# COSE474-2022F: Final Project
# "Fake Detection with CNN"

HyeonSu Hwang

## Abstract

본 연구는 python을 활용한 CNN에 관한 연구로 딥페이크로 만들어진 사진이나 포토샵으로 보정한 사진을 실제 인물의 사진과 분류하고 탐지하는데 연구초점이 있다. 기술의 발전으로 사람의 육안으로는 딥페이크 가짜 사진과 보정 사진을 일반 사진과 구분하지 못하는 시대에 도달했으며 딥러닝의 CNN기술을 활용하여 정확도가 좋은 모델을 만들어 보고자 하였다. 데이터는 실제 얼굴 사진 7만개와 딥페이크로 만들어낸 가짜 얼굴 7만개 사진을 이용하였으며 총 14만개 데이터로 train 10만개, valid 2만개, test 2만개로 나누어 모델을 학습시키고 성능 평가를 진행했다. CNN의 구조는 직접 은닉층을 쌓고 하이퍼 파라미터들을 변경한 구조 두 가지와 전이학습을 이용한 구조를 두 가지를 이용해 성능평가를 실시하였다.

주요용어 : CNN, 딥페이크 , 신경망, 전이학습, VGG16, DenseNet.

## 1. Introduction

최근 기술의 발전으로 사람의 육안으로는 포토샵 기술이나 딥페이크 가짜 얼굴을 구분하지 못하는 시대에 도달했다. 사회적으로 포토샵과 딥페이크를 악용한 범죄 사례가 증가하는 추세이다. 가짜 뉴스, 신종 금융 사기, 포르노 등 세계적으로 수많은 피해 사례들이 발생하고 있다.또한 유튜브나 sns 등에 포토샵이나 딥페이크를 활용한 사진이나 영상을 많이 볼 수 있다. 이는 쉽게 악용되어 범죄 행위로 연결될 여지가 많기 때문에 탐지 모델을 상용화 시킬 필요가 있다.

## 2. Problem definition & chanllenges

가짜 얼굴은 적대관계생성신경망(GAN: Generative Adversarial Network)이라는 기계학습(ML) 기술을 사용하여 새로운 인물을 만들어 낼 수 있지만 포토샵 사진은 저작권 문제로 대량의 데이터를 구하기 힘들다. 딥페이크 탐지 모델을 학습시키고 하나의 작업을 위해 훈련된 모델을 유사 작업 수행 모델의 시작점으로 활용하는 딥러닝 접근법인 전이학습을 이용해 포토샵 탐지에도 사용이 목표.

## 3. Related Works

딥페이크 생성에서의 최고의 결과는 StyleGAN-V2로 얻어졌다(Karras et al., 2020). 최근 Face Warping Features 방법과 Mesoscopic Features방법을 사용한 CNN분류기로 조작탐지 성능을 높였다(Tolosana et al., 2020).

## 4. Datasets

DFDC에서 제공한 FAKE얼굴(StyleGAN에서 생성)(Dolhansky et al., 2020)과 Nvidia가 수집한 REAL얼굴에서 샘플링한 데이터셋을 사용. 총 14만 개로 실제 얼굴 7만개와, 딥페이크로 만들어낸 가짜 얼굴 7만개 사진을 사용하였다.

### 4.1. 데이터 전처리

flow_from_directory메서드의 옵션을 조절해 데이터 전처리 작업을 수행하였다. target_size는 이미지 픽셀의 크기를 조절하는 옵션으로, 학습 속도를 위해 32*32의 입력크기로 진행하였다. ResNet, VGG등 최신 딥러닝 모델이 224*224 를 사용하기 때문에 이후 다른 모델을 사용할 때는 이 크기를 사용하였다. color_mode는 변환될 채널을 선택하는 옵션으로 데이터가 많기 때문에 우선 학습 속도를 위해 1개 채널인 를 사용했고, 이후 모델에서는 정확도를 개선시키 위해 3개 채널인 rgb를 사용하였다. 옵션은 반환될 라벨 배열의 종류를 결정해 주는데 이진 분류이기 때문에를 binary사용했다.

## 5. Method

### 5.1. 기본 CNN

필터 개수는 훈련 속도와 예측 성능의 균형을 맞출 수 있게 32개로 지정했고, 필터 크기는 최근 등장하는 CNN이 모두 3*3을 사용하기 때문에 본 연구에서 3*3을 이용하였다. 입력 이미지 크기는 이미지 픽셀과 마찬가지로, 신경망의 훈련 시간을 줄이기 위해 32*32을 사용했고 이후 정확도 개선을 위해 입력 크기를

증가시켜 진행했다. 채널도 마찬가지로 속도를 위해 우선 1개로 지정했고 이후 3개 채널로 늘려 학습시켰다. 최대 풀링 크기는 일반적으로 가로2, 세로2 크기로 지정하며 입력 레이어의 차원을 반으로 줄였다. 컨볼루션 레이어와 최대풀링 레이어의 활성화 함수는 RELU를 사용하고, 완전연결레이어의 활성화 함수는 이진분류이기 때문에 시그모이드를 사용했다. 손실함수는 이진분류에 적합한 binary_crossentropy, 옵티마이저는 adam, 성능평가 척도는 accuracy로 지정했다.


Figure 1. 컨볼루션 레이어와 최대 풀링 레이어를 함께 구성하여 층을 쌓고 우측에는 완전 연결 레이어를 연결하는 기본적인 CNN을 모델링 하였다. 컨볼루션 레이어는 이미지 내 특징을 식별하는 역할을 하며 최대 풀링 레이어는 각 컨볼루션 레이어 결과의 가중치 개수를 줄여서 모델의 복잡도를 줄이고 과적합을 방지한다. 완전 연결 레이어는 추출한 정보를 바탕으로 최종 분류를 담당한다.

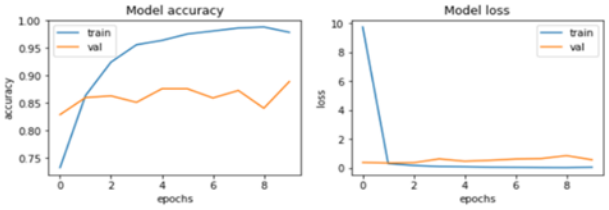


Figure 2. 모델링에서는 데이터가 매우 크기 때문에 큰 크기의 dataset을 효율적으로 학습시키는 fit_generator함수로 학습시켰다. epoch은 10으로 진행하였고, 그림처럼 validation의 정확도가 85%정도로 나오는 것을 확인할 수 있었다.

### 5.2. 6개 블록 CNN

다음은 컨볼루션 레이어와 최대 풀링 레이어 블록을 6개 연결한 모델을 이용해 진행하였다. 입력 이미지 크기는 이전에 32*32를 사용했던 것에 비해 높은 성능을 위해 224*224로 증가시키고 필터의 개수 또한 2배수로 늘려가며 32, 64, 128, 256, 512개로 층을 쌓았다. 채널개수도 이전 grayscale로 채널 1개를 사용했으나 채널 3개인 rgb로 바꿔 모델 성능을 최대한 끌어 올리고자 하였다.

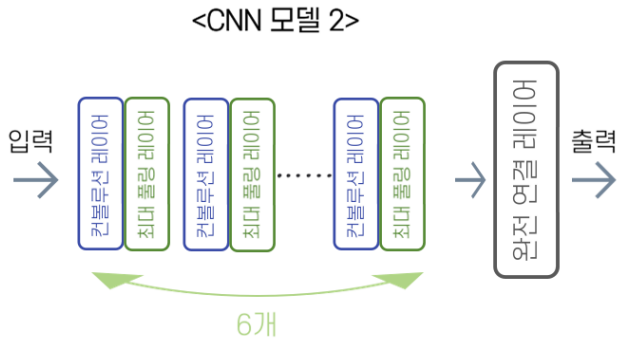


Figure 3. 컨볼루션 레이어와 최대 풀링 레이어 블록을 6개 연결한 모델

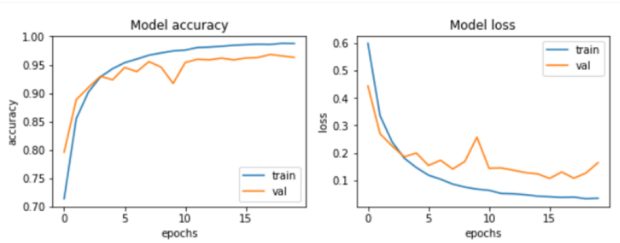


Figure 4. 20epoch으로 충분히 학습한 결과, train 0.9828, validation 0.9628의 정확도를 보이며, 이전 기본 CNN에 비해 성능이 많이 개선되었음을 알 수 있다.

### 5.3. 전이학습-VGG16

전이학습은 특정 대상을 예측하게 훈련시킨 모델을 다른 대상도 예측할 수 있게 바꾸는 기법이다.CNN에서 컨볼루션 레이어와 최대 풀링 레이어 부분을 고정하고 마지막 완전 연결 레이어만 원하는 레이어로 재학습시키면 기존의 잘 만들어진 CNN구조를 이용해 새로운 클래스를 예측하도록 바꿀 수 있다.


Figure 5. 본 연구에서는 케라스에 내장되어 있는 VGG16모델을 불러와 기존의 다중분류를 위한 완전연결 레이어를 분리하고 real과 fake를 구분하기 위한 이진 분류의 완전연결 레이어를 재학습 시키는 과정을 진행했다.

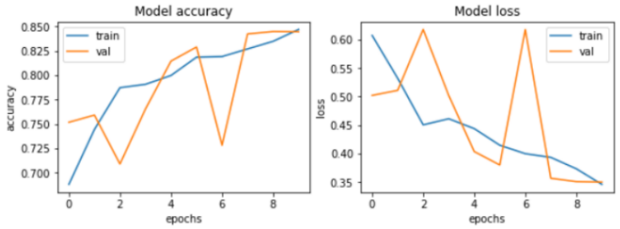


Figure 6. GPU 성능상 시간적 제약이 생겨 절반의 epoch과 훨씬 적은 iteration을 사용하게 되었지만, 그럼에도 84%의 높은 valid accuracy를 보였다.

### 5.4. 전이학습-DenseNet

DenseNet은 ResNet의 개념을 확장하여 나온 최신 딥러닝 구조이다. ResNet에서는 층이 깊을수록 성능이 떨어지는 것을 막기 위해, 입력 데이터를 합성곱 계층을 넘어 출력에 바로 더하는 스킵 연결을 사용한다.

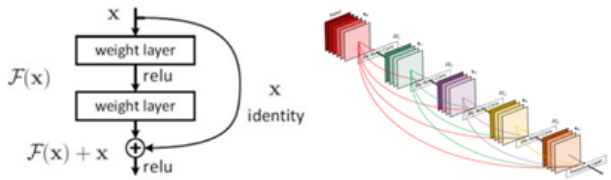


Figure 7. 스킵 연결을 모든 레이어로 확장되게끔 바꾼 것이 DenseNet이다. ResNet보다 기울기 소실 문제를 완화하고 파라미터 수와 연산량이 적다는 장점이 있어 본 연구에서 사용하였다. 모델링에는 이미지넷 가중치와 GlobalAverage풀링을 사용하여 성능을 높였다.

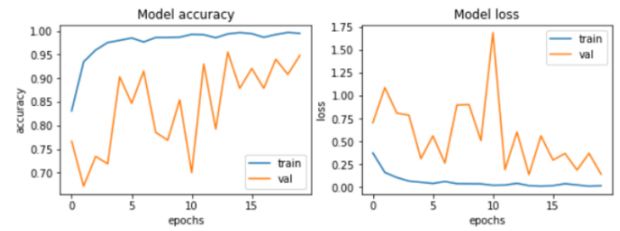


Figure 8. 1 epoch당 너무 많은 시간이 소요되어, 데이터 수를 1/10로 줄이고 epoch수를 20으로 늘려 사용하였다. 모든 데이터를 사용하지 않고도 95%의 valid acc를 보여 뛰어난 성능을 확인할 수 있었다.

## 6. 모델 평가

validation을 기준으로 단순 정확도 수치는 6개의 블록을 쌓은 CNN 모델이 가장 좋았으나 VGG모델과 DenseNet모델 또한 제약이 있음에도 충분히 좋은 성능을 보인다고 판단하고, 각각 test accuracy를 확인하고 가장 좋은 모델의 혼동행렬을 확인하였다.

Test Accuracy

| Method | Classifiers | Accuracy |
|---|---|---|
| CNN_default | CNN | AUC = 83% |
| CNN_6Blocks | CNN | AUC = 83% |
| VGG16 | CNN | AUC = 84% |
| DenseNet | CNN | AUC = 95% |

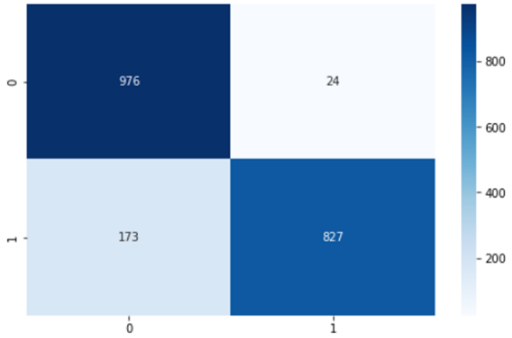


Figure 9. 앞서 valid accuracy와는 다르게 test에서는 6개 블록 CNN의 정확도가 매우 낮아 과적합이 의심되고, DenseNet의 경우 가장 높게 나와 전체 데이터로 학습시킨다면 성능이 더 올라갈 것으로 예상된다. 혼동행렬은 DenseNet의 결과이다.

## 7. 결론

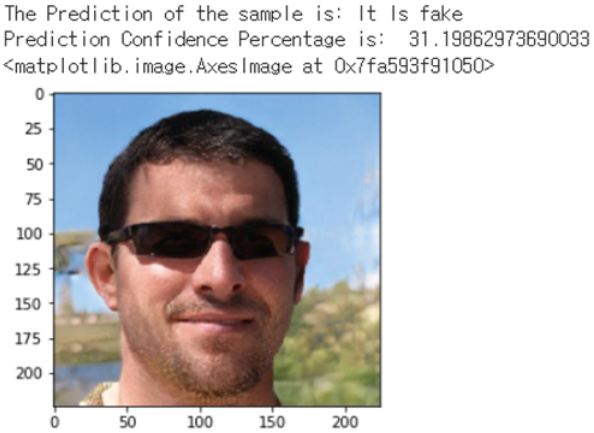


Figure 10. 최종 모델로 DenseNet을 사용하여 딥페이크 사진을 탐지한 결과 fake로 잘 분류 하는 것을 확인할 수 있었다.

다양한 방법으로 CNN모델의 성능을 충분히 높여 train시킨 fake사진뿐 아니라 포토샵으로 보정된 사진을 탐

지하려고 했지만 아쉽게도 포토샵으로 보정된 사진은 탐지 하지 못했다. 포토샵 사진을 탐지하기 위해서는 어느 정도 포토샵 이미지가 데이터에 포함돼야 한다고 판단된다. GAN을 이용해 fake사진을 만들 때 포토샵까지 적용해 fake face를 만들 수 있다면 포토샵의 특징을 CNN이 학습하여 좋은 정확도의 보정 탐지 모델을 만들 수 있을 거라 예상된다.

## References

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397, 2020.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8110–8119, 2020.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64:131–148, 2020.

(Karras et al., 2020) (Tolosana et al., 2020) (Dolhansky et al., 2020)

## A. Appendix

### A.1. GitHub

https://github.com/hyeonsu-hwang/final_project

### A.2. Computing Resource

OS : Windows 11 Pro
CPU GPU : Colab pro
TensorFlow : 2.9.2 version