FakeTalkerDetect: Effective and Practical Realistic Neural Talking Head Detection with a Highly Unbalanced Dataset

Hyeonseong Jeon[1], Youngoh Bang[1] and Simon S. Woo[2]
[1]Department of Artificial Intelligence
[2]Department of Applied Data Science
Sungkyunkwan University, Suwon, South Korea
{cutz, byo7000, swoo}@g.skku.edu

Abstract

Detecting realistic fake images and videos is an increasingly important and urgent problem because they can be maliciously used. In this work, we propose FakeTalkerDetect, which is based on siamese networks to detect the recently proposed realistic talking head with few-shot learning. Unlike conventional methods, we propose to use pretrained models with only a few real images for fine-tuning in siamese networks to effectively detect the fake images in a highly unbalanced data setting. Our FakeTalkerDetect achieves the overall accuracy 98.81% accuracy in detecting fake images generated from the latest neural talking head models. In particular, our preliminary work also demonstrates the effectiveness for the highly unbalanced dataset.

1. Introduction

Significant advancement made in Generative Adversarial Networks (GANs) allow generating highly realistic and sophisticate images such as human faces, scenes, objects, etc. However, these GAN generated images and videos can be misused and maliciously used to fake someone, objects, and facts, and can be exploited to further harm and attack individuals. Recent reported incidents [1, 2] by DeepFake [3] and DeepNude [4] show that these technologies can be abused and misused. To address these challenges, recent research [5, 6] shows various forgery detection techniques. However, it is not easy to cope with newer or different deep fake generation methods, if detection models are developed and trained with older approaches. Recently, Zakharov et al. [7] developed the new highly realistic personalized talking head generation models using few-shot adversarial learning methods. They use embedding layers to extract feature landmarks of human faces and process those into the generator in few-shot GAN, where original training inputs provide attention to middle of generator layers for few-shot learning.

Figure 1. Real images (left) and few-shot GANs generated images (right), where the left image is from the original VoxCeleb2 dataset and the next three images are the generated fake images from the few-shot GAN learned from realistic neural talking head models [7, 8]

In this work, we propose *FakeTalkerDetect* to detect images generated from highly realistic virtual talking heads from few-shot learning by Zakharov et al. [7]. Since Zakharov et al. [7] did not provide the source code, we used the implementation in PyTorch by Grey-Eye [8] to generate neural talking heads for Zakharov et al.'s approach. Our key contributions are summarized as follows: 1) Our preliminary results show that highly realistic fake neural head models can be detected with high accuracy with well known AlexNet [9], 2) We show that pre-trained model as well as siamese networks [10] can achieve higher accuracy with highly unbalanced dataset with the limited number of fake talking head images (e.g., 1%), and 3) We offer open source versions of our preliminary code for use by the broader research community [1].

2. FakeTalkerDetect Design

In this preliminary work, we aim to detect realistic talking heads-like images generated by few-shot adversarial learning [8, 7]. In particular, we focus on detecting highly unbalanced data setting, where we have significantly more real data than fake data. To address this challenge, we utilize a siamese network and few-shot learning for detection.

[1]https://github.com/cutz-j/FakeTalkerDetect

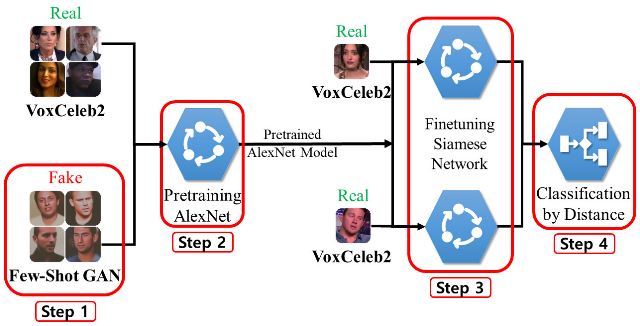

Figure 2. Overview of our *FakeTalkerDetect*

To generate fake images, we use the implementation developed in [8], where has an equal model architecture such as embedder, generator and discriminator layers for few-shot GANs.

Pre-training and Siamese Network. Our detection method, *FakeTalkerDetect*, first pre-trains well-known fake image classification model such as AlexNet [9], using real and fake image pairs, as shown in Step 1 and 2. in Fig. 2. In fact, Step 1 and 2 can be replaced with a pre-trained model. After pre-training, we further focus on improving the detection performance. Since we will have more real images than fake images in a real world detection scenario, we chose a siamese network to detect fake images after training real images as shown in Step 3 in Fig. 2. In Step 3, the siamese network learns two input pairs (e.g., real-real) and evaluates sum of square error of each pair, where the higher error means that they are different classes. Also, we use mean squared error loss function for fine-tuning, where this loss function runs over pairs of samples.

Few-shot Learning. In particular, the siamese network can be effectively learned by few-shot learning, because real and fake images are already discriminated in the pre-trained model. If their pairs have feature mappings, our model can detect the differences. Moreover, our model is more robust and generalizable to evaluate new and unseen data using the siamese distance. That is the main reason we can fine-tune only with real images on our detection model and expect to improve the performance in Step 3 in Fig 2. Finally, we fine-tune our siamese network and measure Euclidean distance to classify real and fake images generated from talking head models.

3. Experimental Results

We compare our model with the baseline AlexNet to compare the detection performance by varying the percentages of fake images in a testset from 50%, 25%, 5% and 1% to measure the performance for highly unbalanced data settings, which are more realistic and practical scenarios.

We use VoxCeleb2 [11] to create few-shot GAN talking head fake images (Step 1) by randomly selecting frames from VoxCeleb2 videos. We use 251,702 images for pre-training in Step 2. For training and fine-tuning siamese network, we only use 1,138 unseen real images and test different number unseen real and fake images for baseline and our model in Step 3. To evaluate the unbalanced data settings, we vary the percentage of fake images in the test set ranging from 50% (33,086), 25% (21,866 images), 5% (17,480 images) to 1% (16,723 images). We compared the performance between pre-trained AlexNet and our model. As the percentage of fake images reduces, the accuracy from the baseline AlexNet decreases. However, ours maintains above 98% accuracy consistently. With only 1% of fake images in a test set, our model achieves 74% in precision compared to 61% in AlexNet as shown in Table 1. The result shows F1 score increases simply by configuring siamese network and tuning the only real images. This demonstrates our model has more discriminative ability to distinguish real and fake images.

Table 1. Performance of pre-trained baseline AlexNet model vs. FakeTalkerDetect (ours) with different testing data sets, where % means the proportion of fake images in each testing data set.

Model	ACC	Recall	Prec.	F1
AlexNet (50% fake)	98.10	0.98	0.98	0.98
FakeTalkerDetect (Ours)	**98.44**	0.98	0.98	0.98
AlexNet (25% fake)	97.13	0.95	0.95	0.96
FakeTalkerDetect (Ours)	**98.64**	**0.98**	**0.98**	**0.98**
AlexNet (5% fake)	96.41	0.98	0.80	0.87
FakeTalkerDetect (Ours)	**98.81**	**0.99**	**0.91**	**0.94**
AlexNet (1% fake)	96.25	0.98	0.61	0.67
FakeTalkerDetect (Ours)	**98.84**	**0.99**	**0.74**	**0.82**

4. Conclusion and Future Work

We propose *FakeTalkerDetect*, which is the siamese network based classifier to detect few-shot GAN-based fake images. Our preliminary result with highly unbalanced datasets shows the promising detection performance. In the future, we plan to experiment with more powerful classification models and complex datasets with triplet loss [12].

Acknowledgement

This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00421, AI Graduate School Support Program), by NRF of Korea by the MSIT (NRF-2017R1C1B5076474 and NRF-2019M3F2A1072217).

References

[1] When seeing is no longer believing. *CNN Business*.

[2] Cao Yin. Altering faces via ai deepfake may be outlawed. *China Daily*, Apr 2019.

[3] Wikipedia. Deepfake. `https://en.wikipedia.org/wiki/Deepfake`, 2019. [Online; accessed 15-July-2019].

[4] lwlodo. Official deepnude algorithm source code. `https://github.com/lwlodo/deep_nude`, 2019.

[5] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

[6] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Gan is a friend or foe?: a framework to detect various fake face images. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1296–1303. ACM, 2019.

[7] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint arXiv:1905.08233*, 2019.

[8] Grey-Eye. Our implementation of "few-shot adversarial learning of realistic neural talking head models". `https://github.com/grey-eye/talking-heads`, 2019.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

A. Examples of Few-Shot GAN generated image

Figure 3. Real images (left) and few-shot GANs generated images (right), where the left image is from the original VoxCeleb2 dataset and the right image is the generated fake images from the few-shot GAN neural talking head models [7, 8]