Performance Analysis of Deepfake Detection Models : Under Dataset Consisting Mainly of Asians

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

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

Abstract

The main objective of our project is to estimate and compare the performance of deep fake detection algorithm, using dataset consisting mainly of asians. Since existing dataset is biased to western ethnics, analyzing the performance of each model regarding asian dataset could be benchmark to understand model's robustness. If possible, our plan is to suggest effective algorithm that can be applied to detect fake videos of various ethnics including asians. Reviewing several relevant papers, we are planning to exploit domain-specific information, and also to construct ethnic-specific detection algorithms.

1. Introduction

Fake videos, also known as *deep fakes*, 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. FakeApp, for example, is one of the free software that enables public to create fake videos by naively using those application.

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.

In this project, we will analyze performance of each fake video detection algorithm to various data set, consisting of asians. Performance metric is "Accuracy", regarding test data set and validation data set. Since existing dataset for such detection algorithm is biased to western ethnics, and also biased regarding gender, it is important to estimate the performance of those algorithm to the other ethnics, namely asian. After estimating performances, we plan to adjust those algorithm using various technic mentioned from relevant paper(will be explained at the section below), such as considering domain-specific information and XceptionNet.

2. Relevant papers

2.1. FaceForensics++: Learning to Detect Manipulated Facial Images

Rossler et al. [7] evaluated state-of-the-art facial forgery detection methods. First, they generated a large-scale input dataset of manipulated videos from two computer graphics-based methods (Face2Face and FaceSwap) and two learning-based methods (DeepFakes and NeuralTextures).

Next, they 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, which can be a good starting point of our work.

2.2. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics

Li et al. [6] pointed out drawbacks current DeepFake datasets are suffering from, and presented a new high-quality challenging DeepFake video dataset, Celeb-DF. They synthesized imperfect visual artifacts that can be found in the current datasets. These include low-quality synthesized faces, visible splicing boundaries, color mismatch, visible parts of the original face, and inconsistent

synthesized face orientations.

Then they generated Celeb-DF using an improved and refined DeepFake synthesis algorithm. For higher resolution of synthesized faces, more layers and dimensions are added to encoder and decoder models. For color match, colors of the training faces are randomly perturbed in each training epoch. For accurate face masks, an additional smoothness mask is created on facial landmarks.

This paper introduced shortcomings of the current DeepFake datasets and some advanced algorithms to correct those flaws of which we need to take account in our research.

2.3. Protecting World leaders Against Deep Fakes

Argawal et al. [2] pointed out that individuals exhibit relatively distinct patterns of facial and head movements when they speak. Using this, they describe a forensic technique that is designed to detect deep fakes of specific individuals, mainly focusing on high-profile individuals that are more prone to deekfake forgery.

The dataset is largely categorized into five parts; actual YouTube video clips, FaceForensics dataset, face-swap week fakes, lip-sync deep fakes, and puppet-master deep fakes. Using this, generative adversarial network (GAN) was trained on Deepfake architecture to swap faces.

Facial detection is done by calculating the correlation of 190 feature-pairs from 20 features (16 action units (AU) provided by OpenFace2, and 4 additional rotation / distance features). The visualization of the 190-dimensional features of individuals show that the POIs are well separated from each other. This proposes that mere action units and head movements can be used to discriminate between individuals. The robustness of this technique comes from the fact that the method is not vulnerable to pixel-level artifacts. Also, it is novel in that it only requires only authentic videos of a POI for training.

In such aspects, this paper gives insight on how to adjust algorithms in a more efficient way for our final project. To add, the context / individual specific training proposed in this paper eludes the possibility that ethnic-specific detection algorithms may be helpful.

3. Potential Datasets for Experiments

Our main goal is to analyze performance of DeepFake detection models, especially on asian datasets. To this end, we will use the dataset provided by collaboration of DACON, Data Science and AI Laboratory (DSAIL), Artificial intelligence Institute (AIIS), Money Brain, and National Information Society Agency (NIA) [1]. The 127GB dataset consists of real and fake classes, which will each be split into train and test sets. Also, the availability of large-scale datasets of DeepFake videos is an enabling factor in the development of DeepFake detection method. Thus we may use the UADFV dataset [8], the DeepFake-TIMIT dataset [5], the FaceForensics++ dataset, the Google DeepFake detection dataset [4], and the Facebook DeepFake detection challenge dataset [3].

References

[1] https://www.dacon.io/competitions/official/235655/data/.

[2] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes, 2019.

[3] Rian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, , and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset, 2019.

[4] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Vic tor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes detection dataset by google and jigsaw.

[5] Pavel Korshunov and Sebastian Marcel. Deepfakes: a new threat to face recognition? assessment and detection, 2018.

[6] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020.

[7] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva Christian Riess, Justus Thies, and Matthias Niebner. Faceforensics++: Learning to detect manipulated facial images, 2019.

[8] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses, 2019.