Grenze International Journal of Engineering and Technology, Jan Issue

A Hybrid Model of Convo-GAN to Detect Fake Images

Sangeeta Saha[1] and Bhawana Rudra[2]
[1-2] National Institute of Technology, Department of Information technology, Karnataka, INDIA
Email: sangeetasaha1025@gmail.com[1], bhawanarudra@nitk.edu.in[2]

***Abstract*—With advancements in the field of Deep Learning, it has become easy to generate face swaps, thereby creating fake images which look extremely realistic, leaving few traces which cannot be detected by bare human eyes. Such images are known as 'DeepFakes' that can be used to create a ruckus and affect the quality of public discourse on sensitive issues, defame an individual's profile, create political distress, blackmail a person or envision fake cyber terrorists. This paper proposes methods to detect fake images with the help of hybrid models having Convolutional Neural Network with Error Level Analysis, Gated Recurrent Unit neural network, Long Short Term Memory recurrent neural network and Generative Adversarial Network respectively. The 2019 'Real and Fake Face Detection' dataset from Kaggle [7] is used to train the models and by experimentation we are able to prove that the combined model of Convolutional Neural Network and Generative Adversarial Network outperforms other models.**

***Index Terms*— DeepFake, Convolutional Neural Network, Long Short Term Memory, Error Level Analysis, fake image detection, Gated Recurrent Unit, Generative Adversarial Network.**

I. INTRODUCTION

Manipulation of images has been a prevalent phenomenon since the advent of photography. There is an immense number of artifact platforms and advanced editing mechanisms that play a huge role in this process. Though it all began with animations and commercial requirements for entertainment purposes, with the advancement in the field of technology, eventually fake multimedia became a common issue, especially with the advent of the DeepFakes [8] that began to defame famous personalities.

Deepfake, i.e., the combination of 'Deep learning' and 'fake' images, is a technique from which one can synthesize a human image with the help of Artificial Intelligence. It is used to combine and superimpose the existing image into the source image. The issue with DeepFakes is that there lies the danger of false propaganda believed as truth. The issue does not only lie in a single DeepFake, but in the hostile power of the resources to create an entire ensemble of DeepFakes, beginning with a core 'false flag' event, and then quickly followed by well-known news anchors and other celebrities or politicians confirming the false events, rioting in the streets, etc. If such an ensemble is unleashed at a critical moment in time, the effect could precipitate to create even a worldwide disaster.

Artificial Intelligence creates DeepFakes and hence, can also be used to detect them. With the help of Deep Learning methodologies, it is now possible to build a model using the existing pictures of a person, that would help detect fake images of the same person, using various real-time face detection algorithms. The backbone of a DeepFake is the deep neural network that is trained on faces and is made to output the facial expressions of the source to the target after appropriate post-processing, resulting in a high level of realism. A Convolutional

Grenze ID: 01.GIJET.7.1.10

Neural Network (CNN) [8] can be used for this process, that tries to reconstruct the input image by learning a lower dimensional representation of the input, which is later used to swap the faces. An encoder is built in order to encode all the pictures using CNN and is used to detect face angles, skin tones, facial expressions, lighting and other information that is important to reconstruct the picture. Later a decoder is used to reconstruct the image.
A Generative Adversarial Network (GAN) [2] can also be used to detect fake images. In GAN, we have a deep network discriminator which is a CNN classifier that is used to distinguish whether facial images are original or created by the computer. When real images are fed to this discriminator, the discriminator trains itself in order to recognize the real images better, and when created images are fed into the same discriminator, it is trained in order to create more realistic images. Also, it can handle partially obstructed faces, i.e., when the face is partially blocked by a hand or the image looks blurry. In such cases, GAN can mask out the obstructed area in the created face and use the part in the target image instead. Even though GAN is powerful, it takes a long time to train the model and also requires a high level of expertise to make it right.

II. Literature Survey

As far as the progress in the field of GANs is concerned [2], an up gradation from blurred to much more realistic faces has taken place in the past few years. In order to create a DeepFake all that is needed is a few thousand of frames consisting of training data that would include the source face as well as the target face [9], [10]. When the training is over, a trained swap face network is created. As a result, the process that starts with the trained network ends with the image we wish to modify. This might take from minutes to hours depending on the complexity of the image that has to be altered [11], [12], [13].
Y. Liao, Y. Wang and Y. Liu [1] used a machine learning technique for DeepFakes detection that consisted of autoencoders and had three different phases. The first phase consisted of the encoder, the second phase was the actual encoding, which they called the bottleneck, and the third phase consisted of the decoder. The encoder was responsible for taking as an input, a high dimensional image, which it compressed down to a representation that was semantically meaningful. The responsibility of the decoder was to reconstruct the encoded input image in order to make an approximation for the same. This approach was a part of unsupervised learning of a semantically meaningful representation of the input image in order to compare the differences between the input and the output images respectively. This provided an error signal which could be used to update the weights in the network during training. The encoder was the common part in both the faces. As a result, it could share weights and learn representations better. The decoders were made to train individually on two different persons. During the production, the first person was taken and encoded, after which the decoder was used on the second person in order to reconstruct the input image that could provide the effect of swapping the faces across the two different images.
Ian Goodfellow [2] presented a piece of work based on the Generative Adversarial Networks or GANs. This was able to provide new ways in which one could train deep neural networks and generate new synthetic data for the purpose of training the deep learning models in order to detect DeepFakes. Zhang et al. [3] proposed that classifiers are able to provide a poor generalization among the GAN models. He presented an AutoGAN model that could generate images. The model consisted of a generator and a discriminator. It could successfully detect GAN images using hand crafted co occurrence features and models for anomaly detection. This was used on images that had faces, where it isolated the faces with the help of face detector models. The work concentrated on GANs generated images that were detected using GANs itself.
Rossler et al. [4] presented methods for the detection of facial manipulations, with the help of CNN classifiers that classified DeepFakes using the same model which was used while generating them. Marra et al. [5] followed a similar approach using image classifiers. Xu et al. [6] presented the approach where the convolution features were extracted in the section of the encoder. Feature extraction was performed on a per frame basis from each raw image. The model was able to generate a caption that encoded as a sequence of total encoded words. The decoder used the LSTM network that produced the caption by generating each word on a context vector. The convolutional features were implemented using a face detector, only on selected portions of the image, i.e., the face, and not on the entire frame.

III. Proposed Methodology

A. Convolutional Neural Network

Convolutional Neural Network (CNN) is a deep learning network that is generally used in order to classify images. CNN uses convolving filters that are predefined, for the purpose of identifying patterns in images and

parts of various objects. It is then trained to detect complete objects like animals, human beings, etc. CNN layers, in their lowest levels, learn to detect edges and corners of the input images. It then learns to detect parts of objects in the next layer followed by successful detection of the entire object in the last layer. CNN consists of the input layer which is responsible for resizing the input image to a given fixed size and to normalize the pixel intensity values of the same, followed by the convolution layer where the process of convolving and subsampling of the image takes place. Extraction of features like edges in the image with the help of specific filters occur in this layer. The next layer is the max pooling layer where the maximum number of operations on the output of the convolution layers is performed in order to retain significant information regarding the patterns and discard the insignificant details about the same. Then comes the fully connected layer that is the final logistic regression layer which maps visual features to desired output functions. The last layer is the output layer that contains the probabilities of each of the classes of image inputs.

B. Error Level Analysis Error

Level Analysis (ELA) is generally used to detect manipulated images by observing the way in which errors propagate when an image is saved as a JPEG, where on every save, some amount of error is added to each pixel of the image, that eventually approaches a maximum error level in the form of a horizontal asymptote. It is considered to be one of the techniques that is used to detect that an image has been manipulated by storing it at a particular level of quality and later calculating the difference from that level of compression. At the time of first saving a JPEG image, it is compressed for the first time. If it is resaved, it is compressed again. As a result, it shows that the original image was first saved when it was clicked with the help of a digital camera and later has been compressed again with the help of an editing software. The image might look the same to the bare eyes, but with the help of this approach, the difference is clearly visible. The technique calculates the average difference of the quantization or luminance value and the chrominance value, as the initial image taken with the help of digital cameras have a high value for ELA, that goes on decreasing gradually with subsequent resaves, thereby increasing the potential error rates.

C. Long Short Term Memory

Long Short Term Memory (LSTM) is a deep learning recurrent neural network that is used in the classification or prediction of temporal data, as this network has the capability to deal with exploding and vanishing gradient problems. LSTM is used in order to solve one of the major issues of recurrent neural networks, i.e., long term dependencies. LSTM helps in the modelling of temporal sequences and their long range dependencies with a better accuracy as compared to the conventional recurrent neural networks. The LSTM consists of special units known as memory blocks in the recurrent hidden layer. These memory blocks contain memory cells with self connections that store the temporal state of the network with the help of special multiplicative units called gates that are used to control the flow of information. Each memory block contains an input gate, an output gate and a forget gate. The input gate helps in controlling the flow of input activations into the memory cell whereas the output gate helps in controlling the output flow of cell activations into the rest of the network. The forget gate addresses one of the major weaknesses of the LSTM model by preventing it from processing continuous input streams that are not segmented into subsequences. It is used to scale the internal state of the cell before adding it as an input to the cell through the self recurrent connections, as a result of which, the cell's memory is reset.

D. Gated Recurrent Unit

Gated Recurrent Unit (GRU) is a gating mechanism in RNNs. GRU consists of two gates, i.e., reset and update gates respectively. The output gate is absent here, unlike LSTM. GRU uses less memory and is therefore, more efficient and less complex than other RNN mechanisms like LSTM, as it uses less training parameters and has lesser number of gates in each unit as compared to LSTM. As a result, it executes faster and trains faster than LSTM.

E. Generative Adversarial Network

Generative Adversarial Network (GAN) proposes a generative model that is meant for natural images which evolves to generate more and more realistic looking data, because of the coupling with an adversarial network. When we do not have a lot of data to solve a particular problem, we can use adversarial networks to generate more data, instead of resorting to the tricks like data augmentation. GANs are the first generative algorithms to give convincingly good results as there is an array of auxiliary reasons for the same, which includes the fact that GANs form a powerful approach in unsupervised learning and that it can generate high quality synthetic data. Also, adversarial training tends to produce sharper and more discrete outputs and not blurry averages, unlike

MSE. This has led to a variety of applications such as super resolution GANs to perform better than MSE and other loss functions.

F. Fake image detection using Convolutional Neural Network

A total of 2000 images are taken as input, 1000 of which are real and 1000 of which are fake. The entire dataset is split into train and test sets, with 80 percent of the real and fake images used for training and the remaining 20 percent for testing respectively. This dataset is used on the CNN model. The CNN consists of a convolutional layer with kernel size of 5x5 and the number of filters as much as 32. The second layer CNN consists of a convolutional layer with the size of the kernel of 5x5 and the number of filters as many as 32, and a Max Pooling layer with a size of 2x2. Second convolutional layer uses the glorot uniform initializer kernel and the ReLU activation function to create existing neurons. The convolutional layer makes a selection so that it can receive useful signals from input data. After that, a dropout of the MaxPooling layer of 0.25 is added in order to prevent overfitting. The next layer is the fully connected layer with the number of neurons as many as 256 and the ReLU activation function. After the fully connected layer, a dropout of 0.5 is added to prevent overfitting. The output layer uses a softmax activation function. RMS Propagation optimizer is used, that is one of the adaptive learning rate methods with a given number of epochs and batch size respectively.

G. Fake image detection using a hybrid model of Error Level Analysis and Convolutional Neural Network

The same dataset is taken and each of the images are converted first to a picture obtained after the Error Level Analysis performed on it. Converting raw data to ELA result images is a method which is used to increase the training efficiency of the CNN model. This efficiency can be achieved because of the image results ELA contains no exaggeration of information like the original image. Features generated by ELA images are already focused on the part of the image that has a level error above limit. Also, the pixels in the image after performing the ELA, tend to be very similar in color contrast with the nearby pixels, so training the CNN model became more efficient. After that, the image size changes. Next step is to normalize the data by dividing each RGB value with the number 255 to perform normalization, so that the CNN converges faster, i.e., reach a global minimum of loss value, because the value of each RGB only ranges between 0 and 1. Next the label on data is turned, where 1 represents tampered and 0 represents real into categorical values.
The entire dataset is then split into train and test sets, with 80 percent of the real and fake images used for training and the remaining 20 percent for testing respectively. This dataset is used on the same CNN model discussed in section III(F). In the architecture used, only two convolutional layers are needed, because the results are generated from the conversion process into an ELA image that can accentuate important features to know whether an image is original or has been modified.

H. Fake image detection using a hybrid model of Convolutional Neural Network and Long Short Term Memory

Again the same dataset is split into train and test sets, with 80 percent of the real and fake images used for training and the remaining 20 percent for testing respectively. This dataset is used on the CNN model discussed in section III(F). Now the data that is output from the output layer of the CNN is passed to a LSTM model for further training. The LSTM model consists of two hidden recurrent layers with 128 units each and the tanh activation function is used to create existing neurons. A dropout of 0.2 is added in order to prevent overfitting. SGD optimizer is used, that is one of the adaptive learning rate methods.

I. Fake image detection using a hybrid model of Convolutional Neural Network and Gated Recurrent Unit

The same dataset is split into train and test sets, with 80 percent of the real and fake images used for training and the remaining 20 percent for testing respectively. This dataset is used on the CNN model discussed in section III(F). Now the data that is output from the output layer of the CNN is passed to a GRU model for further training. The GRU model consists of two hidden recurrent layers with 128 units each and the tanh activation function is used to create existing neurons. A dropout of 0.2 is added in order to prevent overfitting. SGD optimizer is used, that is one of the adaptive learning rate methods.

J. Fake image detection using a hybrid model of Convolutional Neural Network and Generative Adversarial Network

The same dataset is split into train and test sets, with 80 percent of the real and fake images used for training and the remaining 20 percent for testing respectively. This dataset is used on the CNN model discussed in section III(F). Now the data that is output from the output layer of the CNN is passed to a GAN model for further training. The GAN model consists of two hidden layers with 128 and 64 units respectively. The relu and sigmoid activation functions are used to create existing neurons. A dropout of 0.25 is added in order to prevent

overfitting. RMS Propagation optimizer is used, that is one of the adaptive learning rate methods with a given number of epochs and batch size respectively.

IV. Experimental Result

A. Experimental Dataset

The 2019 'Real and Fake Face Detection' dataset from Kaggle is used, that consists of 2000 images, 1000 of which are real and 1000 of which are fake. A CSV file is created that contains two columns, where the first column is the path of each of the images and the second column contains a '0' if the image is real or a '1' if the image is fake.

B. Experimental Environment

We carried out the experiment by creating ipynb files, coding in Python language. The files had been executed on Google Colab setting runtime hardware accelerator to Tensor Processing Unit (TPU) in order to increase the speed of execution as a huge number of simulations and epochs were needed for each dataset to obtain higher accuracy.

C. Experimental Result

From Table I, we can conclude that the accuracy is improved when Error Level Analysis is used in the pre processing step prior to training the dataset on the Convolutional Neural Network, and that the accuracy is even more improved when the training is performed on the Convolutional Neural Network followed by the Long Short Term Memory Recurrent Neural Network or Gated Recurrent Unit. But the accuracy is highest when Generative Adversarial Network is used. The accuracy is improved with Error Level Analysis because of the image results ELA contains no exaggeration of information like the original image. Features generated by ELA images are already focused on the part of the image that has a level error above limit. Also, the pixels in the image after performing the ELA, tend to be very similar in color contrast with the nearby pixels, as a result of which, training the CNN model became more efficient. The accuracy is improved further with the use of LSTM or GRU for training after CNN, because LSTM or GRU both are able to detect spikes in variations in each of the images due to their ability to deal with temporal data better. In such cases where LSTM and GRU provide the same accuracy, it is always better to use GRU as GRU uses less memory and is therefore, more efficient and less complex than other RNN mechanisms like LSTM, the reason being, GRU uses less training parameters and has lesser number of gates in each unit as compared to LSTM. As a result, it executes faster and trains faster than LSTM. With the use of GANs, the accuracy is further improved as it can generate high quality synthetic data and adversarial training tends to produce sharper and more discrete outputs. Moreover, GANs are used in most of the cases to create such fake images, and as a result, are able to detect the same, better than other neural networks.

TABLE I. ACCURACY COMPARISON OF MODELS

Model	Accuracy (in percent)
Convolutional Neural Network	50.76
Hybrid model of Error Level Analysis and Convolutional Neural Network	85.68
Hybrid model of Convolutional Neural Network and Long Short Term Memory	90.35
Hybrid model of Convolutional Neural Network and Gated Recurrent Unit	90.35
Hybrid model of Convolutional Neural Network and Generative Adversarial Network	98.69

V. Conclusion

This paper proposes approaches to detect the visual forgery of images with the help of convolutional neural networks in combination with other methods like error level analysis, and recurrent neural networks like LSTM, GRU and GANs. By experimentation, we are able to conclude that the accuracy with which a fake image is detected increases when a hybrid model is used, that includes other methods along with a training phase involving the use of CNN, instead of using only CNN, and that when GANs are used in combination with CNNs, the accuracy obtained is the highest. The future scope of this work includes performing similar operations on videos instead of just images, that would not only include a still picture, but also a sequence of frames with multiple movements captured in each frame.

REFERENCES

[1] Y. Liao, Y. Wang and Y. Liu, “Graph regularized auto-encoders for image representation”, IEEE Trans. Image Process., vol. 26, no. 6, pp. 2839–2852, Jun. 2016.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative adversarial nets”, in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[3] X. Zhang, S. Karaman and S.-F. Chang, “Detecting and simulating artifacts in GAN fake images”, 2019, arXiv:1907.06515.

[4] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, “FaceForensics++: Learning to detect manipulated facial images”, in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1–11.

[5] F. Marra, D. Gragnaniello, D. Cozzolino and L. Verdoliva, “Detection of GAN-generated fake images over social networks”, in Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR), Apr. 2018, pp. 384–389.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention”, in Proc. Int. Conf. Mach. Learn., Jun. 2015, pp. 2048–2057.

[7] Kaggle—Real and Fake Face Detection. Accessed: Sep. 11, 2020. [Online]. Available: https://www.kaggle.com/ciplab/real-and-fake-facedetection

[8] Mohammed Farukh Hashmi, B. Kiran Kumar Ashish, Avinash G. Keskar, Neeraj Dhanraj Bokde, Jin Hee Yoon and Zong Woo Geem, “An Exploratory Analysis on Visual Counterfeits Using Conv-LSTM Hybrid Architecture“, in IEEE, May. 2020.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies”, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2008, pp. 1–8.

[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation”, 2017, arXiv:1710.10196. [Online]. Available: http://arxiv.org/abs/1710.10196

[11] V. Conotter, E. Bodnari, G. Boato, and H. Farid, “Physiologically-based detection of computer generated faces in video”, in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2014, pp. 248–252.

[12] J. Hyeon Yoo, “Large-scale video classification guided by batch normalized LSTM translator”, 2017, arXiv:1707.04045. [Online]. Available: http://arxiv.org/abs/1707.04045

[13] The Outline: Experts Fear Face Swapping Tech Could Start an International Showdown. Accessed: Dec. 31, 2019. [Online]. Available: https://theoutline.com/post/3179/deepfake-videos-are-freaking-expertsout?zd=1zi=hbmf4sv