International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 1349-1353 www.ijaem.net ISSN: 2395-5252

A Study on Detection of GAN Generated Fake Images over Social Media

VandhanaS.[1] ,VishnupriyaJ.[2],AthiraS.[3], SarithaC.[4], Reshma Mohan A. S.[5]

[5]*Assistant professor , UKF College of Engineering and Technology*
[1,2,3,4]*UG students , UKF College of Engineering and Technology*

Submitted: 01-06-2021 Revised: 14-06-2021 Accepted: 16-06-2021

ABSTRACT -Generative adversarial networks (GANs) can be used to generate a photo-realistic image from a low-dimension random noise. Such types of fake images with unnecessary content can be used on social media networks, it can cause severe and difficult problems. So the aim is to successfully detect fake images, an effective and efficient image forgery detector is necessary. Recent advances in Generative Adversarial Networks (GANs) have mainly shown increasing and immediate success in generating photorealistic images. But they can also raise challenges to visual forensics as well as model attribution. Image to image translation based on GAN is one of the dangerous to learn a mapping between images from a source domain and images from a target domain. The enormous application prospect, including image and vision computing, video and language processing, etc. Besides, the paper also tells that the background of the GAN and its theoretic models and also explains that how to detect GAN generated fake images over social media.

Keywords - GAN, Convolutional neural networks, Image to image translation

I. INTRODUCTION

Recently, deep learning-based generative models, such as variational auto encoders and generative adversarial networks (GANs), have been used to synthesize the photo-realistic images partially or whole content of an image as well as video.

For instance, the cycle GAN can be used to synthesize the fake face image in a pornography video [4]. Furthermore, the GANs can also create speech video with the synthesized facial content or expressions of any famous politician, or any others which becomes causing severe problems to the society, political, and other such activities. Therefore, an effective and immediate fakeface image detection technique is compulsory. In this paper, our previous study is extended to recognize generated fake images effectively and efficiently.

Since deep neural networks have been widely used in areas such as various recognition tasks. We can also adopt such a deep neural network to detect fake images generated by the GANs. Here, we are studied that the deep learning based approach for fake image detection using supervised learning .Also, fake image detection has been treated as a binary classification problem or model (i.e., fake or real image). For instance, in this case the convolution neural network (CNN) network was used to develop the fake image detector.

Fig 1 : Generative Adversarial Networks

Generative Adversarial Networks consists of two models; generative and discriminative.

The Discriminative Model

The discriminative model operates like a normal binary classifier that's able to classify images into different categories. It determines whether an image is real and from a given dataset or is artificially generated.

The Generative Model

The generative modeling is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or pattern in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset. A generative model may be able to sufficiently summarize this data distribution and then be used to generate new variables that plausibly fit into the distribution of the input variables.

This research aim to develop which generated and detect GAN generated fake images. Also tried to improve the accuracy of detection of GAN by preprocessing and segmentation feature extraction. In this work we analyze performance of number of learning - based method in the detection of image -to -image translation.

The generative adversarial network is an approach to generative modeling using deep learning methods such as convolutional neural network. It is a class of machine learning framework, it has two neural network, Generative and Adversarial. Generative which means that there is a network that is constantly generating new data and the second one is Adversarial which means that it involves two network simply means an order of data that is going ahead and generate new data.

GANs consist of two networks, a Generator and Discriminator. They both play an adversarial game where the generator tries to fool the discriminator by generating data similar to those in the training set. The discriminator tries not to be fooled by identifying fake data from real data. They both work simultaneously to learn and train complex data like audio, video or image files. The generator model generates images from random noise and then learns how to generate realistic images. Random noise which is input is sampled using uniform or normal distribution and then it is fed into the generator which generates an image. The generator output which are fake images and the real images from the training set is fed into the discriminator that learns how to differentiate fake images from real images. The output is the probability that the input is real.

II. EXPERIMENTAL PROCEDURE

We compare a number of methods for the detection of fake images.Some are the methods proposed for detection of specifically designed computer graphic images and other are based of CNN architectures.For that first we have trained our adversarial nets using our dataset.For each category dataset includes both real and fake images.For example,here we are taking original images of zebra and horse used to train GAN and corresponding fake images are produced by GAN.

Training phase has two main sub parts.Part one is train the discriminator and freeze the generator,freezing means training as fake,the network does only forward pass.Training of generator while freezing the discriminator is the next part.To train a GAN first we have define our problem.Here we want to generate fake images.Next step is to define architecture of GAN,ie the generator and discriminator are convolution neural networks.Then train the discriminator using orginal images to predict it as real.Then generate fake images using generator and let the discriminator correctly predict them as fake.Train the generator using the output of discriminator.That predictions of discriminator are used for training of generator.Repeat these steps for several times to get accurate results.

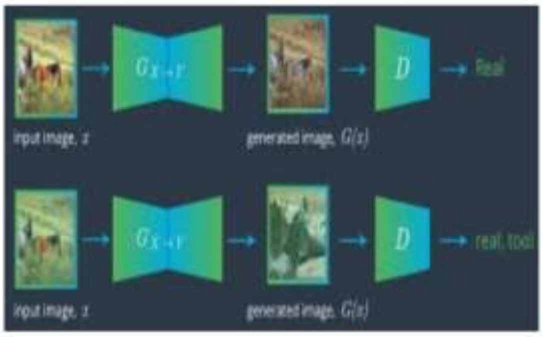

Fig 2 : Working of GAN architecture

GANs are formulated as a minimax game,where discriminator is trying to minimize its loss and generator is trying the loss of discriminator in distinguishing data.Working of GAN can be mathematically expressed as follows,

V=(D,G)=Ex ~ p data(x)[log D(x)]+Ez ~p data(z) [log(1-D(G(Z)))] where,

G-Generator , D-Discriminator, X-Sample from real data, Z-Sample from generator

Pdata(x)-Distribution of real data , pdata(z)-Distribution of generator , D(X)- Discriminator network G(Z)-Generator network

To detect GAN image we compute co-occurrence matrices on the RGB channels of an image.These matrices are usually computed an image residuals by passing the image through many filters and then obtaining difference.However in this paper we compute co-occurrence matrices directly on the image pixels on each of red,green and blue channels and pass them through a convolutional neural network,there by allowing the network to learn important features from the co-occurrence matrices.The first step is to compute co-occurrence matrices on RGB channels to obtain 3x256x256 tensor.This tensor is then passed through a multi layer deep convoloutional neural network;Convo layer+ReLu layer+max pooling layer+dense layer+sigmoid layer.Input layer in CNN should contain image data.Image data is represented by three dimensional matrix.Convo layer is sometimes called feature extraction layer because features of image are get extracted within this layer.Convo layer also contains ReLu activation to make all negative value to zero.ReLu refers to rectifier unit,the most commonly deployen activation function for the outputs of CNN neurons.Pooling layer is used to reduce the spatial volume of input image after convolution layer.Dense layer is the regular deeply connected neural network layer.A sigmoid layer applies a sigmoid function to the input such that the output is bounded in the interval (0,1) .Fully connected layer involves weights,biases and neurons.It is used to classify images between different category by training .Output layer contains the label which is in the form of one hot encoded.

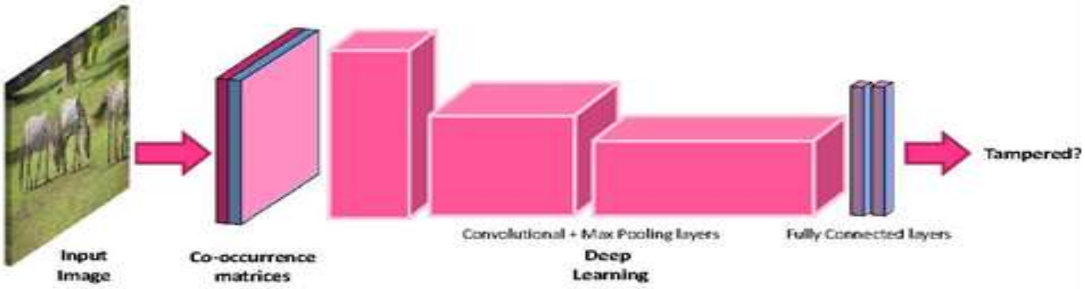

Fig 3 : Convolutional Neutral Network

DATASET

CycleGAN dataset:In this experiment 30000 natural images in a format we downloaded from the cyclegan dataset. We randomly selected 1500 natural images for our experiment.This dataset contains unpaired image to image translations of various objects and scenes such as horses to zebra,summer to winter,images to paintings (Monet,VanGough) and others that were generated using cycle consistent GAN frame work. We followed instructions provided by authors as detailed in https://githhub.com/junyanz/CycleGAN. Dataset includes more than 36k 256x256.

III. RESULT AND DISCUSSION

In order to train and validate the detectors under comparison. We built a very large dataset of samples of different sections from image -to -image translation. The dataset consists both the real and fake images. For example, horse to zebra subset includes all the original images of horses and zebras used to train the GAN and the corresponding fakes (horses and zebras respectively) generated by the GAN itself once trained. First we group concerns the translation of natural images and includes horse to zebra large collection. Second group is related to the generation of images from labeled maps of cityscapes in this case only the real photos and generated once are include in the dataset.

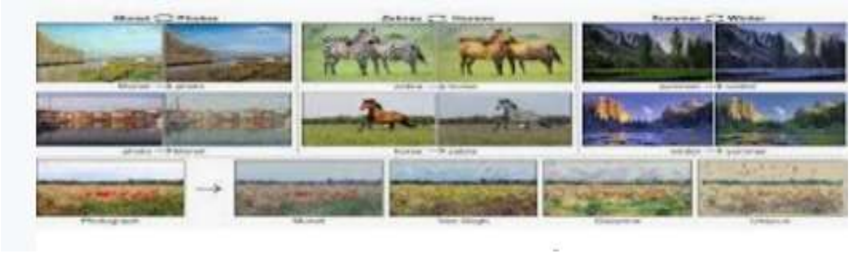

Fig 4 : Categories of image to image translation

IV. CONCLUSION

We have presented a study on the detection of images manipulated by GAN based image-to-image translation. Then Several detectors perform very well on original images, but some of them show dramatic impairments on Twitter-like compressed images. Then Robustness is better preserved by deep networks, especially Xception Net, which keeps working reasonably well even in the presence of training-test mismatching. Future research, besides extending the analysis to more manipulations and detectors, will study cross method performance, possibly after transfer learning, w.r.t. other synthetic image generators. Moreover, we will test the performance in real world scenarios involving different social networks Finally , we can understand that GAN is a amazing technology in society and also in medical field .It will helps to detecting fake data from real data.Thereby, we have to analyse the original images quickly. So we can easily detect the magic behind the fake data.

In this paper, we proposed a novel method to detect GAN generated fake images using a combination of pixel co-occurrence matrices and deep learning. Co-occurrence matrices are computed on the color channels of an image and then trained using a deep convolutional neural network to distinguish GAN generated fake images from real ones. Experiments on two diverse GAN datasets show that our approach is both effective and generalizable. In future, we will consider localizing the manipulated pixelsin GAN generated fake images.

REERENCES

[1]. 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.

[2]. V. Schetinger, M. Oliveira, R. da Silva, and T. Carvalho, “Humans are easily fooled by digital images,” Computers & Graphics, vol. 68, pp. 142–151, 2017.

[3]. D. Cozzolino, D. Gragnaniello, and L.Verdoliva, “Image forgery detection through residual-based local descriptors and block-matching,” in IEEE Conference on Image Processing (ICIP), October 2014, pp. 5297–5301.

[4]. S. Fan, T.-T. Ng, B. Koenig, J. Herberg, M. Jiang, Z. Shen, and Q. Zhao, “Image visual realism: From human perception to machine computation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, in press, 2017.

[5]. S. Lyu and H. Farid, “How realistic is photorealistic?” IEEE Transactions on Signal Processing, vol. 53, no. 2, pp. 845 – 850, 2005.

[6]. R. Wu, X. Li, and B. Yang, “Identifying computer generated graphics via histogram features,” in IEEE ICIP, 2011, pp.1933–1936.

[7]. S. Dehnie, H. Sencar, and N. Memon, “Digital image foren- sics for identifying computer generated and digital camera images,” in IEEE ICIP, 2006, pp. 2313–2316.

[8]. A. Dirik, S. Bayram, H. Sencar, and N. Memon, “New features to identify computer generated images,” in IEEE ICIP, Oct 2006, pp. IV–433–IV–436.

[9]. A. Gallagher and T. Chen, “Image authentication by detecting traces of demosaicing,” in IEEE CVPR Workshops, June 2008, pp. 1–8.

[10]. J.-F. Lalonde and A. Efros, “Using color compatibility for assessing image realism,” in IEEE ICCV, Oct 2007, pp. 1–8.