Model Card - Boken

Model Details

- Classifier developed by the team named "Forensics" That was the top performer in the DeeperForensics [1, 2] challenge.
- Ensemble of 3 2D CNN's. Each model uses an EfficientNet [3] architecture as an encoder.
- MTCNN face detector used in training and inference. 15 (224x224 pixel) frames used per video.
- Data augmentation was key to training the model. These included: color saturation change (CS), color contrast change(CC), local block-wise (BW), white Gaussian noise in color components (GNC), Gaussian blur (GB) and JPEG compression (JPEG).

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 [4], NTH [5], FSGAN [6] and StyleGAN [7].
- 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

- This model was trained on a number of preexisting DeepFake datasets: DeeperForensics-1.0 [1], UADFV [8], Deep Fake Detection [9], FaceForensics++ [10], Celeb-DF [11], and the preview release of the DeepFake detection dataset [12].

Evaluation Data

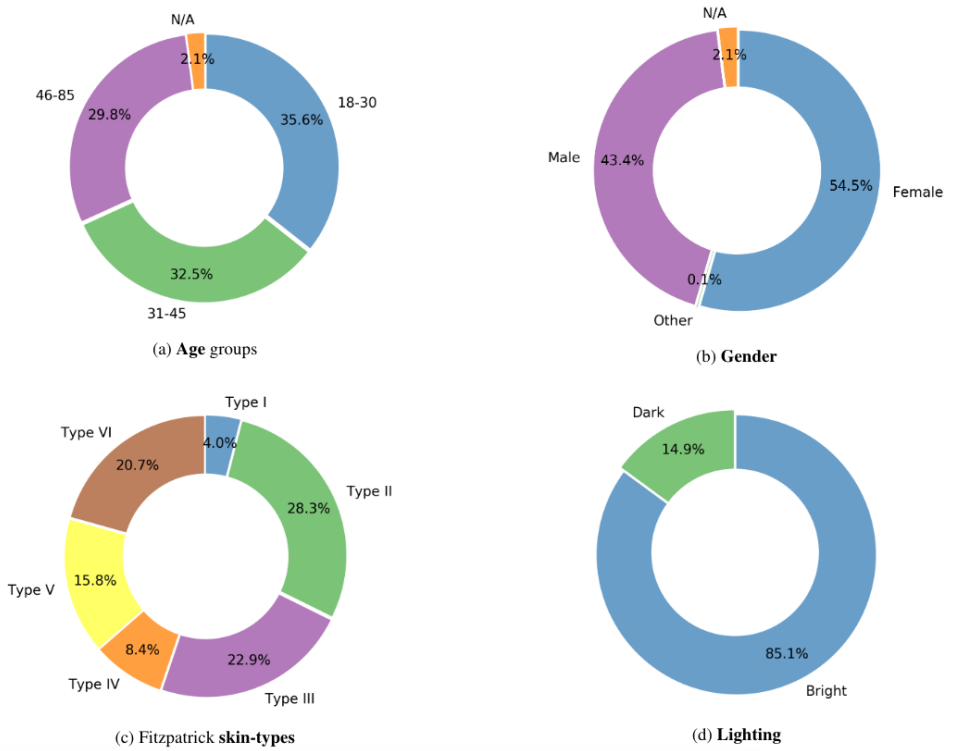
- A validation set was used to compute the leaderboard during the development stage of the DFDC competition [13]. This set consisted of 4000 ten second videos of which 50% included Deepfakes. 214 unique subjects were used in this split, with zero overlap with the training data. Additionally, the validation 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 [1] that was used to score that competition. The dataset features three appealing properties: good quality, large scale, and high diversity. The original videos were taken from the FaceFornesics++ [10] 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



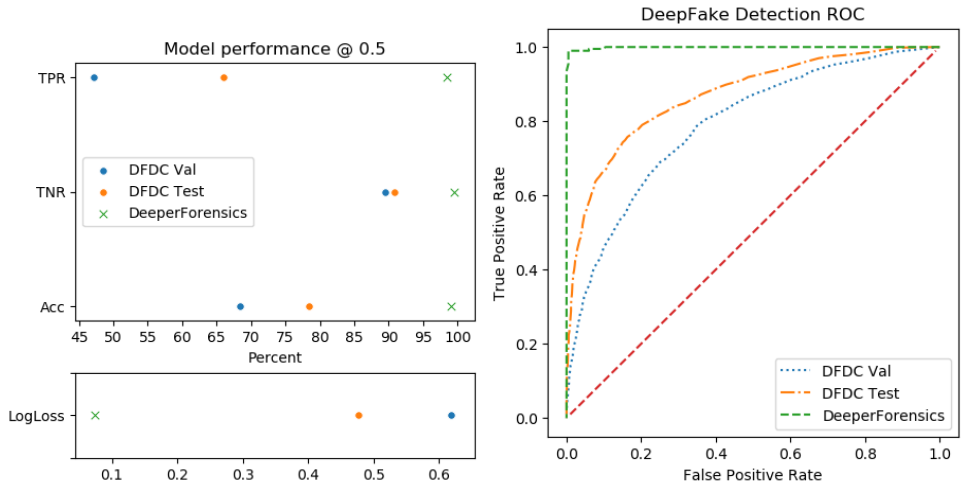
(a) **Age** groups

(b) **Gender**

(c) Fitzpatrick **skin-types**

(d) **Lighting**

Figure 1: Subject distribution by a) Age, b) Gender, c) Fitzpatrick skin-type and d) Lighting conditions for the DFDC dataset [13]



(a) (b) 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] 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

[2] L. Jiang, Z. Guo, W. Wu, Z. Liu, Z. Liu, C. C. Loy, S. Yang, Y. Xiong, W. Xia, B. Chen, P. Zhuang, S. Li, S. Chen, T. Yao, S. Ding, J. Li, F. Huang, L. Cao, R. Ji, C. Lu, and G. Tan, "Deeperforensics challenge 2020 on real-world face forgery detection: Methods and results," 2021.

[3] 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

[4] 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.

[5] 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

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

[7] 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.

[8] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8261–8265.

[9] N. Dufour and A. Gully, "Contributing data to deepfake detection research," *Google AI Blog*, vol. 1, no. 2, p. 3, 2019.

[10] 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

[11] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3207–3216.

[12] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019.

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