ECEA 2019 5th International Electronic Conference on Entropy and Its Applications
18–30 November 2019

sciforum

Conference Proceedings Paper

Exposing Face-Swap Images based on Deep Learning and ELA Detection

Zhang Weiguo and Zhao Chenggang *

College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China

Abstract: The new development in Artificial Intelligence, has significantly improved the quality and efficiency in generating fake face images; for example, the face manipulations by DeepFake is so realistic that it is difficult to distinguish the authenticity—either automatically or by humans. In order to enhance the efficiency of distinguishing facial images generated by AI from real facial ones, a novel model has been developed based on deep learning and ELA detection, which is related to entropy and information theory, such as cross-entropy loss function in the final Softmax layer, normalized mutual information in image preprocessing and some applications of encoder based on information theory. Due to the limitations of computing resources and production time, DeepFake algorithm can only generate limited resolutions, resulting in two different image compression ratios between the fake face area as the foreground and the original area as the background, which would leave distinctive artifacts. By using the error level analysis detection method, we can detect the presence or absence of different image compression ratios, and then use CNN to detect whether the image is fake. Experiments show that the training efficiency of CNN model can be significantly improved by using the ELA method. And the detection accuracy rate can reach more than 97% based on CNN architecture of this method. Compared to the state-of-the-art models, the proposed model has the advantages such as fewer layers, shorter training time and higher efficiency.

Keywords: artificial intelligence; deep learning; DeepFake detection; ELA detection

1. Introduction

Nowadays, with the popularization of smartphones and various face-swap applications, the manipulation of visual content is becoming more and more common, which has become one of the most critical topics in the digital society. Faces are the main focus of visual content manipulation. There are many reasons for this focus. First of all, face reconstruction and tracking is a relatively mature field in computer vision [1], which is the basis of these editing methods. Then the human faces play a key role in communications, because human face can emphasize and convey some certain information in its own ways [2].

The root of the problem comes from the new generation of generative deep neural networks [3], which are capable of synthesizing videos from large volume of training data with minimum manual editing. And the appearance of DeepFake [4] greatly reduces the threshold of face forgery techniques. DeepFake replaces the face in an original video with the face of another person by using generative adversary networks(GANs) [5]. Because the GAN models were trained using tens of thousands of images, it is possible to generate realistic faces that can be spliced into the original video in an almost perfect way. Through suitable post-processing, the resulting video can achieve higher authenticity.

In addition to DeepFake technology, there are also Fake2Face [6] and Faceswap [7] as prominent representatives for facial manipulations. Recently, it became popular with wide-spread consumer-level applications like ZAO in China. While face swapping based on simple computer graphics or

deep learning is running in real time, DeepFakes need to be trained for every pair of videos, which is a time-consuming and resource-demanding task.

Before the emergence of fake video, it was generally believed that videos are reliable and dependable, and video evidences were widely used in multimedia forensics. However, after the prevalence of fake videos, people's psychological security zone is broken. There is widespread concern that once such fake videos are used for court proof, press and publication, political elections, television and entertainment, it is difficult to estimate the impact on people's lives. And some people even think that this technology could hinder the development of society. In this case, detection and identification of such fake videos, whether for digital media forensics, or the ordinary people's lives, have become extremely urgent.

In this paper, we describe a novel model based on the deep learning and ELA detection, which can effectively distinguish facial images generated by AI from real facial ones. Our experiment is based on a characteristic of DeepFake principle: due to the limitations of computing resources and production time, the DeepFake algorithm can only generate limited resolutions, resulting in two different image compression ratios between the fake face area as the foreground and the original area as the background, which would leave distinctive artifacts.

By using the error level analysis (ELA) detection method, our model can capture such artifacts. Because the entire image should have roughly the same compression level for JPEG formats. However, if a part of the image has been modified, such as copy and paste, and other removal operations, there will be a significant error level between the tampered part and the surrounding part. At this time, ELA images with different error levels can be generated by ELA method, and the tampered part will be displayed as obvious white color.

By using the ELA detection method, we can detect the presence or absence of the image different compression ratios [8]. For the generated ELA image of the real face and fake face, we will input it into the special convolutional neural network model and train a binary classifier to distinguish whether the image is fake.

2. Related Work

2.1. AI-based Video Synthesis Algorithms

With 3D computer graphics-based methods, it is easy to generate realistic images/video. Recently, the new deep learning algorithms have developed rapidly, especially those based on the generative adversary networks (GAN). Goodfellow et al. [9] first proposed the new generative adversary networks (GANs), which usually consists of two networks: generator and discriminator. Face2Face, proposed by Thies et al. [6], is an advanced real-time facial reenactment system, which can change the facial movements in video streams, such as videos from the movies.

Recently some facial image synthesis methods based on deep learning techniques have been proposed. Most of these techniques have the problem of low image resolution. Karras et al [10] use progressively growing of GAN to improve image quality. Their results include high-quality facial synthesis.

2.2. GAN Generated Image/Video Detection

With the popularity of face-swap applications, detecting GAN generated images/videos technology has also made some progress. Li et al. [11] observed that DeepFake faces lack realistic eye blinking, because the image collected through the Internet typically does not include the photo of closed eyes. Therefore, the lack of eye blinking is detected with a CNN/RNN model to expose DeepFake videos. However, this method can be invalid by purposely adding images with eyes closed in training.

Li et al. [12] used the color difference between GAN generated images and real images in non-RGB color spaces to classify them. Afchar et al. [13] trained convolution neural network to directly

classify real faces and fake faces generated by DeepFake and Face2face [6]. Although it showed promising performance, the overall approach has its drawbacks. In particular, it needs both true and false images as training data, and generating the fake images using the AI-based synthesis algorithms is low efficient than simple mechanism for training data generation in our method. Because extracting features directly from the original image, it needs to go through too many training cycles, resulting in low efficiency.

2.3. Image Tampering Detection

As the reliable evidence of judicial identification, digital image authentication technology has made a series of achievements in the field of image tampering detection. Previous methods can be classified according to the image features they aim at, such as CFA pattern analysis, local noise estimation, double JPEG localization. Bianchi et al. [14] proposed a probability model for estimates DCT coefficients and quantization factors. FUD et al. [15] determined whether the image has been tampered by estimating quality factor. Ferrara et al. [16] proposed a model to estimate the camera filter mode based on the difference of the variance of prediction error between CFA existing areas (authentic areas) and CFA absent areas (tempered areas). After the Gaussian Mixture Model (GMM) classification, the tampered regions can be localized.

3. Methods

In this section, we will describe the method of detecting facial images forgery in detail. First of all, we analyzed the principle of DeepFake generating face and simulate the process of affine transformation generating a fake face. Then the data sets of real face and fake face are processed by ELA method, and the resulting ELA image will highlight some parts of the original image where the error level is higher than the threshold value, that is, the affine transformation introduced artifacts. Finally, a binary classifier is trained by convolutional neural network (CNN) to distinguish whether the image is fake.

3.1. Data Sets Preprocessing

We analyzed the process of generating fake face by DeepFake. The principle of DeepFake is shown as Figure1. Due to the limitations of computing resources and production time, DeepFake algorithm can only generate limited resolutions, and then perform affine transformation on those generating images, such as scaling, rotation and shearing, to match and cover the original face that they will replace (see Figure1g–h). This will result in two different image compression ratios between the fake face area as the foreground and the original area as the background, which would leave obvious artifacts.

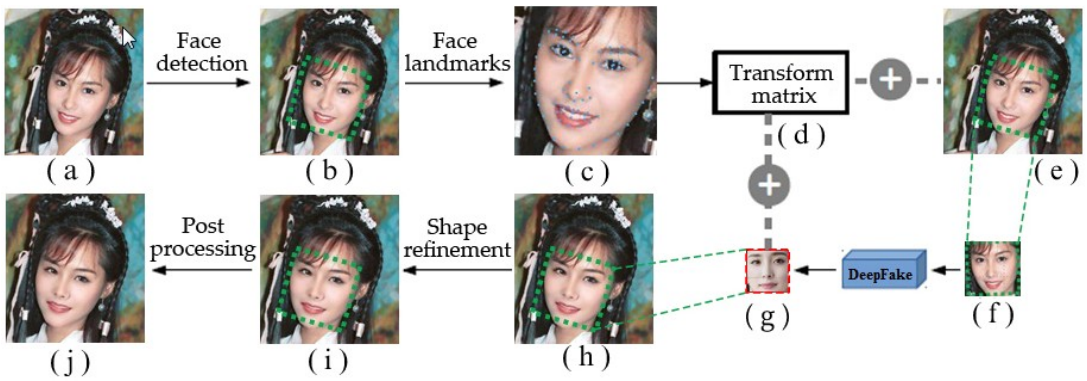

Figure 1. Overview of the DeepFake production pipeline. (**a**) the source image. (**b**) the green box is the detected face area. (**c**) blue points are the face landmarks. (**d**) calculate the transform matrix to wrap face region in (**e**) to the normalized region (**f**). (**g**) the face image generated by neural network,

and will be used to cover the source image (a). (**h**) Synthesized face wraped back using the same transform matrix. (**i**) post-processing, such as applying boundary smoothing to composite image. (**j**) the final synthesized image.

Our purpose here is to detect the artifacts introduced by the affine face wraping steps in DeepFake production pipeline. On the other hand, due to DeepFakes need to be trained for each pair of videos, which is a time-consuming and resource-demanding task, we did not use DeepFake algorithm to create negative examples. Instead, we simplified the process of generating negative examples by simulating the process of generating face in DeepFake (Figure 1).

Specifically, we take the following steps to generate negative examples, as shown in Figure 2: First, we detect faces in the original image, extract face landmarks from each detected face area, calculate the transform matrix according to the landmarks. Then apply Gaussian blur to the adjusted face. According to the inverse of transform matrix, the face is wrapped back to the original angle and cover on the original face.

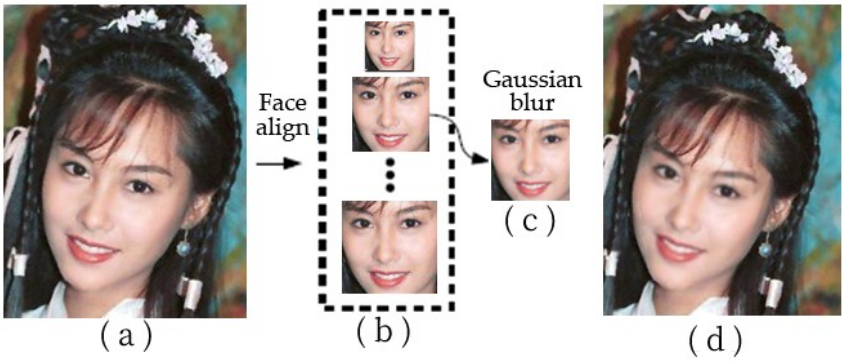

Figure 2. Overview of the process of generating a negative example. (**a**)is the original image. Use dlib to extract the face in (**a**) and align face with different scals as (**b**). We randomly selsect a scale of face in (**b**) and apply Gaussian blur as (**c**), and after the affine transform, cover the original image to generate a negative example.

In order to simulate more different resolutions of face affine transform in reality, we align faces into multiple scales and randomly select one scale to enlarge the training diversity. At the same time, we also use image enhancement technology to simulate different post-processing technology that may exist in the DeepFake process. Our approach also further deals with the shape of the face area affine transformation to cope with the different post-processing techniques.

In addition, we also need to preprocess the images before training. Since the input of convolutional neural network is 128 x 128, the size of the image area is not large. Therefore it is important to retain the most effective and prominent signs of forgery area as our region of interest (RoI). Analyzed as above, there are many traces forgery marks in the surrounding region involved affine transform, thereby retain the rectangular region composed of the convex hull of facial landmarks (except the contour of the cheek) and the surrounding area, and remove the remaining part of the image.

Specifically, for all positive and negative examples in the dataset, we only keep the above rectangular region, which is slightly larger than the external rectangular region of the convex landmarks of the face (except the contour of the cheek). We determine the RoIs using face landmarks, as $[y_0 - \hat{y}_0, x_0 - \hat{x}_0, y_1 + \hat{y}_1, x_1 + \hat{x}_1]$, where y_0, x_0, y_1, x_1 indicates the minimum bounding box b which you can cover all the facial landmarks except the cheek contour. The variables $\hat{y}_0$, $\hat{x}_0$, $\hat{y}_1$, $\hat{x}_1$ are random values between $[0, \frac{h}{5}]$ and $[0, \frac{w}{8}]$, where h, w are the height and width of b respectively. The RoIs are resized to 128 × 128 for the next ELA processing.

3.2. ELA Processing

Error level analysis method is one of the techniques for detecting tampering image. ELA method can obtain the compression distortion during lossy image compression. This method detects image tampering by storing images at a specific level of quality and calculating the ratio between compression levels [17]. Typically, this method is performed on images with lossy compression formats, such as JPEG.

When saving images in JPEG format, it will be independent "lossy compression" in units of 8 × 8 pixels. After lossy compression of JPEG, there are significant differences between the ELA of the original area and the ELA of the spliced or modified. If the image has not been modified, the compression difference of each 8 × 8 pixel region is similar. We will check the "compression feature" of the tested image with an 8 × 8 pixel grid. So if the image is saved as a whole, the compression feature of the adjacent grid should be an approximately high-frequency white distribution.

On the contrary, if it is saved after editing or modification, the ELA distribution between the grids will have obvious difference characteristics, which is shown as discontinuous high-frequency white distribution. The more times the images are stored or edited, the lower the ELA. The ELA processing effect is shown in Figure 3.

Figure 3. Samples of the original image and tampered images and their ELA results: The first line are the original image and its ELA image. It can be seen that the compression ratio of the whole image remains the same. The second line is the tempered image and its ELA image. It can be seen that the compression ratio between the tampered face as the foreground and the original image as the background are quite different.

4. Experiments

In this section, we first introduce our dataset, then evaluate our model on it. In addition, we visualize present our results, in order to better understand the proposed model.

4.1. Dataset

Although there are some datasets [18–21] for image tampering detection, they are not suitable for large-scale facial tampering detection. Because there are not enough tampering samples concentrated in facial areas. The Columbia Image Splicing dataset [18] and CASIA [19,20] are large but most of the tampered areas are not human faces. The DSI-1 dataset [21] focuses on facial tampering but the total number of tampered images is only 25. Therefore, it is difficult to train deep learning methods on these datasets to detect facial tampering.

To do this, we used the MUCT database, which consists of 3755 facial images and 76 manual facial landmarks. Each compressed file in the data corresponds to a camera, providing more diverse lighting, age, and race than the currently available 2D face database.

We take the 3755 "jpg" format face images in the database as the examples, and the negative examples can be generated by simulating the DeepFake algorithm, as shown in Figure 2, but it requires us to train and run DeepFake, which is a time-consuming and resource-demanding algorithm. Therefore, we use the method in 3.1 to generate negative examples dynamically and train them. Dynamic means that instead of generating all the negative examples in advance before the training process, we randomly select half of the positive examples for each training batch and convert them to negative examples according to the process in Figure 2, so as to make the training data more diverse.

4.2. Experiment Setup

For the 128 × 128 ROI region images generated in the previous step, we use ELA method to process them and get their ELA images. The CNN model that we trained uses these ELA images, rather than the original ones. Converting the original image to ELA image is a method to improve the training efficiency of CNN model. Because the ELA image does not contain as much information as the original image, it can improve the efficiency.

The feature generated by ELA image focuses on the part of the original image where the error level is higher than the threshold value. In addition, the pixels in the ELA image are often quite different from the nearby pixels, and even the contrast is very obvious, so the image processed by ELA makes the training CNN model more effective.

Therefore, we train a CNN model to extract the features of the ELA images, then detect whether the input image is forged or not. In the architecture we use, only two convolution layers are required, because the ELA images generated during the conversion process can highlight the characteristics of the original image where the error level is higher than the threshold value. So it is easier to determine whether the image is fake.

The maximum accuracy of the results obtained by our proposed method is 97%. The image of the accuracy curve and the loss function curve can be seen in Figure 4a. The confusion matrix of verification data is shown in Figure 4b.

As shown in Figure 4, our model achieves the best accuracy in the ninth cycle. From the first nine cycles later, verify that the value of the loss function starts to be flat and eventually began to increase, which is a sign of over-fitting. This is also a recognition method of ending training in advance during training, that is, when the verification accuracy value begins to decrease or the verification loss value starts to increase, the training will be stopped.

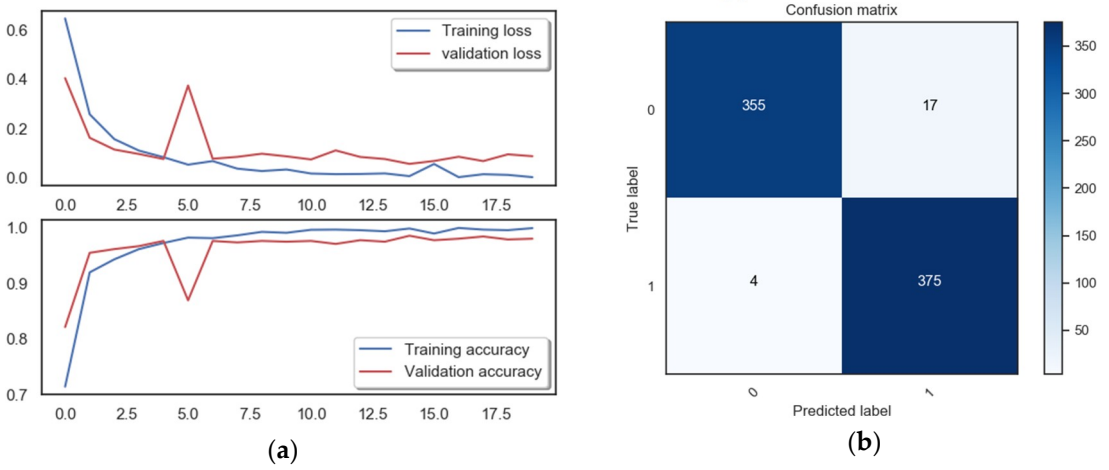

Figure 4. This is the experimental result images: (**a**) The image of the accuracy curve and the loss function curve; (**b**) The confusion matrix of verification data.

4.3. Comparison with Other Methods

We compare our method with the method [22] of training directly using CNN without ELA processing. The code for this method is available from the public implementation on GitHub [23]. In this method, the positive and negative examples images are directly input to the network model for training. And this method train 4 CNN models—VGG16, ResNet50, ResNet101 and ResNet152. The AUC performance on VGG16, ResNet50, ResNet101 and ResNet152 reached 83.3%, 97.4%, 95.4%, 93.8%.

However, compared with our method, this method has the following problems:

1. Deep learning training model is lack of explanatory and cannot explain the deep principle of identification forgery. Our ELA method can explain the principle.
2. This method is too complicated to train. On the one hand, if there is no GPU environment, it will lead to a long training period. On the other hand, it also requires a larger number of samples to participate in the training. Our method using two layers convolution, a MaxPooling layer, a fully connected layer, an output layer with Softmax can reach 97% accuracy, and greatly reduce the training time and the training period.

The advantages of our model are as follows: the number of training periods required to achieve convergence is significantly reduced, because the image features processed by ELA make the training more efficient, and accelerate the convergence of CNN model. On the other hand, the accuracy of our classification results is very high. This indicates that the features in the image processed by ELA can be successfully used to classify whether the image is fake. Experiments show that the training efficiency of CNN model can be significantly improved by using the ELA method.

5. Conclusions

The new development in AI, has significantly improving the quality and efficiency in generating false face. In this work, we studied a new model based on the deep learning, which can effectively distinguish facial images generated by AI from real facial ones.

We evaluated our method proved its effectiveness in practice. This indicates that the features in the image processed by ELA can be successfully used to classify whether the image is fake. Experiments show that the training efficiency of CNN model can be significantly improved by using the ELA method.

As the technology behind DeepFake continues to develop, we will continue to improve detection methods. We want to evaluate and improve the robustness of our detection methods for video compression.

References

[1] M. Zollh¨ofer, J. Thies, D. Bradley, P. Garrido, T. Beeler, P. P´eerez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. 2018. 1, 2

[2] C. Frith. Role of facial expressions in social interactions. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1535), Dec. 2009. 1

[3] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in NIPS, 2017, pp. 700–708.

[4] Deepfakes Github, howpublished = https://github.com/deepfakes/faceswap, note = Accessed: 2018-10-29. 1, 5, 9

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in NIPS, 2014.

[6] J. Thies, M. Zollh¨ofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2387–2395, June 2016. 1, 2, 3, 4, 5, 6, 10

[7] Faceswap. https://github.com/MarekKowalski/FaceSwap/. 1

[8] N. Krawetz, "A pictures worth digital image analysis and forensics," Black Hat Briefings , hlm. 1-31, 2007.
[9] 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.
[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. CoRR, abs/1710.10196, 2017. 3
[11] Yuezun Li, Ming-Ching Chang, and Siwei Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," in IEEE International Workshop on Information Forensics and Security (WIFS), 2018.
[12] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang, "Detection of deep network generated images using disparities in，color components," arXiv preprint arXiv:1808.07276, 2018.]
[13] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Mesonet: a compact facial video forgery detection network," in IEEE International Workshop on Information Forensics and Security (WIFS), 2018.
[14] T. Bianchi, A. De Rosa, and A. Piva. Improved dct coefficient analysis for forgery localization in jpeg images. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 2444–2447. IEEE, 2011. 1, 2, 3, 6, 7
[15] FUD, SHI Y Q, SU W. A, generalized Benford's law fox JPEG coefficients and its applications in image forensic [C] //Electronic Imaging. International Society for Optics and Photonics,2007
[16] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva. Image forgery localization via fine-grained analysis of cfa artifacts. IEEE Transactions on Information Forensics and Security, 7(5):1566–1577, 2012. 1, 2, 3, 6
[17] N. Krawetz, "A pictures worth digital image analysis and forensics," Black Hat Briefings , hlm. 1-31, 2007.
[18] T.-T. Ng, J. Hsu, and S.-F. Chang. Columbia image splicing detection evaluation dataset, 2009. 2, 5
[19] J. Dong, W. Wang, and T. Tan. Casia image tampering detection evaluation database 2010. In http://forensics.idealtest.org. 2, 5
[20] J. Dong,W.Wang, and T. Tan. Casia image tampering detection evaluation database. In Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on, pages 422–426. IEEE, 2013. 2, 5
[21] T. J. De Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. de Rezende Rocha. Exposing digital image forgeries by illumination color classification. ieee transactions on information forensics and security, 8(7):1182–1194, 2013. 1,2, 3, 5
[22] Yuezun Li, Siwei Lyu .Exposing DeepFake Videos By Detecting Face Warping Artifacts .[J] arXiv preprint arXiv:1811.00656. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),2019
[23] CVPRW2019_Face_Artifacts: https://github.com/danmohaha/CVPRW2019_Face_Artifacts