Proceedings of the 39th Chinese Control Conference
July 27-29, 2020, Shenyang, China

DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network

Xu Chang[1,2], Jian Wu[1,2], Tongfeng Yang[1] , Guorui Feng[1,2]

1. School of Cyber Security, Shandong University of Political Science and Law, Jinan 250014

2. Key Laboratory of Evidence-Identifying in Universities of Shandong(Shandong University of Political Science and Law), Jinan 250014, China

E-mail: changxumail@163.com

Abstract: DeepFake can forge high-quality tampered images and videos that are consistent with the distribution of real data. Its rapid development causes people's panic and reflection. In this paper we presents an improved VGG network named NA-VGG to detect DeepFake face image, which was based on image noise and image augmentation. Firstly, In order to learn the tampering artifacts that may not be seen in RGB channels, SRM filter layer is used to highlight the image noise features; Secondly, the image noise map is augmented to weaken the face features. Finally, the augmented noise images are input into the network to train and judge whether the image is forged. The experimental results using the Celeb-DF dataset have shown that NA-VGG made great improvements than other state-of-the-art fake image detectors.

Key Words: DeepFake, Image Detection, VGG

1 Introduction

With the continuous development of artificial intelligence technology, especially the emergence of deep learning, image and video editing becomes more and more easy. Different from common tampering techniques such as spreading, copy move, and remove. DeepFake relies on the technology of deep learning. Through the algorithm of deep learning, it can identify the photos of different angles, postures and expressions of the target characters (such as celebrities, politicians, etc.), and then continuously train to automatically generate the fake pictures, and cover them to the faces of the original video characters to form the " DeepFake videos " [1]. Compared with PS and other image modification and tampering technologies, the reason why "DeepFake is worrisome is that it is a combination of high authenticity, pervasiveness and rapid evolution [2]. In November 2017, DeepFake has been widely used on Reddit for its production of many pornographic videos in the United States, which has attracted attention from all walks of life and gained a great reputation on the Internet. In January 2018, the application using DeepFake was officially launched, which further intensified the spread of DeepFake videos. The object of face swapping has also rapidly expanded from celebrities and politicians to friends, classmates and colleagues. The development of DeepFake naturally triggered people's panic and reflection. Therefore, the major technology enterprises also began to make joint action with the academic community to avoid further negative impact on the fierce discussion of whether and how to regulate the DeepFake technology.

*This work is supported by Open Fund of the Key Lab of Forensic Science,Ministry of Justice,China (Academy of Forensic Science);Program for Young Innovative Research Team and Big Data and Artificial Intelligence Legal Research Collaborative Innovation Center in Shandong University of Political Science and Law; Projects of Shandong Province Higher Educational Science and Technology Program under Grant No.J16LN19, J18KA357, J18KA383; Shandong Province Soft Science Research Project under Grant No.2019RKB01369.

Fig. 1. Deepfake images of Tommy Lee Jones and Ian McKellen

With the continuous improvement of computer computing power and the continuous reduction of hardware price, as well as the high integration of deep learning tools such as tensorflow[3] and keras [4], technicians with certain professional foundation can forge high-quality forged images and videos consistent with the real data distribution through convolution automatic encoder [5] and Generative adversarial networks (GAN) [6].Deepfake image as shown in Fig. 1 In addition, smart phones and desktop applications such as deepnude [7], faceapp and fakeapp [9] make it easy for the general public without a deep technical background to produce videos or pictures that are hard to identify with the naked eye. In particular, counterfeiting some shocking content on social media can be spread quickly without verification.

Deepfake videos on the Internet are mainly produced in the following three ways, including computer apps, service portals, and market services. Among them, computer apps refer to the tools used to create DeepFake videos, most of which have the most of these apps provide 'facewatching' capabilities; online service portals provide users with service

Authorized licensed use limited to: Lovely Professional University - Phagwara. Downloaded on February 24,2024 at 14:47:51 UTC from IEEE Xplore. Restrictions apply.

portals functions as online businesses for generating and selling custom deepfakes. Users can upload photos by uploading Market services refer to individual deep fake creators who advertise their services on forums and online markets.

So far, the published DeepFake tools have been widely used to produce pornographic videos of fake celebrities or revenge porn, which seriously infringes the personal rights and property rights of citizens. Such porn has been banned by reddit, twitter, ponhub and other websites [8]. On the other hand, they have also been used to produce fake news, such as forging politicians' specific statements, creating political tensions and destroying society stability, national security and international order. At present, governments are also considering these issues. In January 2020, Facebook has deleted DeepFake videos that according its standards [9]; DeepFake will destroy the trust mechanism of the social community. Human beings will enter a post truth era, and the truth will no longer exist. People only believe what they are willing to believe. As a result, social consensus is difficult to aggregate and the trust model will be broken. AI is like a double-edged sword. It can also bring harm. It can help [10]. DeepFake makes it more and more difficult to distinguish the real image from the processed image. More and more researches are devoted to the fight against counterfeiting in the digital world. We need to use all available tools to identify the true and the false, and prevent the criminals from using the tampered images for immoral business or political activities. But AI can also be used to detect fake images that human eyes can't see. Making and recognizing face swapping is like a cat and mouse game. The Deepfake technology is changing with each passing day, and the counterfeiting technology should also be iterative. In the future, in the face of increasingly severe trust crisis, it is urgent to need effective DeepFake image detection methods and technologies. It is becoming more and more important and challenging to study how to distinguish whether a picture or video is true.

It is for this reason that we proposed a DeepFake image detection VGG network (NA-VGG) based on image noise and image augmentation, as shown in Fig.2. Our contribution of this work is summarized as follows: first. The RGB image to be detected is used to highlight the image noise information through SRM filter layer to obtain the image noise characteristic image as the input of the network. Second, the noise image is flipped horizontally / vertically, Third, in view of the poor visual quality of a large number of depth forgery data sets, which cannot truly reflect the depth forgery image spread on the network, this experiment uses the Celeb-DF dataset [11], the experimental results show that that NA-VGG has great improvements in detecting DeepFake face images.

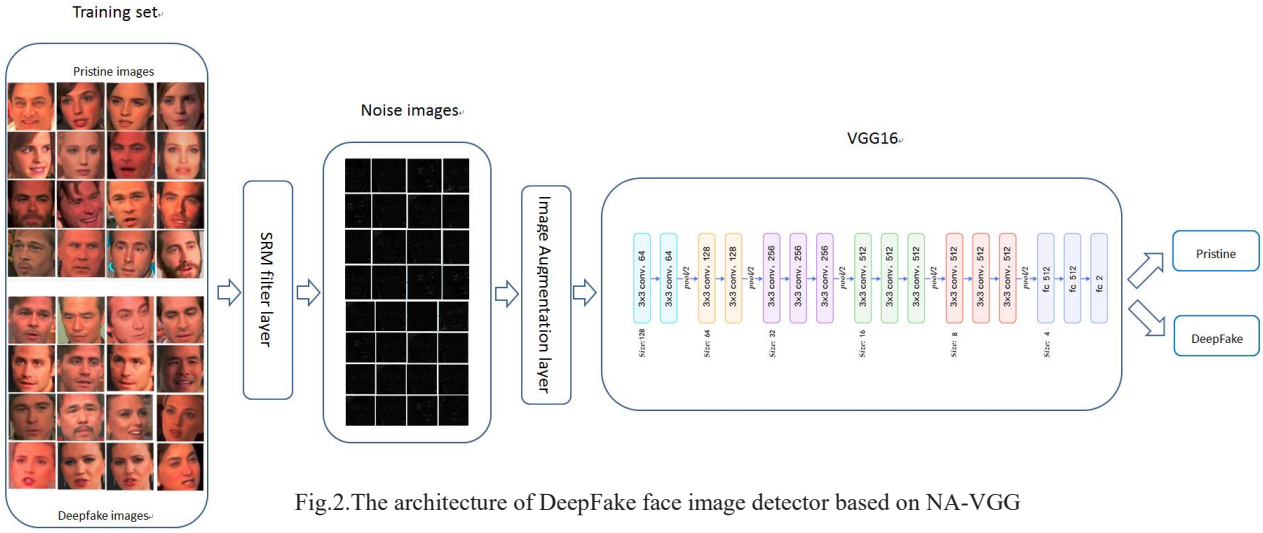

Fig.2.The architecture of DeepFake face image detector based on NA-VGG

2 Related Work

Digital Media Forensics. The purpose of digital media forensics technology [9] is to automatically evaluate the integrity of images or videos, and judge whether they have been tampered, synthesized, spliced or not only by the digital media itself. It is a passive forensics technology, which does not need to deal with media information in advance. Its universality and practicability make it an important research field. Traditional passive forensics technology of digital media mainly includes passive forensics based on traces left during forgery, passive forensics based on the consistency of digital media imaging equipment, and passive forensics based on the statistical characteristics of digital media itself [12]. Different from the traditional image editing tools such as PS, the emergence and continuous development of DeepFake technology, zero technology users can use desktop applications or apps to tamper with the image with one key, which is not only convenient to use, but also difficult to identify the authenticity of the faked image.In 2018, David Güera et al. [12] proposes a temporal-aware pipeline to automatically detect deepfake videos, convolutional neural network (CNN) is used to extract frame level features, which are then used to train a recurrent neural network (RNN);Peng Zhou et al. [20] proposes a two-stream Faster R-CNN network and train it end-to-end to detect the tampered regions given a manipulated image. In 2019,Chih-Chung Hsu et al. [23]proposes a deep learning-based methd to detect the fake image by combining the contrastive loss;Yuezun Li et al. [24] Based on the existing DeepFake algorithm, only limited resolution images can be generated, and distortion is needed to match the original face in the source video. It is proved that convolutional neural network (CNN) can effectively capture the artifacts left in the distortion matching process. The

Authorized licensed use limited to: Lovely Professional University - Phagwara. Downloaded on February 24,2024 at 14:47:51 UTC from IEEE Xplore. Restrictions apply.

above DeepFake algorithm improves the detection accuracy of DeepFake image from different perspectives, but most of the datasets used are of rough quality, and most of the tampering marks can be seen by the naked eye, resulting in its weak detection and weak generalization ability. With the continuous development of DeepFake technology, the forgery image quality is getting higher and higher, which leads to the high quality circulated in the re detection network of the above methods The performance of image decreases rapidly.

Generative Adversarial Networks(GAN). Deep learning has made breakthroughs in computer vision, voice and other application fields [14]. Compared with traditional machine learning, deep learning has a good representation ability, and it can automatically obtain abstract features. The model of deep learning can be roughly divided into discriminant model and generative model. Among them, the generative model has a direct impact on the real world modeling performance, and requires a lot of prior knowledge; in addition, because the real world data is too complex, it needs too much computation to fit the model, which makes the generative model become a very challenging and difficult to solve machine learning problem. Until 2014, Goodfellow et al. [15] proposed a new generation model inspired by the two person zero sum game in the game theory (that is, the sum of the two people's interests is zero, and the gains of one person are exactly the losses of the other person). The network system is composed of a generator G and a discriminator D. Both the generator and the discriminator can use the depth neural network which is currently hot in research [16]. The optimization process of GAN is a two-player minimax game with value function V (D,G) as formula (1), and the ultimate goal of optimization is to achieve the Nash equilibrium [17], which avoids the calculation of the distribution function brought by repeated application of Markov chain learning mechanism, and does not need the lower limit of variation or approximate inference, thus greatly improving Application efficiency [18]. At present, GAN has been continuously improved, such as DCGAN[13], CoGAN[22], ProGAN[25], StyleGAN[26] and other GAN networks, which have been successfully applied in the field of image generation and video generation. The quality of generation has been continuously improved, and it is difficult for the naked eye to recognize.

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{data}(x)}\left[\log D(x)\right] + \mathrm{E}_{z\sim p_z(z)}\left[\log\left(1-D(G(z))\right)\right] \quad (1)$$

where x represents the real picture, z represents the noise of input G, G (z) represents the picture generated by G network, D (x) represents the probability that D network judges whether the real picture is real,D (g (z)) is the probability that D judges whether the picture generated by G is real.

Convolutional neural network(CNN). CNN is a kind of feedforward neural network. It is one of the most representative neural networks in the field of deep learning. It has made many breakthroughs in the field of image analysis and processing. Compared with other neural network structures, CNN needs relatively few parameters, which makes it can be widely used and has gradually become a commercial application, almost wherever there are images, there will be CNN. The current popular CNN include LeNet5, AlexNet, AlexNet, ResNet, GoogLeNet and VGG[19], among which VGG is developed by researchers from the Visual Geometry Group of Oxford University and Google DeepMind Company. VGG explores the relationship between the depth of CNN and its performance, and successfully constructs a CNN of 16-19 layers. It is proved that increasing the depth of the network can affect the final performance of the network to a certain extent. Colleagues who reduce the error rate enhance the mobility, and the generalization of migrating to other image data is also very good. Therefore, this paper uses VGG16 to extract the Noise features of the DeepFake images, and improves and optimizes the network structure to detect the DeepFake face images.

3 NA-VGG for Deepfake Image Detection

3.1 SRM Filter Layer

Simple image splicing, moving and other tampering operations have obvious differences in contrast. The authenticity can be judged by extracting relevant features. While the high-quality DeepFake face image is different, and its fidelity is very high. RGB channel is not enough to deal with different forgeries [20], especially in the case of deep forgeries without obvious splicing boundary and contrast, the effect of feature extraction using RGB channel is not good. Therefore, this paper focuses on image noise rather than semantic image content to determine the authenticity of the image, uses SRM filter in image forensics [21] to extract local noise feature map from RGB image, and takes the local noise distribution data of the image as the network input, and then uses the noise feature to provide the basis for image processing for authenticity classification.

The SRM filter kernel used in this paper, its weights are shown in Fig.3, sets the kernel size of SRM filter layer in noise flow as 5 × 5 × 3, and the output channel size of SRM layer as 3. Figure 4 shows the noise characteristic of the image to be examined after passing through the SRM layer. It can be seen that the acquired image is not the content of the image, but highlights the image noise, which contains the tampering artifacts that may not be seen in the RGB channel.

$$\frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix}$$

Fig. 3. The SRM filter kernel used to extract noise features

3.2 Image Augmentation

Image Augmentation can translate, flip, rotate, zoom and enhance the existing image data to generate new images for training or testing. This operation can increase the number of pictures by several times, thus greatly reducing the possibility of over fitting. In this paper, we use the ImageDataGenerator class of Keras framework to augment the image to be examined, mainly through the random horizontal flip function of horizontal flip and vertical flip

Authorized licensed use limited to: Lovely Professional University - Phagwara. Downloaded on February 24,2024 at 14:47:51 UTC from IEEE Xplore. Restrictions apply.

function to weaken the features of human face, highlighting the detection of DeepFake trace features.

Fig. 4. Noise features of test images

3.3 VGG for Noise Feature Extraction

In our work, VGG16 is selected as the basic architecture. The convolution layer and pooling layer of VGG16 can be divided into different block, numbered Block 1 ~ Block 5 from the front to the back. Each block contains several volume layers and a pool layer, including 13 volume layers, 3 fully connected layers, and 5 pool layers. We add SRM filter layer and image Augment layer in front of VGG16 network, and propose a kind of VGG network based on noise and image augmentation (NA-VGG).

4 Experiments

4.1 Dataset

The development and evaluation of a DeepFake image detection algorithm need to meet two conditions: one is the need for large-scale datasets for training; the other is that the data in the datasets should truly reflect the quality of DeepFake images or videos on the Internet. Celeb-DF [22] is a new deepfake video dataset proposed by Yuezun Li et al. The dataset contains 408 original videos obtained from YouTube, and 795 DeepFake videos are synthesized from these real videos. Table 1 shows the average AUC performance of Celeb-DF dataset and UADFV, DeepFake-TIMIT(LQ,HQ), FF++ / DF, Celeb-DF and other datasets in different DeepFake detection methods. It can be seen that, compared with the previous DeepFake datasets, because of the high quality of DeepFake video in the Celeb-DF, the characteristics of artifacts are not obvious, and some detection models will reduce the performance on the Celeb-DF, resulting in the low detection accuracy.

Table 1: Average AUC performance of different methods on each dataset

Database	Average AUC performance
UADFV	78.7%
DeepFake-TIMIT(LQ)	73. 8%
DeepFake-TIMIT(HQ)	66.6 %
FF++ / DF	76.1 %
Celeb-DF	48.7 %

4.2 Image Dataset Preprocessing

In this paper, Celeb-DF is used for training evaluation. Firstly, the image is extracted from the DeepFake video. In this paper, Python + Opencv is used to extract the image by frame in the video. In order to ensure that the proportion of real image and DeepFake image is basically balanced, so in folder Celeb-real, 40 discontinuous images are captured for each video, and 10 discontinuous videos are captured for each video by Celeb-synthesis. 12416 training set images (including 5334 original images and 7082 DeepFake images) are extracted, 1376 verification set images (including 573 original images and 803 DeepFake images) and 1376 test set images (including 552 original images and 552 DeepFake images) are extracted 824 DeepFake images). The images of training set and test set have no repetition and are taken from different videos. They are disjoints sets. Secondly, we use classifier "haarcascade_frontalface_alt.xml" of OpenCV for face location and capture, and save the training set, verification set and test set of DeepFake detection.

4.3 Parameter Settings

Resizing of every image to 128*128,The optimizer is set to SGD for end-to-end training of the complete model with a learning rate of 0.01,decay of 1e-6, momentum of 0.9, and nesterov of True . The loss function is set to categorical_crossentropy as formula (2).

$$L(y,\hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N}\left(y_{ij} * \log\left(\hat{y}_{ij}\right)\right) \qquad (2)$$

where M represents the total number of categories, n represents the number of samples, Yij is the actual result, y^ is the predicted result.

4.4 Results and Analysis

The experimental results are shown in Statistical table of the comparison of average AUC score of each detection method on Celeb-DF (see Table 3). The results show that the accuracy of our method in detecting DeepFake images is much higher than that of several DeepFake detection models using Celeb-DF data set in reference [22]. Other DeepFake detection models are trained in other datasets with obvious artifact features such as low resolution, color mismatch and visible boundary. The learned features may not be applicable in high-quality DeepFake dataset Celeb-DF, resulting in performance degradation. In addition, from the experimental results, it can be seen that the SRM filter can enhance the image noise by 16.8% compared with the VGG16 network, and the image augmentation by 12.5%, which shows that the SRM filter can highlight the image noise feature, and the image augmentation is effective to improve the detection accuracy.

Table 2: AUC performance of different methods on Celeb-DF

Methods	Average AUC performance
Average of Several Methods[22]	48.7%
Two-Stream[10]	55.7%
VGG16	56.4 %
SRM filter +VGG16	73.2 %
NA-VGG	85.7 %

Authorized licensed use limited to: Lovely Professional University - Phagwara. Downloaded on February 24,2024 at 14:47:51 UTC from IEEE Xplore. Restrictions apply.

5 Conclusion

In this paper we have presented a VGG network based on noise and image augmentation (NA-VGG) to detect the DeepFake face image. Firstly, the RGB image to be detected is used to highlight the image noise information through SRM filter layer, and then the image noise characteristic image is obtained as the input of the network. Secondly, the image noise is flipped horizontally / vertically, and weakened by data augmentation .Finally, the experimental results using the Celeb-DF have shown that by using SRM filtering to highlight image noise and image augmentation to weaken face features, we can learn tampering artifacts that may not be seen in RGB channels. The results show that NA-VGG made great improvements in detecting DeepFake face images. In future work, we plan to introduce Siamese network and RGB feature during training to further improve the accuracy.

References

[1] J.Y. Zhu, T. Park, P. Isola and A.A. Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In: Proceedings of the IEEE International Conference on Computer Vision. 2017:2223-2232.

[2] L.S. Wang, On the integrated regulation of "deep forgery" intelligent technology, Oriental Law, 2019:1–14.

[3] M. Abadi et al, Tensorflow: A system for large-scale machine learning. Proceedings of the USENIX Conference on Operating Systems Design and Implementation, 16:265–283, 2016.

[4] F. Chollet, et al, https://github.com/fchollet/keras. Keras, 2015.

[5] A. Tewari et al, Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017: 1274–1283.

[6] G. Antipov, M. Baccouche, and J.-L. Dugelay, Face aging with conditional generative adversarial networks. arXiv:1702.01983, 2017.

[7] The state of deepfake,deeptrace, http://www.deeptracelabs.com ,2019:4-7.

[8] D. Güera , Edward J. Delp, Deepfake Video Detection Using Recurrent Neural Networks. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance(AVSS), 2018:1–6.

[9] SoHu, Facebook banned deepfake video before the presidential election,http://www.sohu.com/a/365259905_99-956743,2020.

[10] P. Zhou, X.T. Han, V. I. Morariu and L.S. Davis , Two-stream neural networks for tampered face detection, IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[11] Y. Li, X. Yang, P. Sun, H.G. Qi and S.W. Lyu, Celeb-DF: A New Dataset for DeepFake Forensics, arXiv preprint arXiv:1909.12962, 2019.

[12] R.C. Chen, Research on passive forensics of digital media based on object edge analysis.Hunan University,2012.

[13] R. Alec, M. Luke, C. Soumith, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, arXiv preprint arXiv:1511.06434,2016.

[14] W.L. Wang, Z.R. Li. Research progress of generative countermeasure network, Journal on Communications, 2018.

[15] I. Goodfellow, J. Pouget-abadie, M. Mirza, et al, Generative adversarial nets[C]//International Conference on Neural Information Processing Systems. 2014: 2672–2680.

[16] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning. Cam-bridge, UK: MIT Press, 2016.

[17] K.F. Wang,C. Gou, Y.J. Duan, Y.L. Lin, X.H. Zheng ,F.Y.Wang, Generative Adversarial Networks: The State of the Art and Beyond.Acta Automatica Sinica, 43(3),2017.

[18] I. Goodfellow, Generative adversarial networks[J], arXiv: arXiv1701.00160, 2017.

[19] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2014.

[20] Peng Zhou, Xintong Han, Vlad I. Morariu, Larry S. Davis, Learning Rich Features for Image Manipulation Detection, arXiv: arXiv 1805.04953,2018.

[21] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 7(3): 868-882,2012.

[22] M.Y. Liu, O. Tuzel, Coupled Generative Adversarial Networks, arXiv:1606.07536,2016.

[23] Chih-Chung Hsu , Yi-Xiu Zhuang, and Chia-Yen Lee, Deep fake image detection based on pairwise learning, Preprints, 2019.

[24] Yuezun Li, Siwei Lyu.Exposing, DeepFake videos by detecting face warping artifacts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019:46–52.

[25] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive Growing of GANs for Improved Quality, Stability, and Variation, arXiv:1710.10196,2018.

[26] T. Karras, S. Laine, T. Aila, A Style-Based Generator Architecture for Generative Adversarial Networks, arXiv:1812.04948,2019.

[27] Y.J. Zhang, A Course of Image Processing and Analysis(2nd Edition).Beijing:Posts & Telecom Press,2016.

Authorized licensed use limited to: Lovely Professional University - Phagwara. Downloaded on February 24,2024 at 14:47:51 UTC from IEEE Xplore. Restrictions apply.