Status Report : Detecting Racial and Gender Bias in Deep Fake Detectors

Ammar Ahmad, Sherry Courington, Trivikram Ranga, Cosmin Ciausu
ammar977@bu.edu, scouring@bu.edu, tvranga@bu.edu, ccosmin@bu.edu

Image taken from link

1. Task

Our goal is to detect racial and gender bias in recent state of art deep fake detectors. We do so by running the various deep fake detectors on the FakeAVCeleb dataset which has videos of different racial and gender groups.

2. Related Work

A lot of research ([5]) has been done in finding bias in deep learning datasets but very few works have focused on detecting bias in deep fake detectors and deep fake detection datasets. Trinh et al [4] have examined bias and fairness in deepfake datasets and models.

3. Approach

We plan to take pre-trained recent state of art deep fake detectors and run them on the FakeAVCeleb dataset. We then try to analyze the difference in the performance of the detectors on different racial and gender groups. In this way, we can bias towards certain groups. We plan to extract images from the videos available in the dataset and work on those. So, we will be using deep fake detectors for images.

We surveyed the website paperswithcode.com and several research papers and the deep fake detection benchmarks and selected state of the art models that have open source implementations available. The models that we will be using are:

- XceptionNet (trained on DFDC)
- XceptionNet (trained on FaceForensics ++)
- EfficientNetB4 (trained on DFDC)
- EfficientNetB4 (trained on FaceForensics ++)
- EfficientNetB4ST (trained on DFDC)
- EfficientNetB4ST (trained on FaceForensics ++)
- EfficientNetAutoAttnB4 (trained on DFDC)
- EfficientNetAutoAttnB4(trained on FaceForensics ++)
- EfficientNetAutoAttnB4ST (trained on DFDC)
- EfficientNetAutoAttnB4ST (trained on FaceForensics ++)
- Cross Efficient Vision Transformer

We are borrowing these models from the following repositories and papers

1. N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini and S. Tubaro, "Video Face Manipulation Detection Through Ensemble of CNNs," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 5012-5019, doi: 10.1109/ICPR48806.2021.9412711. https://github.com/polimi-ispl/icpr2020dfdc
2. Davide Coccomini, Nicola Messina, Claudio Gennaro, Fabrizio Falchi. “Combining EfficientNet and Vision Transformers for Video Deepfake Detection”, arXiv.org,2021. https://github.com/davide-coccomini/Combining-EfficientNet-and-Vision-Transformers-for-Video-Deepfake-Detection

Model[1] makes use of ensemble CNN models, more precisely EfficientNetB4 CNN architectures, with an attention mechanism to see which local part of the image contributes the most to the predictions of the CNN models.

Model[2] uses an Efficient Net CNN architecture combined with vision transformers, but contrary to [1], does not use ensemble models.

For additional code, we will be writing code to evaluate these models on our dataset. Then we will be computing different metrics according to [8].

We can detect bias by analyzing the performance of the models on different groups. E.g A big drop in the performance (Accuracy and AUC) on a group indicates a bias towards that group.

For processing videos, the models that we have chosen use a different number of frames from each video. Some use only 1 frame. We are also using one single frame from the video, extracting a face from it and then passing that face to the model after all the transformations. For extracting faces, we are using BlazeFace.

After detection of the source and type of the bias, if there is any detected, we can then focus on mitigation techniques for these biases, see [8] for input, representation, and prediction quality specific methods for mitigating bias. A few worth mentioning here is adversarial training, or explicit constraints to the loss function.

4. Dataset and Metric

We are going to be using the FakeAVCeleb dataset which has over 490 real videos and over 20000 fake videos. The videos are divided into the following four categories :

1. Real video real audio
2. Real video, fake audio
3. Fake video real audio
4. Fake video, fake audio

The videos are also categorized by racial (African, Asian, Caucasian,) and gender (men, woman) groups.

Even though the dataset has videos, right now we will only be evaluating the visual content and not the audio as the models that we have selected for evaluation take images as input.

We will be discussing whether there is any discrimation present in the models that falls in one of the following categories

1. Discrimination via Input
2. Discrimination via Representation
3. Prediction Quality Disparity

To find the above mentioned discrimination, we will be evaluating the following metrics among different ethnic and racial groups

1. Accuracy
2. Precision / Recall
3. ROC/AUC curves using False Positive rate and True positive rate
4. Demographic Parity
5. Equality of opportunity

5. Preliminary Results

Uptil now we have been able to make XceptionNet and EfficientNetB4 run on our dataset after all the preprocessing steps. We are currently running these models on the complete dataset where we are using one frame from each video.

The process for extracting frames and cropped faces for method[7] still requires some additional work, especially since the implementation was fine-tuned for different structure based datasets than ours(FORENSICS/DFDC).

An early look at the deep fakes predictions for model[1] for a few race and gender specific samples can be found in the github, the link is given section 7.

As an early demo, 2 experiments were performed on African male and female images as well 2 Caucasian male and female images.

6. Detailed Timeline and Roles

Task	Deadline	Who
Test Cross Efficient Vision Transformer on FakeAVCeleb	04/22/22	Cosmin
Test XceptionNet on FakeAVCeleb	04/22/22	Trivikram
Test EfficientNetB4 on FakeAVCeleb	04/22/22	Sherry
Test EfficientNetAutoAttnB4ST on FakeAVCeleb	04/22/22	Ammar
Write code to calculate custom metrics	04/22/22	TBD
Calculate all metrics and analyze the results	04/22/22	TBD
If bias present, then find out the cause of the bias	04/28/22	TBD
Suggest bias mitigation approach	05/3/22	TBD

7. Preliminary Code

We have been using SCC and our local PCs to run our code.

Our git code can be found here, under dev branch, for frames extraction and data input preparation:
https://github.com/ammar977/Biased_Deepfake_Detectors

References

1) Hasam Khalid, Shahroz Tariq, Minha Kim, Simon S. Woo FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset, arXiv.org, 2021.
2) Davide Cozzolino, Diego Gragnaniello, Giovanni Poggi, Luisa Verdoliva. Towards Universal GAN Image Detection.arXiv.org, 2021.
3) Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, Luisa Verdoliva, Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. arXiv.org, 2021.
4) Loc Trinh, Yan Liu An Examination of Fairness of AI Models for Deep fake Detection. arXiv.org, 2021.
5) Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. arXiv.org. 2019
6) N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini and S. Tubaro, "Video Face Manipulation Detection Through Ensemble of CNNs," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 5012-5019, doi: 10.1109/ICPR48806.2021.9412711.
7) Davide Coccomini, Nicola Messina, Claudio Gennaro, Fabrizio Falchi. “Combining EfficientNet and Vision Transformers for Video Deepfake Detection”, arXiv.org,2021.
8) Mengnan Du, Fan Yang, Na Zou, Xia Hu, “Fairness in Deep Learning: A Computational Perspective”. arXiv.org 2020.