NTech Lab

Model Details

- Classifier developed by NTech-Lab as part of the DeepFake Detection Challenge [1]. July 2020, v1.
- Ensemble of 3 CNN's built on the EfficientNet-B7 [2] architecture. The two classifiers operate on individual faces in the videos while the third is predicts on a sequence of 7 frames by replacing by adding a 3-D CNN block to each block in the EfficientNet model. on the isolated faces in the frames.
- SSD [3] face detector used in training and inference.
- Key to the training procedure is the use of mixup [4]p augmentation where for each fake face a corresponding real face form the original video is selected. A linear combination of the two images is used in training. item Other key augmentations included video compression: frame rate (FPS) in the range 15-30, rescaling faces (0.2, 1.0), etc.

Intended Use

- Intended to be used to detect video manipulated using a series of Deepfake algorithms. Training data included the following algorithms: DFAE, MM/NN face swap [5], NTH [6], FSGAN [7] and StyleGAN [8].
- Not suitable for detection of fully synthetic images or Deepfake audio.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; software factors such image compression and environmental factors of lighting and humidity.
- Due to PII, evaluation is not done against any factors independently. Dataset reports, however, general distribution of the participants age, gender, Fitzpatrick skin-types and lighting conditions. These are shown in figure 1

Metrics

- Evaluation metrics include true positive rate (TPR) and true negative rates (TNR), model accuracy and the raw Cross Entropy loss used to score the competition. These values, calculated at a decision threshold of p=0.5, are shown in 2a.
- Figure 2b shows the Receiver-Operator Curve (ROC) for Deepfake detection.

Training Data

- Training data consisted of 119,154 ten second video clips containing 486 unique subjects collected by Facebook AI. Of the total number of videos 100,000 contained Deepfakes. The StyleGAN generation method was not included in the training dataset.
- Videos that had been manipulated using any of the algorithms were labeled "fake". A fraction of the videos also included audio swapping using TTS Skins voice conversion [9]. These were not labelled as fake in the competition.

Evaluation Data

- A validation set was used to compute the leaderboard during the development stage of the competition. This set consisted of 4000 ten second videos of which 50% included Deepfakes. 214 unique subjects were used in this split, with zero overlap with the training data. Additionally, the validation set included one unseen generation method for Deepfakes: StyleGAN. Augmentations (distractors, geometric and color transforms, frame rate changes, etc.) were applied to 79% of the videos.

- A private test set was used to compute the final competition scores. This set consists of 10,000 ten second clips, including 50% Deepfakes and two previously unseen augmentations:dog mask and flower crown filter. Unlike the validation set, 50% of the test set included organic content found on the internet, however due to copyright and privacy these were not released and so the metrics reported in this model card include only 5000, half real and half deepfakes.
- The model was also tested against the DeeperForensics Challenge 2020 test dataset [10] that was used to score the competition. The dataset features three appealing properties: good quality, large scale, and high diversity. The original videos were taken from the FaceFornesics++ [11] dataset, which used roughly 1000 sequences extracted from videos scrapped from YouTube. The original videos were manipulated using 100 paid actors with four typical skin tones across 26 countries. Their eight expressions (i.e., neutral, angry, happy, sad, surprise, contempt, disgust, fear) are recorded under nine lighting conditions by seven cameras at different locations. In addition, seven types of real-world perturbations at five intensity levels are applied to obtain a more challenging benchmark.

Caveats and Recommendations

- This model is intended for detection of face swaps in video only and can not be used to detect deep fake audio or still images.
- This model is not intended for the detection of traditional "cheepfake" manipulations.
- There is no assurance that this detector will generalize beyond the algorithms studies in the training, validation and tests set, including algorithms that might be developed in the future.

Quantitative Analyses

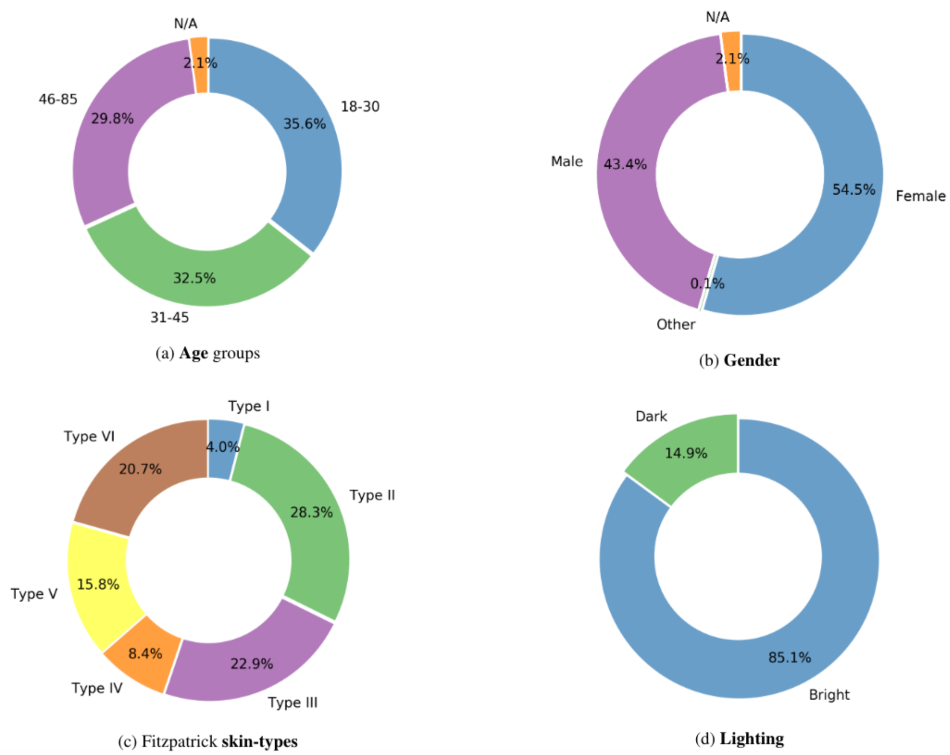
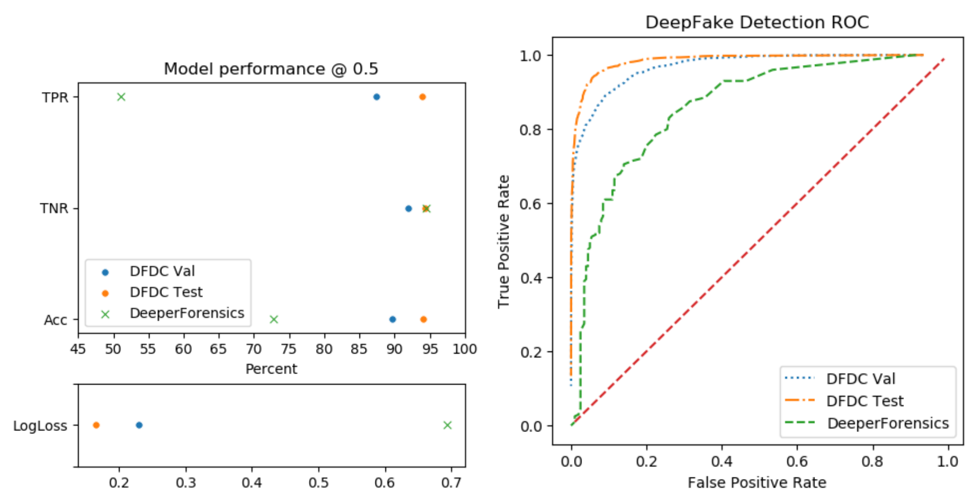


Figure 1: Subject distribution by a) Age, b) Gender, c) Fitzpatrick skin-type and d) Lighting conditions



(a) (b)

Figure 2: a) True Positive Rate(TPR), True Negative Rate(TNR), Accuracy (ACC) at a classification threshold of p=0.5 and Binary cross entropy (LogLoss) against validation and test sets. b) Receiver operator curve for DeepFake detection.

References

[1] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The deepfake detection challenge dataset,” 2020.

[2] Q. Xie, E. H. Hovy, M. Luong, and Q. V. Le, “Self-training with noisy student improves imagenet classification,” *CoRR*, vol. abs/1911.04252, 2019. [Online]. Available: http://arxiv.org/abs/1911.04252

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, “SSD: single shot multibox detector,” *CoRR*, vol. abs/1512.02325, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325

[4] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *CoRR*, vol. abs/1710.09412, 2017. [Online]. Available: http://arxiv.org/abs/1710.09412

[5] D. Huang and F. De La Torre, “Facial action transfer with personalized bilinear regression,” in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 144–158.

[6] E. Zakharov, A. Shysheya, E. Burkov, and V. S. Lempitsky, “Few-shot adversarial learning of realistic neural talking head models,” *CoRR*, vol. abs/1905.08233, 2019. [Online]. Available: http://arxiv.org/abs/1905.08233

[7] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," 2019.

[8] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4396–4405.

[9] A. Polyak, L. Wolf, and Y. Taigman, "TTS skins: Speaker conversion via ASR," *CoRR*, vol. abs/1904.08983, 2019. [Online]. Available: http://arxiv.org/abs/1904.08983

[10] L. Jiang, W. Wu, R. Li, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," *CoRR*, vol. abs/2001.03024, 2020. [Online]. Available: http://arxiv.org/abs/2001.03024

[11] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," *CoRR*, vol. abs/1901.08971, 2019. [Online]. Available: http://arxiv.org/abs/1901.08971