Towards Reducing Bias in Gender Classification

Komal Kumar Teru and Aishik Chakraborty

McGill University

Objectives

Reducing bias in Gender Classification

- Use Fader Networks to generate race invariant representations of input images
- Use such representations to classify images.
- Evaluate and compare against other models.

Motivation

The task under consideration is gender classification, which involves looking at images of faces and classifying whether the face under consideration belongs to a male or a female person. There is a lot of talk in the Machine Learning(ML) community about the biased and stereotyping ML systems that we build everyday. This type of bias stems from the fact that the datasets that we build and train these models on are imperfect and contain these stereotypes and inherent bias within them.

Recent work by [1] shows how gender classification performance of state of the art systems depend on the skin color of the subjects under consideration. The gender classifier works substantially better for lighter skin subjects. To motivate the seriousness of the situation, the authors in [1] use the example that someone can be wrongfully accused of some crime if such face recognition systems are used in practice.

The above work indicates the presence of bias in such systems, however, the authors do not provide new ways to solve the issue. In this work, we plan to use Fader Networks [2] to generate rich race-invariant image representations to help us alleviate the issue.

Fader Network Architecture

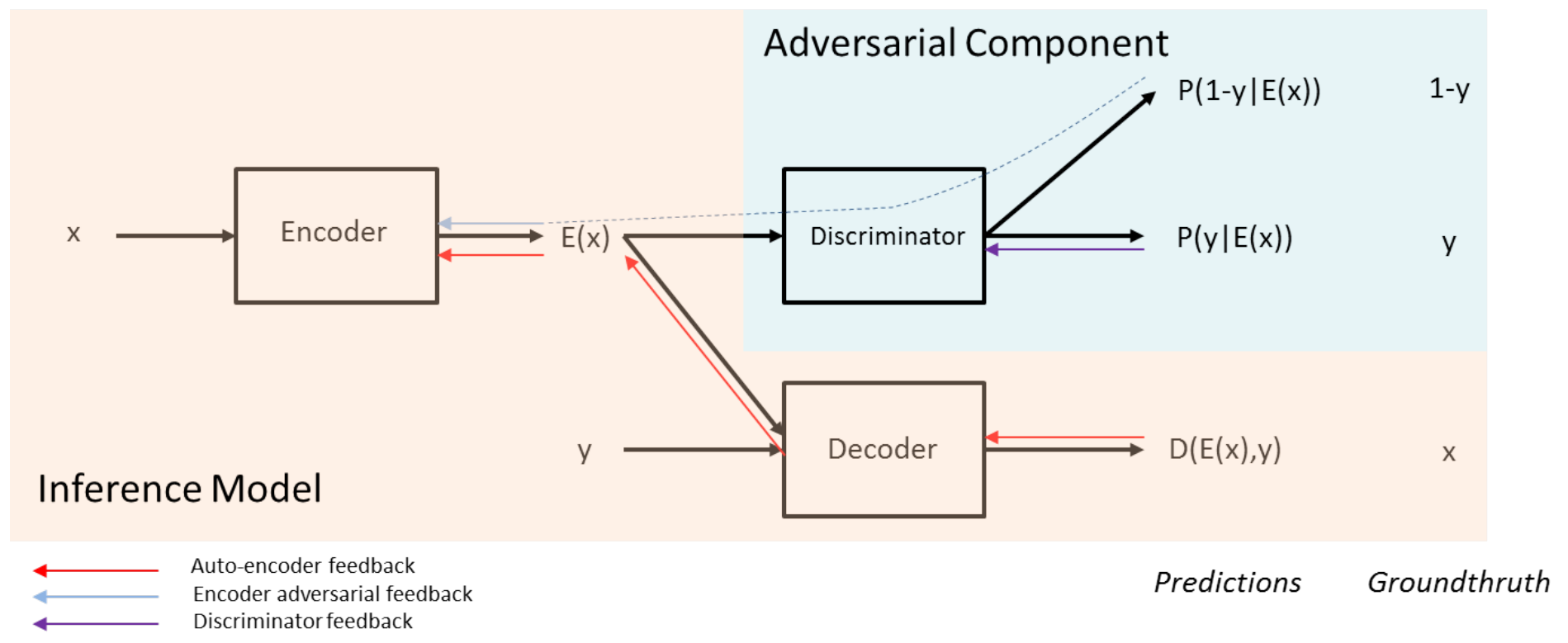

Figure 1: FaderNet Architecture. Image taken from [2]

The adversarial objective:

$$L(\theta_{enc}, \theta_{dec}|\theta_{dis}) = -\frac{1}{m} {}^{\Sigma}(x,y)\in D$$
$$\left\|D_{\theta_{dec}}(E_{\theta_{enc}(x,y)}) - x\right\|^2 - \lambda_E \log P_{\theta_{dis}}(1 - y|E_{\theta_{enc}(x)})$$

UTKFace Dataset Description

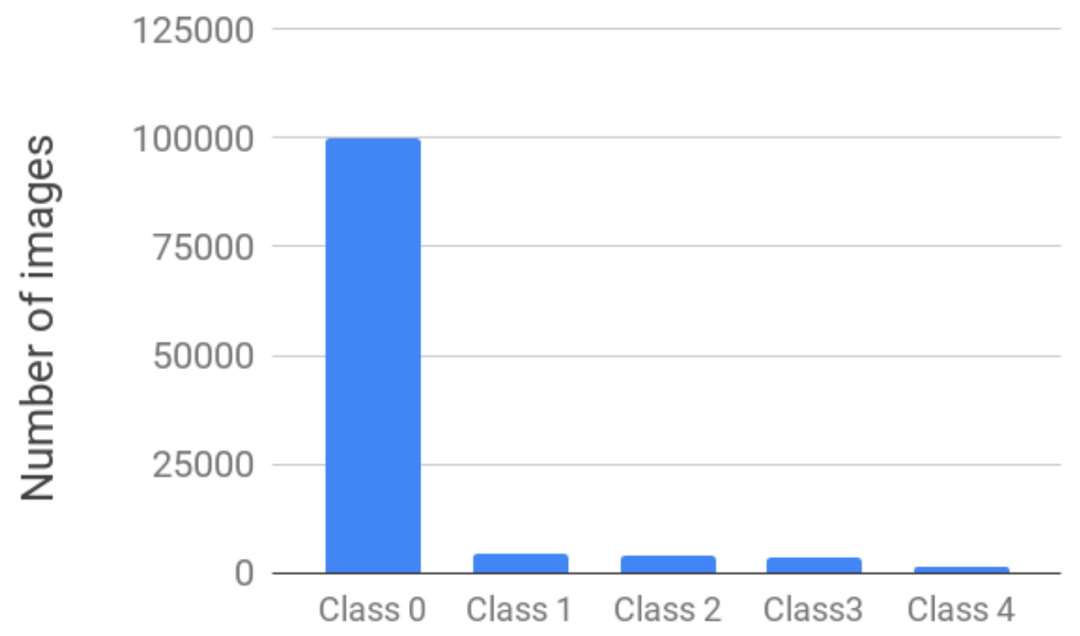

Figure 2: UTKFace Dataset Statistics

- The dataset consists of 23070 images
- The dataset annotated with age(0-116), gender(male/female) and ethnicity(White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern)) information.

Figure 3: UTKFace Examples. Image taken from `http://aicip.eecs.utk.edu/wiki/UTKFace`

Gender Classifier

We train a classifier two times with two different input image representations. To make the comparison fair, we bring the image representations to a common space.

- Representation containing full information obtained from a pre-trained vanilla Autoencoder
- Race invariant representation derived from pre-trained FaderNet

These representations are passed into a simple CNN with the following topology

- :C_{128} - C_{64} - C_{16} - C_8 - FC_2. Each convolutions layer is followed by batchnorm and relu non-linearity. We also apply dropout to the first, third, fourth conv layers and the FC layer.

Experiments

Preprocessing We re-size the images to 256×256 from the original size of 200 × 200 and extract the attributes from the file names. There were a few wrongly labelled images which we corrected manually.

Train, Test Split We first create the test set by randomly selecting 474 images from the dataset for each of the 5 races under consideration. Thus, we get a test set of size 2370 which is about 10% of the whole dataset. We also keep another 10% of the data for validation and the rest for training.

Getting Attribute Invariant Representation We run Fader Nets on our whole dataset of images with race as the input attribute to get *race invariant representations* of the image.

Gender Classification We train the gender classifier using the standard Categorical Cross Entropy Loss of the training set. We perform early stopping based on the performance on the validation set. All our evaluations have been done on the test set.

Results

Ablations	**Vanilla**	**Fader**
All classes	89.51	84.83
Class 0	92.41	85.65
Class 1	91.56 (0.85)	86.08 (-0.43)
Class 2	84.07 (**8.34**)	80.90 (**4.75**)
Class 3	90.93 (1.48)	87.66 (-2.01)
Class 4	88.61 (3.8)	83.86 (1.79)

Table 1: Classwise Evaluation Results on the test set

Conclusions

- Fader Networks help us obtain good race-invariant representations which help us mitigating bias to some extent in minority classes.
- In Fader CNN, the difference between the accuracy of the majority class and some of the minority classes dips as compared to Simple CNN.

References

[1] Buolamwini, J., and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (2018), pp. 77–91.

[2] Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems* (2017), pp. 5967–5976.