Research Proposal

Limin Deng
College of Engineering & Computer Science
the Australian National University
Canberra ACT 2600 Australia
u6849956@anu.edu.au

Abstract

I intend to participate in Deepfake Detection Challenge as my research project. This competition was released on Kaggle on December 11, 2019 and its submission ends on March 31, 2020. It aims to identify fake videos with face or voice swap (known as deepfake) with total prizes of $1,000,000.
Detecting deepfakes has attracted researchers' attention all over the world due to its significance in protecting social security and wide applications. For example, it ensures the authenticity of contents on social media, helps police and insurance identify the authenticity of the video, and defends human rights.
My current rank is 36% on Kaggle (14th Dec 2019). I intend to 1) label sample videos at first, 2) generate fake images by pure linear transformation or deep neural network, 3) train network on fake and real images on binary classifiers. 4) train sample training videos. 5) if the model and weight are acceptable, then train real training dataset on a large scale.
Source code: empty
keywords: deepfake, faceswap, voiceswap, forgery detection

1 Introduction

Information security becomes significant in the artificial intelligence age. As the algorithms invented by human become more and more powerful, activities in copying, transporting, and faking information becomes more and more frequent. Face and voice, as the most distinguishable features of a human, are used in many applications, such as face authentication on the smartphone or in bank and voice identity validation on the tax hotline. Thus the security issue of faking faces and voices becomes a big concern for society, with the potential to destroy the whole security system.
For now, the methods of reconstructing information mainly evolve in three lines. They are deep neural networks including Generative Adversarial Network(GANs) [1] and Variational Encoder(VAE) [2], pulse neural networks based on biology, and quantitative calculations. Methods to manipulate face include *faceswap* and *face reenactment*. Research on voice synthesize includes *Tacotron series*[1]. For this competition, my research will focus on faceswap and voiceswap.

Datasets. The official datasets[2] include 400 sample training videos, full training videos of 470 GB, and 400 test videos as the validation set. Each video is equal in length(10 seconds) but not equal in size. Other datasets in face detection include *VoxCeleb1* [3], *VoxCeleb2*[3], *FaceForensics++*[4], *CelebA*[5], *WilderFace*[4], and *CNNFacePoint*[5]

[1] https://google.github.io/tacotron/

[2] https://deepfakedetectionchallenge.ai

[3] http://www.robots.ox.ac.uk/ vgg/data/voxceleb/

[4] http://shuoyang1213.me/WIDERFACE/

[5] http://mmlab.ie.cuhk.edu.hk/archive/CNN_FacePoint.htm

Metrics. The official metric for the competition is log loss

$$LogLoss = -\frac{1}{n}\sum_{n=1}^{n}[y_i log(\hat{y_i}) + (1 - y_i) log(1 - \hat{y_i})] \tag{1}$$

where 1) n is the number of videos being predicted. 2) $\hat{y_i}$the predicted probability of the video being FAKE. 3) y_i is 1 if the video is FAKE, 0 if REAL. 4) log() is the natural (base e) logarithm.

2 Related Works

Neural network develops very rapidly. From famous backpropagation (BP)[6] in 1988 to the convolutional neural network (CNN) based on LeNet[7] in computer vision and recurrent neural network (RNN) in natural language processing. Diverse variants emerged these years.

The Generative Adversarial Network (GAN) proposed by Goodfellow et al. in 2014 is a groundbreaking milestone for image reconstruction. Based on this, all kinds of GAN emerge such as DCGAN [8], FSGAN [8], WGAN[9], PGGAN [10], BigGAN [11], and StyleGAN [12]. Variational Encoder (VAE) such as DFC-VAE [13] also contributes to image reconstruction.

Some face detection networks would be useful in this research to detect whether the generated image is a person or belong to the same person. Multi-task cascaded convolutional networks (MTCNN)[14] is for face detection and face alignment by using 3 cascade stages on a course-to-fine strategy. The result will be marked with a face bounding box and landmarks. MTCNN can be used to extract face to accelerate training. FaceNet [15] is mainly for face identity detection. Other face detection networks for comparison include Fast R-CNN [16], Faster R-CNN [17], YOLO v1 [18], YOLO v2 [19], and YOLO v3 [20].

Approaches specifically targeted deepfake have capturing unreal eye-blinking [21], unnatural head poses [22], and inconsistent color space [23]. There are 3 methods to process video. One is to train frame by frame, the second is to use a recurrent neural network [24], and the third is to put all frames into one stack and detect inconsistency [25].

3 Experiments

3.1 Methods

Ffmpeg [6] can convert video to images or vice versa. I will train on the sample training dataset. If the network and weight are proved effective, I will train on full training videos to get the final weight.

It is a binary classification problem. In the algorithm, I may choose 1) Logistic Regression, 2) Supported Vector Machine, 3) K-means Clustering.

3.2 Results

placehoder

4 Discussion and Future Works

placehoder

Acknowledgments

placehoder

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[6] http://ffmpeg.org/

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[3] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.

[5] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale celebfaces attributes (celeba) dataset," *Retrieved August*, vol. 15, p. 2018, 2018.

[6] D. E. Rumelhart, G. E. Hinton, R. J. Williams *et al.*, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.

[7] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[11] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[12] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[13] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 1133–1141.

[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[16] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[19] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[20] ——, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[21] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," *arXiv preprint arXiv:1806.02877*, 2018.

[22] 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.

[23] H. Li, B. Li, S. Tan, and J. Huang, "Detection of deep network generated images using disparities in color components," *arXiv preprint arXiv:1808.07276*, 2018.

[24] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, p. 1, 2019.

[25] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 38–45.