Deep Experiments on Deepfake Detection

Yebin Lee, Boyoung Han, and Juhyeon Jung,

I. Introduction

A. Background and Motivation

Deepfake is a portmanteau of 'deep learning' and 'fake'. Deepfake is a technology that creates a manipulated image by superimposing another image on a video through deep learning (Chawal, 2019). March 2022, a video clip of the Ukrainian president declaring his surrender to Russia was distributed, drawing much attention (Wakefield, 2022). However, the video is fake video utilized Deepfake technology. This technology can cause national, political and social problems.

B. Purpose

Hany Farid, Professor of California State University at Berkeley, pointed out that Ukrainan president surrender video was fake when he saw that the resolution was lowered to hide the distortion caused by the manipulation process (Metz, 2022).

This project is aiming to investigate the effect of image change on Deepfake detection problem. Deepfake detection performance is compared using original (color) images, images with color changes, and images with changes in saturation.

II. Related Work

Shad et al. (2021) implemented eight CNN models to detect deepfake images and make a comparative analysis. In this research, models are trained by dataset from Kaggle, which had 70,000 images from the Flickr dataset and 70,000 images produced by styleGAN. The eight CNN models utilized are as follows: DenseNet121, DenseNet169, DenseNet201, VGG16, VGG19, VGGFace, ResNet50, and custom CNN which was proposed by authors. The models are evaluated by five metrics: accuracy, precision, recall, F1-score, and area under the ROC (receiver operating characteristic) curve. Amongst all the models, VGGFace performed the best, with 99% accuracy. The second highest performance models were ResNet50 and DenseNet121 with 97% accuracy.

III. Data

This project intends to use the Deepfake Detection Challenge dataset from Kaggle (Dolhansky et al. 2019). In this project, we define the corresponding image as an Original Image (OI), an image with RGB values adjusted from the original image as an RGB Transformed Image (RTI), and finally an image with HSV values adjusted as an HSV Transformated Image (HTI).

Fig. 1: Original Image(OI) Example

Autors are with the Department of Data Science, Seoultech, Republic of Korea
E-mail: {yebin, byhan2253, jjh990307}@ds.seoultech.ac.kr
2022 Spring Semester, Neural Networks and Deep Learning

IV. Methodology

We will proceed with the experiment as the best model in our environment among VGGFace (Fig. 3), ResNet50 (Fig. 4), and DenseNet121 (Fig. 5), which had good performance in previous studies. From now on, we will configure input data with various conditions and see how each condition affects model performance.

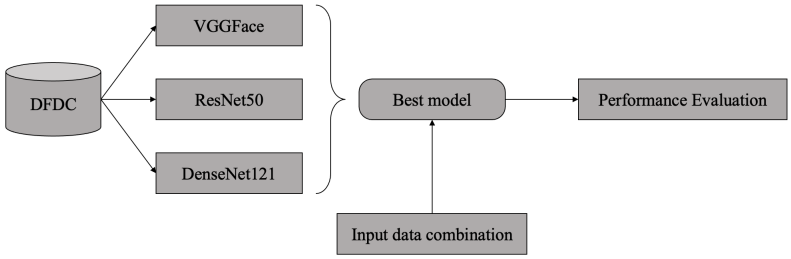

Fig. 2: Project Framework

Input data combination is like below:

1) Original Image(OI) : Full color Image, Data from Kaggle
2) RGB Transformed Image(RTI) : Modify RGB ratio to make black & white image
3) HSV Transformed Image(HTI) : Modify HSV ratio to make low resolution image
4) OI + RTI
5) OI + HTI
6) RTI + HTI
7) OI + RTI + HTI

V. Evaluation

In this project, we will use training accuracy, validation accuracy, test accuracy, loss curve and confusion matrix to evaluate models. Through this evaluation, we are going to find out how image change affects Deepfake detection performance.

Appendix A
Architecture of Models

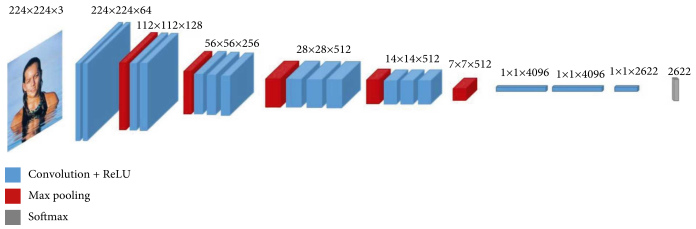

Fig. 3: VGGFace

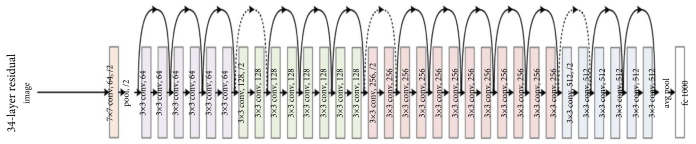

Fig. 4: ResNet50

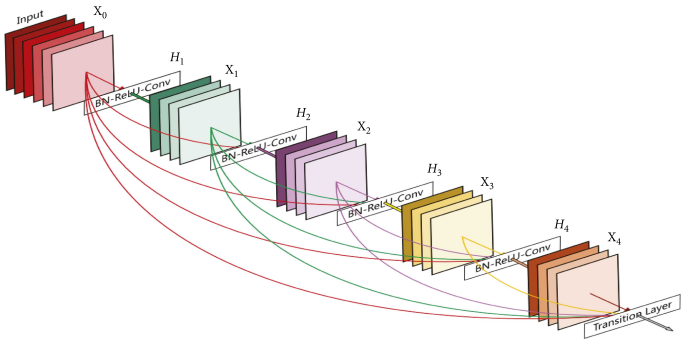

Fig. 5: DenseNet121

REFERENCES

[1] Wakefield, B. J. (2022, 18 maart). Deepfake presidents used in Russia-Ukraine war. BBC News. Geraadpleegd op 1 april 2022, van https://www.bbc.com/news/technology-60780142
[2] R. Chawla, "Deepfakes : How a pervert shook the world." international journal of advance research and development, vol 4, issue 6, p.4-8, 2019.
[3] Metz, R. (2022, March 16). Facebook and YouTube say they removed Zelensky deepfake. CNN Business. Retrieved April 1, 2022, from https://edition.cnn.com/2022/03/16/tech/deepfake-zelensky-facebook-meta/index.html.
[4] Shad, H. S., Rizvee, M., Roza, N. T., Hoq, S. M., Monirujjaman Khan, M., Singh, A., ... & Bourouis, S. (2021). Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network. Computational Intelligence and Neuroscience, 2021.
[5] Salpekar, O. DeepFake Image Detection.
[6] Deepfake Detection Challenge. Kaggle. (n.d.). Retrieved April 1, 2022, from https://www.kaggle.com/competitions/deepfake-detection-challenge
[7] Do, N. T., Na, I. S., Kim, S. H. (2018). Forensics face detection from GANs using convolutional neural network. ISITC, 2018, 376-379.
[8] Serengil, S. (2022, 20 januari). Deep Face Recognition with VGG-Face in Keras — sefiks.com. Sefik Ilkin Serengil. Geraadpleegd op 7 april 2022, van https://sefiks.com/2018/08/06/deep-face-recognition-with-keras/
[9] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854.

Yebin Lee
Feb 2021, Industrial Information System Engineering at Seoultech (BS)
March 2021 ~, Data Science at Seoultech (ME)
Research Area : Data Mining, Machine Learning, Natural Language Processing

Boyoung Han
Feb 2019, IT Management at Seoultech (BS)
March 2021 ~, Data Science at Seoultech (ME)
Research Area : Homomorphic Encryption, Machine Learning

Juhyeon Jung
Aug 2021, Industrial Information System Engineering at Seoultech (BS)
Sep 2021 ~, Data Science at Seoultech (ME)
Research Area :Data Mining, Machine Learning, Recommendation System