Landmark DeepFake Detection

Shin Woochul
Dept. of Economics
Seoul National University
swc0620@snu.ac.kr

Sung Seokrin
Dept. of Economics
Seoul National University
ssr0827@snu.ac.kr

Park Jinsol
Dept. of Liberal Studies
Seoul National University
solvely95@snu.ac.kr

Abstract

In this project, we discuss three models for DeepFake detection: Steganalysis, MesoNet, VGGNet. We analyze the performance of models using whole face images, and cropped parts of face images (left eye, nose, and mouth). We discuss the reason why each model shows such performance, and suggest future improvements that can be made to DeepFake detection methods.

1. Introduction

Fake videos, also known as *DeepFake*, is a serious problem deteriorating social trust, safety, and security. Recent advances in deep learning have made it significantly easier to generate sophisticated fake videos. It is clear that the importance of detecting manipulation in images and videos to counter DeepFake is growing faster and faster.

Since majority of the target of fake videos are politicians, actors, and other world leader, generating fake videos might spread incorrect information or rumors through the society. Spurious words shown from the fake video of corporation leader may affect significantly in the stock market. Fake video of military leader, or other nation leader may have worldwide impact regarding each national interest. Therefore, fake videos pose significant threat to our society, security, and also our democracy.

The motivation of this project is to explore efficient algorithm favorable to DeepFake detection. Noticing that many previous experiments take the input face images as whole, we hypothesize that reducing the input facial data to regional landmarks may increase models' performance. We use both whole face dataset and cropped images of the face (eyes, nose, and mouth) to see if there is any change in performance. Performance metric is "Accuracy", regarding validation and test dataset. After analyzing performance, we discuss the reason behind the preformance of each model and suggest further improvements that can be made to data preprocessing and detection models.

2. Dataset

2.1. Whole Face Images

Figure 1. Sample whole face images from dataset

We use Kaggle dataset [9] consisting of all 70k real faces from the Flickr dataset collected by Nvidia, as well as 70k fake faces sampled from the 1 million fake faces generated by StyleGAN that was provied by Bojan. We randomly select 1,000 real and fake images respectively from this (total 2,000 images). The size of images is 256 pixels each.

Dataset is split into train and validation set by ratio of 0.9:0.1. Additional 100 real and fake images respectively (total 200 images) are used as test set.

2.2. Cropped parts

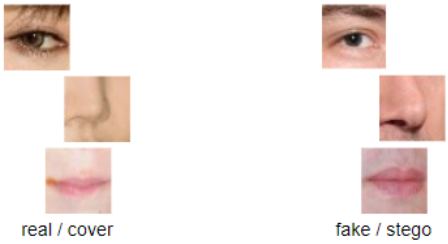

Figure 2. Sample cropped parts images from dataset

We crop images using MTCNN [2], extract left eye, nose and mouth. We randomly select 1,000 real and fake images respectively from Kaggle dataset. (total 6,000 images; 1,000 real and fake images respectively for 3 spatial parts)

Dataset is split into train and validation set by ratio of 0.9:0.1. Additional 300 real and fake images respectively (total 600 images; 100 real and fake images respectively for 3 spatial parts) are used as test set.

3. Background and related work

There has been many recent papers applying state-of-the-art steganalysis and deep learning networks to DeepFake detection. FaceForensics++: Learning to Detect Manipulated Facial Images [6] evaluated the overall performance of facial forgery detectors based on steganalysis features and on learned features. A forgery detection from steganalysis features achieved high performance on RAW input data, but showed relatively low performance for compressed videos such as low quality videos. Among forgery detection from learned features, XceptionNet, a network architecture using a traditional CNN trained on ImageNet based on separable convolutions with residual connections, achieved relatively high performance on both high and low quality images, as it benefits from its pre-training on ImageNet. In this respect, this paper offers insight into what state-of-the-art facial forgery detection methods can be tested in advance.

4. Detection methods and experiments

4.1. Steganalysis

Steganalysis is the study of detecting messages in files, images, or videos concealed by steganography. Steganalysis and DeepFake detection methods share analogous properties in that they both try to detect payload encoded to the original images. This includes, for example, checking altered pixel values in relations to neighbour pixels. We implement steganalysis model following typical machine learning process. First, we extract features from input images. Next, we train classifier with extracted features. Last, we predict which class test data belong using the trained classifier.

4.1.1 Feature Extraction

We extract features following Spatial Rich Model(SRM) by Fridrich *et al*. [4] [5] The key idea of SRM is "merging smaller submodels" to extract various features. The major weakness Steganalysis by Subtractive Pixel Adjacency Matrix(SPAM) [8] exposes is that it only caputres first and second Markov chain. In contrast, SRM sweeps through input image with multiple submodels. Constructing submodels as

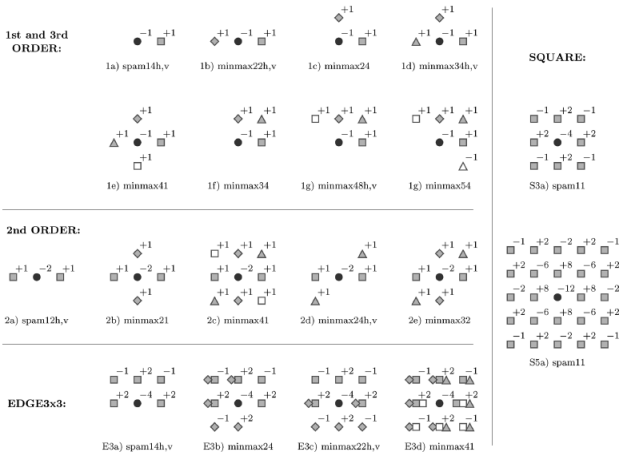

Figure 3. Submodels of residuals used in SRM. Each submodel represent relations between a pixel and its neighbouring pixels. For example, 1a) sapm14h,v represent first order Markov chain [4]

filters for convenience, we get the residuals of pixels representing various relations between neighbouring pixels (See Figure 3). The reason why SRM uses residuals is to make each submodel robust to noise. We quantize residuals with q to make residuals more sensitive to edges and textures.

Since merging multiple submodels results in high dimensions and poor computational complexity, SRM concentrates on decreasing dimensions and curbing residuals' dynamic range. This process includes two steps: truncation and co-occurence symmetrization. To reduce dimension of co-occurence matrix, we only focus on co-occurence of residuals along the horizontal and vertical directions, where residuals are within range of truncation $\{-T, T\}$. (*e.g.* $T = 2$ results in 4 dimensions, 625 elements) Then we use co-occurence symmetrization between co-occurence matrices of submodels to reduce number of co-occurence matrices. Thus constructed co-occurence matrices are to represent features of input image.

4.1.2 Ensemble Classifier

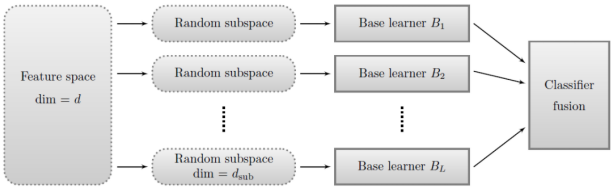

Figure 4. Structure map of Ensemble Classifier [3]

We use ensemble classifier as described in Kodovsky *et al*. [3] [5] Method used is bagging (bootstrap aggregating). From feature space extracted before, we sample

random subspace with replacement. We learn each base learner from each random subspace. Each base learner is trained with FLD (Fisher Linear Discrimminant). The goal of FLD is to maximize the difference of means of data of two classes and minimize variance of data of each class. We calculate covariance matrices, means and do eigendecomposition to compute the generalized eigenvector. This eigenvector is used as a tool for projection to lower dimension for classification. We use OOB(Out-Of-Bag) error to determine number of base learners L and dimensions of subspaces d_{sub}.

4.1.3 Performance

Table 1. Performance of Steganalysis model. (Accuracy in %)

	train accuracy	validation accuracy	test accuracy
whole face	98.1838	91.0891	97
all cropped parts	98.8764	45	55
left eye	99.4382	85	88
nose	99.1759	90	88
mouth	99.1816	84	86

DeepFake detection using train set as whole face and spatial region worked marvelously and showed high accuracy. However DeepFake detection using train set as spatial region showed relatively low scores. This is because there were less features extracted from the input image. We conclude that there is a trade-off between test accuracy and computational workload (represented by number of features).

4.2. MesoNet

MesoNet is deep neural network designed to detect image tampering inside the video. Since it focuses at the detection inside the video, main purpose of this algorithm is to detect it in more efficient and automatic way.

4.2.1 Basic structure

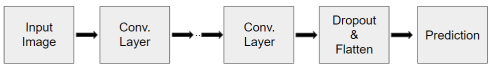

Figure 5. Basic structure of MesoNet

This model is basically formed with several convolutional layers. However, unlike ordinary deep learning network, it uses small number of layers as it is targeting the efficient detection algorithm which could shows reasonable performance even when processing videos. From the paper [1], authors propose several types of model, which are Meso4, MesoInception. In this project, we used Meso4 as baseline model to analyze the performance. As the name shows, Meso4 consists of 4 convolutional layers in total with dropout and flatten algorithms added. Thus, cropped face images are input to this model, which are processed via several convolutional layers that will labeling each input as deepfake image or real image. Then, generated classifier is used to validation set to detect and analyze the performance of the model.

4.2.2 Performance

Table 2. Performance of MesoNet. (Accuracy in %)

	test accuracy
whole face	70
left eye	63
nose	61
mouth	61

The chart above shows the performance of MesoNet with changing train sets, which are whole face, eyes, nose, and mouth. Data shows that there are huge difference in accuracy between models trained by 'whole face' and 'part of the face'. Therefore, We could figure out that amount of information is critical for detection accuracy. That is, using only landmark of face as a parameter of detecting video forgering is not enough to show reasonable performance.

4.3. VGGNet

One of the famous CNN architecture models is the VGGNet. It was noted at ILSVRC in 2014 along with GoogLeNet, but unfortunately it stayed at second place. However, the structure of VGGNet is relatively very simple compared to GoogLeNet, and thus, we thought it was suitable for us to understand the model, and make further modifications or preprocessing of images to allow this model to be a more effective deep fake detector.

4.3.1 Basic structure

First, let us go quickly over the structure of the VGGNet. Originally, as the research team mentioned in their paper "Very Deep Convolutional Networks for Large-Scale image Recognition" [7], the goal of VGGNet was to study the effect of the 'depth' of a network. Thus, they used a 3x3 receptive field and experiment on models with different depths. Spatial pooling is done by five max-pooling layers with a 2x2 pixel window and stride 2. The team mentioned that by using a stack of two 3x3 convolutional layer, it has an effective receptive field of 5x5, but has less parameters.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 6. Basic structure of VGGNet

4.3.2 Performance

Table 3. Performance of VGGNet. (Accuracy in %)

	test accuracy
whole face	51
left eye	45
nose	42
mouth	43

Using this VGGNet, our team ran the aforementioned datasets. Accuracy for model trained with whole face, eye, nose, and mouth showed 51, 45, 42, and 43 percent accuracy respectively. Like the other models and methods that our team used previously, it can be seen that training the model with whole face images showed the highest accuracy among the four cases. However, overall, the accuracy was lower in all aspects compared to other models, and we concluded that this was because VGGNet is not a deep fake-specific model. It is just a convolutional neural net model used for classification. Thus, by training and testing with VGGNet, our team learned that detecting deep fake images with mere CNN models is not enough, and therefore there is a need of well-formed and elaborate techniques to detect deep fake images.

5. Conclusion

We sought to improve DeepFake detection performance through reducing the input facial data to regional landmarks. However, it proved out that this landmark-approach did not boost performace. We learned and conclude that overall-viewing of whole facial images is more important. We further suggest three points for future improvements that can be made to DeepFake detection methods. First, assigning weights to each landmark and updating loss differently could make difference. Second, adding more landmakrs might enhance performance, Third, we hope that landmark-approach can be further developed to bring a notable performance in future studies.

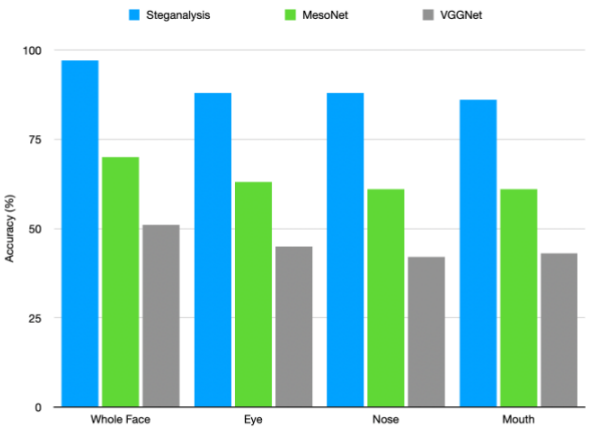

Figure 7. Performance of three DeepFake models

References

[1] Darius Afchar. Mesonet: a compact facial video forgery detection network. *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[2] ipazc. MTCNN. `https://github.com/ipazc/mtcnn`.

[3] Vojtech Holub Jan Kodovsky, Jessica Fridrich. Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on Information Forensics and Security*, 7(2):432–444, Apr 2012.

[4] Jan Kodovsky Jessica Fridrich. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, Jun 2012.

[5] Daniel Lerch. Aletheia. `https://github.com/daniellerch/aletheia`.

[6] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. 2019.

[7] Andrew Zisserman Simonyan, Karen. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[8] Jessica Fridrich Tomas Penvy, Patrick Bas. Steganalysis by subtractive pixel adjacency matrix. *Proceedings of the 11th ACM Multimedia & Security Workshop*, pages 75–84, Sep 2009.

[9] xhlulu. 140k Real and Fake Faces - 70k real faces (from Flickr) and 70k fake faces (GAN-generated). `https://www.kaggle.com/xhlulu/140k-real-and-fake-faces`.