Machine Learning based Signal Processing for Detecting GAN Generated DeepFake Images

1st Suryansh Singh Rawat
Department of Electronics and Electrical Engineering
Birla Institute of Technology and Science, Pilani
Goa, India
f20180021@goa.bits-pilani.ac.in

***Abstract*—Recent advancements in the field of GANs (Generative adversarial networks) such as StyleGAN, have made it possible to generate realistic looking facial images. These GAN generated DeepFake images which are indistinguishable to the naked eye, are constructed from nothing but a low dimensional random noise, provided a training set. Artificially fabricated images like these can cause fraudulent and inappropriate informations to float in media, causing severe problems. Technologies such as GANs are evolving at an exponential rate and with this we are faced with an unprecedented potential for violation of basic human rights. Hence, we need AI based solutions that help us deal with the misuse of such technologies and come up with innovative approaches to detect and correctly classify images/videos as real or fake. In this paper, we will propose an efficient CNN architecture to detect StyleGAN based DeepFake images, which performs at par with the current state-of-the-art methods.**

***Index Terms*—GANs, DeepFakes, Fake Face Detection, Convolutional Neural Networks**

I. Introduction

DeepFake, an unheard-of concept to the general masses until 2017, but now is one of the most talked about topics in the world of AI. Its influence dominates the social media world. DeepFake utilizes both supervised machine learning (ML) methods such as convolutional neural networks (CNNs) and generative adversarial networks (GANs) along with unsupervised ML methods such as autoencoders to create fabricated images, videos, and audio. Recent additions to the basic DeepFake framework have involved the use of generative adversarial networks (GAN) [1] which are generative models that learn the distribution of the data without any supervision. Given these innovations in the DeepFake images and videos now being synthesized, there is now a serious forensics problem in how we can distinguish between fake images from real ones. In 2017, a Reddit user trained machine learning algorithms using explicit films and the celebrity actresses to create several adult videos which were uploaded on the subreddit "r/deepfakes" and went viral in no time [2]. It is actually threatening that privacy can be infiltrated by methods like this and be manipulated into something that is far from reality.However, every coin has two faces and we must acknowledge the positive side as well. Recently GANs are used for generating Speech-driven facial animation in which using speech signals and only a still image of a person, we can automatically synthesize a talking character [3]. But we have to agree to that fact that the number of illegitimate uses of DeepFakes dominates that of the positive ones by a large factor. We belong to a society where most of our devices use technologies such as face recognition and naturally we rely on them. Images form a major part of the data that is available on the web and hence it is necessary for people to be aware of such high end technologies and not blindingly trust any information just because it's in image form. A research study was recently published which tried to measure people's ability to recognize whether a photo had been doctored or not, purely by visual inspection [4]. A significant percentage of photos i.e 62%-66% were classified correctly. When asked about localizing the manipulation in the images, the users proved to be worse at that. Finding the truth in digital domain therefore has become increasingly critical and hence our project addresses the need for the application of machine learning methods to forensically ascertain the fake versus truth data.

II. Related Works

In image forgery detection, a traditional method of active schemes proposed by *Chang et. al* [5] is used in which an externally additive signal like a unique watermark is embedded in the source image without visual artifacts. To verify if the image is manipulated or not, a process is performed on the target image which includes the extraction of watermark and restore. The image of the watermark which is extracted during this process is used to verify the manipulated regions of the target image. However, in the case of GANs since we don't get the source image for the generated image, the method mentioned by *Chang et. al* can't be used as the watermark extraction process would fail.

While most of the advancements and research on DeepFake images is done in the image domain, a classical approach was proposed by *Frank et. al* [6] in the field of frequency domain. GANs may have become highly advanced, but it is observed that these GAN generated images consist of several distinguishable artifacts in the frequency domains which are mainly caused due to the upsamping during generation. These artifacts can be easily identified in the frequency spectrum and can be used as a major parameter for detecting DeepFake images.

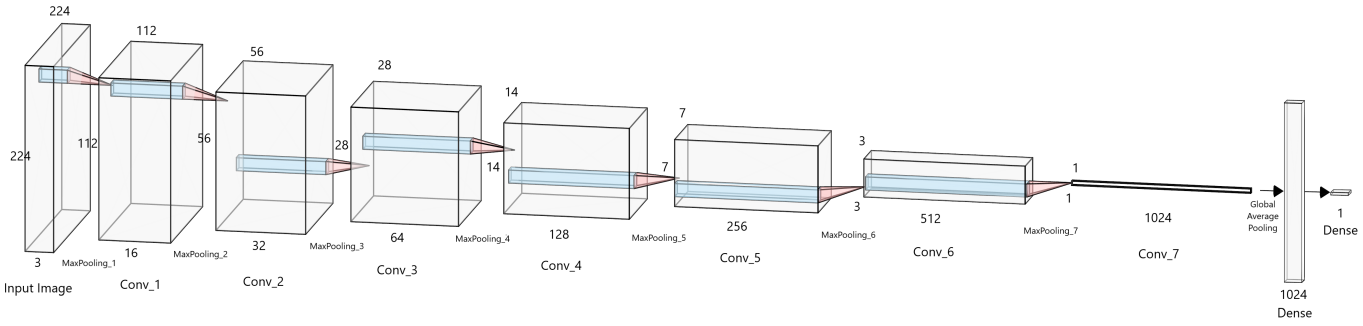

Fig. 1. Proposed CNN architecture

III. Proposed Model

Since the generators and discriminators of the StyleGAN architecture display the use of convolutional layers [7], it was natural for us to use a CNN based approach to train our model to detect StyleGAN based fake images. Hence, we carefully designed a CNN architecture as displayed in Fig 1. The input of our proposed model will be an RGB image which will be fed into 7 convolutional layers (3×3 size and 1×1 stride) each paired with batch normalization layer and max-pooling layer (2×2 size). The activation function used for all the convolutional layers was Rectified Linear Unit (ReLU). We also applied dropout for each convolutional layer to minimize over fitting. In addition to applying a dropout, we also used padding in order to allow the kernel to have more space to inspect the image. Finally, after the convolutional layers, we also added a dense layer at the end with sigmoid activation since this was a binary classification problem. The optimisation function in the proposed architecture is Adam optimization with a learning rate of 0.0001, decay constant of 1e-6 and the batch size for training the model is set to 64.

IV. Dataset Information

The name of the dataset which we used for this paper is "140k Real and Fake Faces" which is a publicly available dataset on Kaggle [8]. This dataset consists of 70k REAL face images from the Flickr dataset collected by Nvidia, along with 70k fake face images sampled from the 1 Million FAKE faces (generated by StyleGAN). the dataset consists of an equal ratio of real images to fake images, each of which has a dimension of 256×256 Fig 2.

Fig. 2. Fake images from the training dataset and image size 256×256

V. Environments and Results

A. Environments

We built the system on a Windows environment with Python 3.5. In the system, we use the Keras library based on TensorFlow for developing Deep Learning models. The training machine has the core i5-8250U CPU with an Intel(R) UHD Graphics 620 with 8GB RAM.

B. Data Splitting and Preprocessing

Here we are using the StyleGAN generated image dataset with high resolution and good quality real and fake face images that include 140k images. The input dataset is first preprocessed before putting it as an input in our proposed model. We resized the images to 224×224 as doing this would retain the features of the original image and we can also use this size of images as an input for pre-trained models.

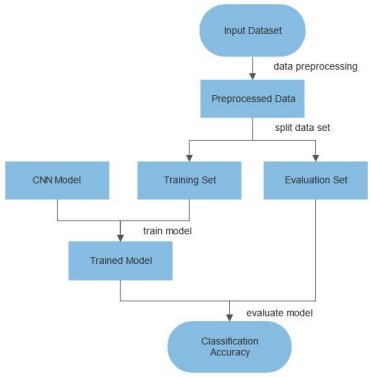

Fig. 3. Overall process flowchart

This dataset is equally divided into Real and Fake images i.e. 70k Real and 70k Fake images. To train our model we used 100k images in total. After separating the training set images, the remaining images were divided into two sets of 20k images each for validation dataset and test dataset. Fig 3. shows us the concise summary of the overall processes from preprocessing the images till the evaluation of our model in the form of a flowchart.

C. Assessment Model

For the assessment of our model, we will compare our performance with metrics such as precision, recall and accuracy, which are defined as:

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

Here, TP is the number of true positives, denoting that a real image was predicted as real, FP is the number of false positives, denoting that a fake image was predicted as real, FN is the number of false negatives, showing that a real image was predicted as a fake and TN is the number of true negatives, showing that a fake image was predicted as a fake.

We will also be using Area under the ROC Curve (AUROC) as a metric to test the performance of our model. ROC curve is a graphical plot to show the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

D. Results

In the model proposed by *Hsu et. al* [9] the values of precision and recall R were 0.930 and 0.936 respectively for SA-GAN (Self Attention GAN). The VGGFace architecture proposed by Nhu et al. [10] gave an accuracy of 80% and an AUROC value of 0.807 with a dataset created using PG-GAN and DC-GAN. In contrast to that, from (1) and (2) We achieved the values of precision and recall as 0.9843 and 0.9710 respectively and the accuracy from (3) comes out to be 97.77% which can also be clearly seen from the Confusion Matrix in the Fig 4. The AUROC curve shows a value of 0.9977 as displayed in the Fig 5.

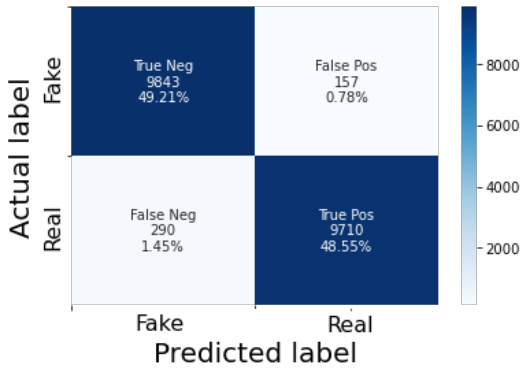

Fig. 4. Confusion matrix of the evaluation dataset

VI. Conclusion

In summary, we present a CNN architecture which is computationally cheaper and performs at par with the current state-of-the-art methods. We used DeepFake image dataset which was created using StyleGAN to train our model and make it reliable enough to classify real and fake images correctly. Our results based on the evaluation dataset, points to the fact that even though the current GAN based techniques have evolved to an extent that they produce realistic looking images, some statistical artifacts such as noise would be inevitably introduced which would serve as an evidence for detecting the fake images.

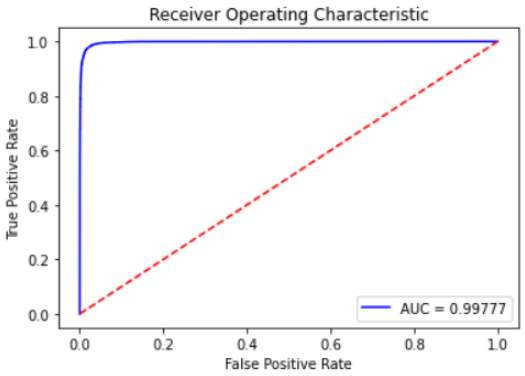

Fig. 5. ROC curve of the evaluation dataset

VII. Future Work

Our future work will be more focused towards comparison of different state-of-the-art deep learning architectures with our custom CNN architecture. We will also compute the performance of our model by using different kernel dimensions to observe a trend in accuracy. Along with that we will also look into the performance levels by using augmented data and grayscale data to train our model and try to reason its behaviour on test dataset.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Westerlund, Mika. "The emergence of deepfake technology: A review." Technology Innovation Management Review 9.11 (2019).

[3] Vougioukas, Konstantinos, Stavros Petridis, and Maja Pantic. "End-to-end speech-driven facial animation with temporal gans." arXiv preprint arXiv:1805.09313 (2018).

[4] S. Nightingale, K. Wade, and D. Watson, "Can people identify original and manipulated photos of real-world scenes?" *Cognitive Research: Principles and Implications*, pp. 2–30,2017

[5] Chang, H.T.; Hsu, C.C.; Yeh, C.H.; Shen, D.F. Image authentication with tampering localization based on watermark embedding in wavelet domain. Opt. Eng. 2009, 48, 057002.

[6] Frank, Joel, et al. "Leveraging frequency analysis for deep fake image recognition." International Conference on Machine Learning. PMLR, 2020.

[7] Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[8] "140k Real and Fake Faces" https://www.kaggle.com/xhlulu/140k-real-and-fake-faces , 2020

[9] Hsu, Chih-Chung, Yi-Xiu Zhuang, and Chia-Yen Lee. "Deep fake image detection based on pairwise learning." Applied Sciences 10.1 (2020): 370.

[10] Tai Do Nhu, In Na, and S.H. Kim. Forensics face detection from gans using convolutional neural network, 2018.