).

Much complexity, however, still exists when using AGR as a feedback system. In the study which the makeup feedback system was developed. Transgender individuals reported that the stakes of being misgendered felt lower with a machine compared to a human. At the same time they felt that there were still many issues with recommendation from a machine that could lead to damage to mental health and an aversion to experimenting in new styles which the machine did not rate high. Core to some of these issues is the inherent lack of interpretability of such models with one participant expressing that they wished heatmaps could be applied to final make-up recommendation showing how the system made decisions (Chong et. al, 2021).

2.1 Questions on Benefits of an Interpretability System

The development of a heatmap for gender classification systems and investigating what insights it could provide to transgender people is the core question of this project. The higher level questions we pursue are: (1) What insights would a method of interpretability give into how a gender classification model contributes to these social issues? (2) Could interpretability allow us to make recommendations to people, particularly transgender people, on how they might "fool" a gender classification model?

3 Gradient-Weighted Class Activation Mapping (Grad-CAM)

For our visualization technique we decided to use the Grad-CAM algorithm. We decided to use this algorithm as it produces class-discriminative (i.e. localizes cateogries in the image) and high resolution visual explanations as to why a convolutional neural network predicts a target category. This is superior to other visualization techniques such as deconvolution or pixel-space gradient visualization which trade-off localization for higher resolution. Further, Grad-CAM is directly applicable to CNN's with fully connected layers and requires no re-training of the model (Selvaraju et. al, 2019). This makes it quite an efficient algorithm to run to produce large numbers of heatmaps.

3.1 Algorithm

Given an input image let c be the target class we are producing the heatmap for and y^c be the corresponding score given by the model. Let A^k be the k-th feature map activation of the last convolutional layer. Let Z be the total number of elements in A^k. Then the following algorithm gives $L^c_{\text{Grad}-\text{CAM}} \in \mathbb{R}^{u\times v}$ where u and v are the dimensions of the image. The heatmap depicts more weighted features in red and less weighted features in blue. For an example refer to figure 2.

Algorithm 1 Grad-CAM

```
  $a^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$
  $L^c_{\text{Grad}-\text{CAM}} = \text{ReLU}(\sum_k a^c_k A^k)$
```

4 Data and Model Training

4.1 Dataset Description

For our experiments we used the GENDER-COLOR-FERET dataset from Mivia Lab of the University of Salerno. This is a balanced subset of the COLOR-FERET dataset made between 1993 and 1996 by the US Department of Defense, adapted for gender recognition purposes. We used this dataset because it maintains a high degree of consistency with individuals having been photographed in very similar lighting and photography setups ("Face Recognition Technology (FERET)," 2017). All individuals face directly towards the camera and are visible from the top of their chest upwards. Finally, all background has been removed from the images. These features of the dataset significantly reduce any pre-processing we would have to do. In total the dataset contains 836 images.

We used the 50/50 training/testing split which was recommended by the dataset authors.

The only data processing done was resizing images from their original 512 x 768 dimensions to 256 x 384 in order to speed up training. We resized images using the PIL library's thumbnail method which resizes images using bicubic interpolation while maintaining the original aspect ratio.

A drawback of this dataset is that it is binary in terms of gender and contains no labels for non-binary or transgender individuals. The dataset further has no

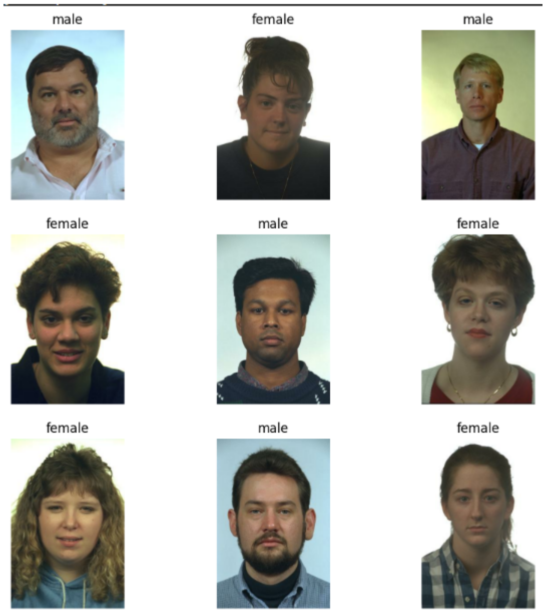

Figure 1: Example Pictures from GENDER-COLOR-FERET

labels for race but a cursory glance through the images shows that it is somewhat diverse though we do not know to what extent.

4.2 Model Architecture

For our model architecture we used the EfficientNetB0 network pre-trained on the ImageNet dataset with the top fully connected layers removed as our base model. We then added a global average layer and a final dense prediction layer to complete the model.

4.3 Transfer Learning Methodology

Our transfer learning methodology was split into a training and fine-tuning phases. In the initial training phase the base model weights were frozen and only the final dense layer was allowed to update weights. In the fine-tuning phase the top convolutional layer of the base model was unfrozen and allowed to update weights. Training for the final dense layer continued from the weights achieved in the first phase. For both training phases the training dataset was divided into a 80/20 training/validation split. Throughout all training a batch size of 32 was used.

In the initial training phase we used a learning rate of $\eta = 0.0001$ with the Adam optimizer and Binary Cross Entropy loss function. Training for 30 epochs we were able to achieve a binary accuracy of 81.82% and a loss of 0.4347 on the heldout test set.

In the fine tuning phase we used a learning rate of $\eta = 0.00001$ with the same optimizer and loss function. Training for 10 epochs we were able to achieve a binary accuracy of 88.04% and a loss of 0.3856 on the heldout test set.

Training Phase	**Binary Accuracy**	**Loss**
Initial	81.82%	0.4347
Fine-Tuning	88.04%	0.3856

Table 1: Training Results

We believe that with further data processing techniques such as augmentation and hyper-parameter tuning that the accuracy of the model could be increased. For our purposes, however, having an accuracy of 88% shows that the classifier's decision making method is not random which is sufficient.

5 Results

We apply the Grad-CAM algorithm to all images in the heldout test set (the set which our model achieves 88% accuracy on).

5.1 Misclassifying Women at a Higher Rate

The first bias to note is that the model mis-classifies women at a higher rate than men. This is a particularly interesting bias as the training and test sets are equally balanced between women and men so it is not completely obvious where such a bias may be coming from.

Gender	**Number Misclassified**	**Error Rate**
Women	33	15.79%
Men	10	4.78%

Table 2: Misclassification by Gender

One potential explanation for this bias is that when the model is classifying women it is not looking for "female traits" but rather finding a dearth of "male traits." This hypothesis is somewhat supported by the Grad-CAM heatmaps. As figure 2 shows, we find that for correctly classified women the heatmap focuses on less of their face overall, often tuning into just one area such as the nose bridge or area underneath the eye. Meanwhile for correctly classified men the heatmap is looking at multiple features such as the forehead, area under the eyes and nose, and upper clothing.

Further, to support the theory that women are being classified based on the absence of masculine traits, in figure 3 we see there are women who are classified correctly that exhibit heatmaps strongly concentrated on their collar area even

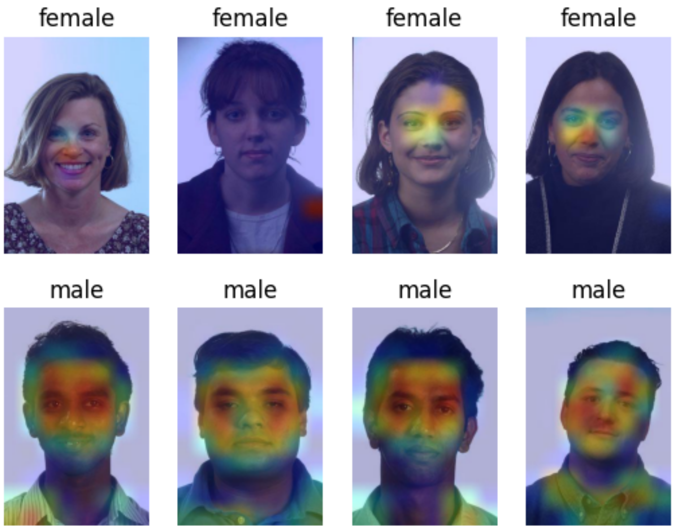

Figure 2: Examples of correctly classified men and women. The model looks at less features of women than men when making decisions.

though there is no features of note there. This would, however, be exactly the area on a man where the shirt collar or shirt button would be placed. Thus, one might conclude that they are being classified based on the absence of these attributes.

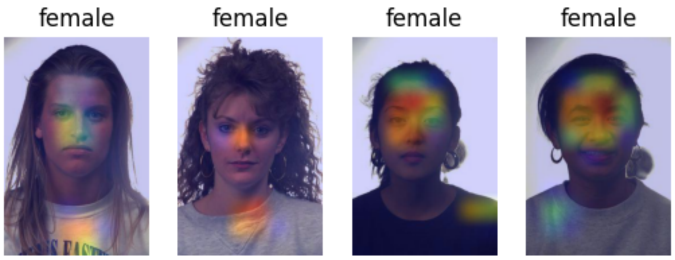

Figure 3: Examples of women correctly classified based on the absence of a shirt collar and/or button

5.2 Button Bias

The outsized importance this model seemed to be placing on shirt collars and buttons revealed an important bias in the dataset. Because our image data was collected in a formal setting participants tended to wear professional clothing. As a result most men wore collared shirts. While women wore collared shirts as well, a cursory glance at the image data shows that men wore collared shirts at a much higher proportion to women.

One way that this bias manifested in the model was the misclassification of women based on the fact that they were wearing collared shirts with buttons. We first hypothesized that the classifier had observed the fact that on women's shirts and men's shirts collar buttons appear on opposite sides. To test this hypothesis we gathered seven women who were wearing buttons that had been misclassified as men by the model. As shown in figure 4, the heatmaps for these women all have strong areas around the collar button.

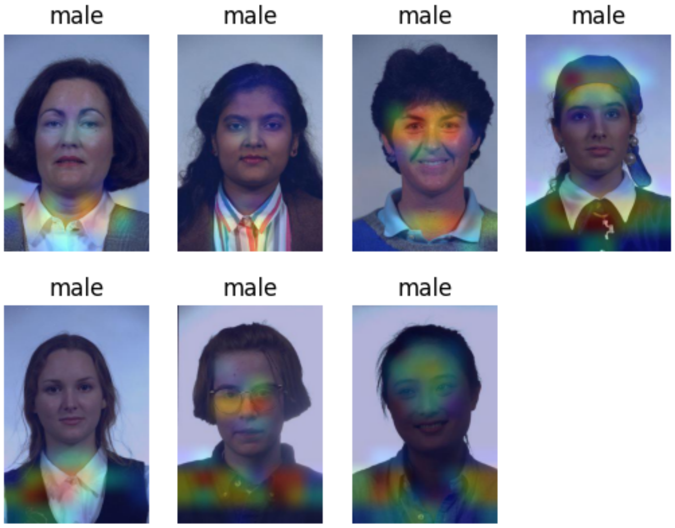

Figure 4: Women wearing buttoned shirts misclassified as men

We then flipped the images such that the buttons would appear on the opposite side of the shirt. However this resulted in all the women being misclassified once more and had no qualitative change in the heatmaps.

We then hypothesized that the presence of the button itself was contributing to the misclassification and that the position on the shirt was a moot point. To test this hypothesis we black out the collar areas of these images and run them through the model. This resulted in five out of seven women being classified correctly. Thus we conclude that the model was not indeed picking up on the difference between men's and women's shirts but might have picked up on the bias within the dataset that men wore more collared shirts than women.

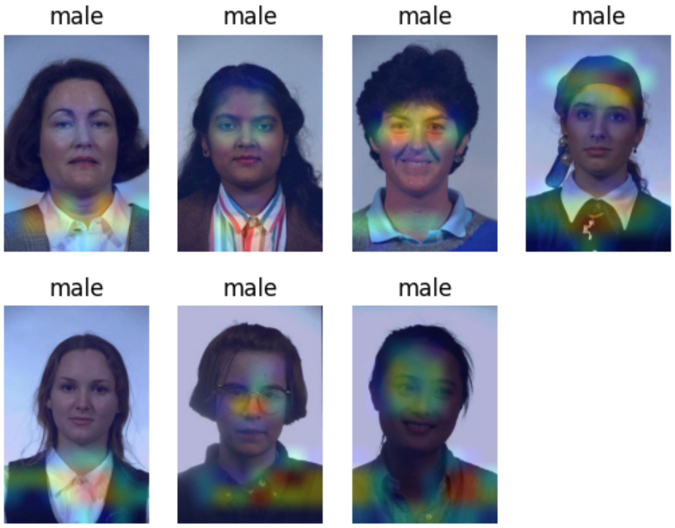

Figure 5: Flipped images from figure 4. All women are still misclassified as men.

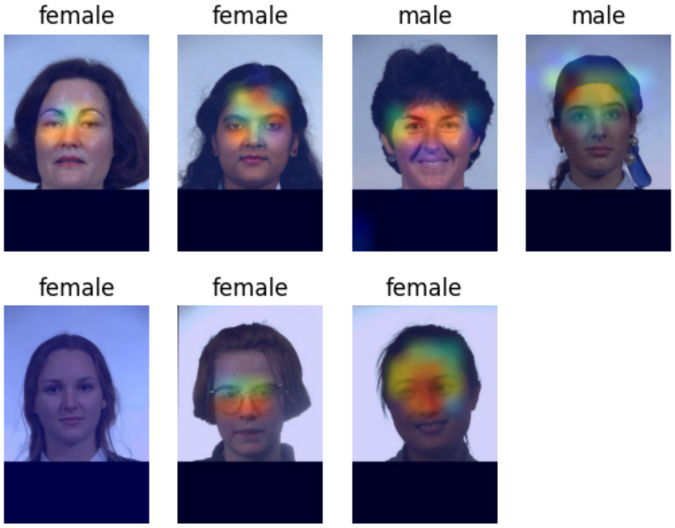

Figure 6: Blacking out the collar area results in five out of seven being classified correctly.

5.3 Potential Recommendations for Transgender People

As our experiments show, Grad-CAM as a visualization technique can help find biases within gender datasets that can then be potentially used by transgender people to “fool” an automatic gender recognition system. In the case of our GENDER-COLOR-FERET dataset we were able to use Grad-CAM to find a dataset bias as it pertained to shirt collars. As our last experiment showed, some women who wore shirt collars were classified by the system as men when the shirt collar was present and as women when it was not. This is a small relatively non-invasive intervention to clothing that a transgender person could use to come off as male to a gender recognition system.

However we have no reason to believe that this particular recommendation will generalize to a broader range of models. In fact this recommendation is contingent on the pre-processing that we did on our data-set. For example, if we had initially cropped to only the face area for all training images then this collar bias would no longer be applicable.

6 Further Research

In order to make more general recommendations to transgender people as to how they may “fool” AGR systems more visualization must be applied to a larger range of models. Of particular interest for research would be commercial models that would be present in public spaces.

We can also perform human studies to understand how interpretability methods affect how transgender people interact with or view automatic gender recognition systems.

Finally, to build on the work of Chong et. al, we can attempt to incorporate our heatmap system into a make-up recommendation system and study whether that is an effective tool to help transgender people pass in public.

Appendix A: Code Implementation of Grad-CAM

In our code we used the Grad-CAM implementation presented in the official Keras documentation (“Keras Documentation: Grad-Cam Class Activation Visualization.”).

```
def make_gradcam_heatmap(img_array,
                         model,
                         last_conv_layer_name,
                         pred_index=None):
    grad_model = Model(
        [model.inputs],
        [model.get_layer(last_conv_layer_name).output,
        model.output]
    )

    with tf.GradientTape() as tape:
        last_conv_layer_output, preds = grad_model(img_array)
        if pred_index is None:
```

```
        pred_index = tf.argmax(preds[0])
    class_channel = preds[:, pred_index]

grads = tape.gradient(class_channel, last_conv_layer_output)

pooled_grads = tf.reduce_mean(grads, axis=(0, 1, 2))

last_conv_layer_output = last_conv_layer_output[0]
heatmap = last_conv_layer_output @ pooled_grads[..., tf.newaxis]
heatmap = tf.squeeze(heatmap)

heatmap = tf.maximum(heatmap, 0) / tf.math.reduce_max(heatmap)
return heatmap.numpy()
```

Bibliography

Ahmed, Alex A. “Bridging Social Critique and Design.” Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

Boulamwini, Joy, and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” Proceedings of Machine Learning Research, vol. 81, 2018, pp. 1–15.

Chong, Toby, et al. “Exploring a Makeup Support System for Transgender Passing Based on Automatic Gender Recognition.” Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021.

“Face Recognition Technology (FERET).” NIST, 13 July 2017, www.nist.gov/programs-projects/face-recognition-technology-feret.

Haimson, Oliver L., et al. “Designing Trans Technology.” Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.

“Keras Documentation: Grad-Cam Class Activation Visualization.” Keras, keras.io/examples/vision/grad_cam/. Accessed 25 June 2023.

Keyes, Os. “The Misgendering Machines.” Proceedings of the ACM on Human-Computer Interaction, vol. 2, no. CSCW, 2018, pp. 1–22.

Krishnan, Anoop, et al. Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups, 2020.

Selvaraju, Ramprasaath R., et al. “Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization.” International Journal of Computer Vision, vol. 128, no. 2, 2019, pp. 336–359.

Sewak, Mohit, et al. Practical Convolutional Neural Networks: Implement Advanced Deep Learning Models Using Python. Packt Publishing, 2018.

Vincent, James. “Automatic Gender Recognition Tech Is Dangerous, Say Campaigners: It’s Time to Ban It.” The Verge, 14 Apr. 2021.