2024-1 기계학습

<CNN 을 사용한
DeepFake 매체 판별 모델 제작>

과 목 명 : 기계학습

담당교수 : 손경아 교수님

팀 명 : 1 조

팀 원 : 201820699 김동령

201921185 이동현

202022349 박요셉

202127164 신현욱

202221129 장다희

아주대학교

www.ajou.ac.kr

1. 프로젝트 개요

1-1. 프로젝트 배경

디지털 시대의 발전과 함께 딥 페이크(Deepfake) 기술이 주목받고 있다. 딥 페이크는 GAN(Generative Adversarial Network)을 활용하여 원본과 구분하기 어려운 가짜 이미지를 생성하는 기술이다. 이러한 기술은 엔터테인먼트와 예술 분야에서 유용하게 사용될 수 있지만, 정치인의 허위 영상이나 음란물 생성 등 악의적인 목적으로 사용될 경우 심각한 사회적 문제를 일으킬 수 있다. 따라서, 딥 페이크 탐지 기술의 발전이 필수적이다.

1-2. 프로젝트 목표

본 프로젝트의 목표는 인물 이미지가 실제 이미지(REAL)인지 딥 페이크 이미지(FAKE)인지 분류하는 딥 페이크 탐지 모델을 개발하는 것이다. 이를 위해 다양한 환경에서 촬영된 실제 인물 이미지와 딥 페이크 이미지를 학습 및 테스트 과정에 이용하며, 최종적으로 팀원들의 이미지를 활용하여 검증한다.

2. 데이터

2-1. 데이터 수집

모델 제작에 사용할 데이터는 Kaggle 에서 수집하였다. 주요 데이터셋은 "deepfake_faces"와 "deepfake and real images"이다. 이 데이터들은 불균형 하므로, 데이터 언더샘플링을 통해 REAL 비율을 1:2 와 1:1 로 맞추어 학습하였다.

2-2. EDA 및 전처리

데이터셋 1 - <deepfake_faces>

전체 이미지 개수	95634
REAL(비율)	16293
FAKE(비율)	79341

[Under Sampling Dataset]

전체 이미지 개수	24000
REAL(비율)	8000
FAKE(비율)	16000

데이터셋 2 - <deepfakeand real images>

전체 이미지 개수	190281
REAL(비율)	60937
FAKE(비율)	129344

[Under Sampling Dataset]

전체 이미지 개수	24000
REAL(비율)	8000
FAKE(비율)	16000

- 데이터 언더샘플링 (Data Undersampling)

전체 데이터셋은 총 95.634 개의 이미지로 구성되어 있으며, 이중 REAL 데이터는 약 16,000 개, FAKE 데이터는 약 79,000 개이다. 이처럼 불균형한 데이터셋은 모델의 성능에 부정적인 영향을 미칠 수 있으며, 모델 학습에 과정에 다양한 문제를 발생시킬 가능성이 있기 때문에 최종적으로 REAL 데이터는 8,000 개, FAKE 데이터는 16,000 개로 언더샘플링을 진행하였다.

- 데이터 분할 (Data Splitting)

전처리된 데이터셋은 훈련, 검증, 테스트 데이터셋 6:2:2 의 비율로 데이터 분할을 진행한다.

- 이미지 리사이징 (Image Resizing)

CNN 모델은 고정된 크기의 입력을 요구하기 때문에 이를 위해 모든 이미지를 224*224 픽셀 크기로 리사이징을 진행한다.

3. 방법론 및 모델 구조

3-1. 방법론

본 프로젝트에서는 정의한 문제를 해결하기 위해 자체 설계한 CNN 모델을 사용하는 방법과 사전 학습된 모델(pre-trained model)을 미세 조정(fine-tuning)하는 방법 중 하나를 선택해야 했으며, 모델 선택 후에도 하이퍼파라미터와 내부 구조 등 다양한 요소에 대해 결정을 내려야 했다. 수많은 경우의 수 중에서 가장 적합한 모델 구성을 신속하게 찾아내기 위해 각 조원이 동일한 데이터셋을 이용하면서도 서로 다른 종류의 모델을 선택하여 학습을 진행하였다. 또한, 각 모델에 대해 다양한 레이어 구조와 하이퍼파라미터 구성을 시도해보아 최적의 성능을 내는 모델의 구성을 최종 모델로 채택하고자 하였다.

우선, '데이터셋 1'을 이용하여 다섯 종류의 서로 다른 모델을 학습시키고 테스트하였다. 그러나, 정확도가 기대에 미치지 못했다. 이는 '데이터셋 1'의 이미지가 일반적인 얼굴 사진이 아닌 영상에서 캡처된 얼굴 사진이기 때문에 데이터 품질이 낮아 정확도가 낮게 나왔다는 가설을 세웠고, 이를 해결하기 위해 영상이 아닌 얼굴 사진 데이터셋을 새로 도입하였다.

새로 도입한 '데이터셋 2'로 다섯 종류의 모델을 학습시키고 테스트하였다. 또한, 각 모델을 '데이터셋 1'과 '데이터셋 2'를 혼합한 데이터셋으로도 학습 및 테스트하였다. 그러나 데이터셋을 혼합한 경우 성능이 좋지 않았으며, '데이터셋 2'만으로 훈련된 다섯 종류의 모델이 '데이터셋 1'로 훈련된 다섯 종류의 모델보다 우수한 정확도를 나타내었다.

더 나아가, 정확도 뿐만 아니라 precision, recall, F1-score 등의 다른 지표들도 개선하고자 기존 학습 데이터셋 비율을 REAL=1:2 에서 REAL=1:1 로 맞추어 FAKE 데이터를

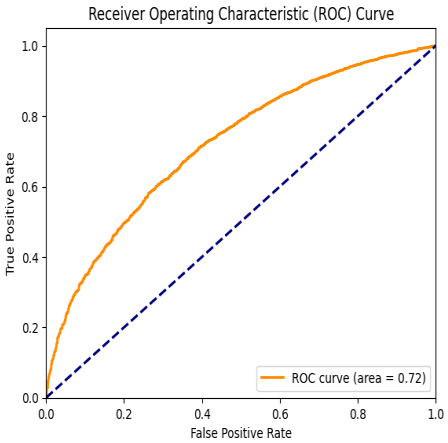
언더샘플링하였다. 그 결과, 정확도를 포함한 다른 지표에서도 안정적인 결과를 얻을 수 있었다.

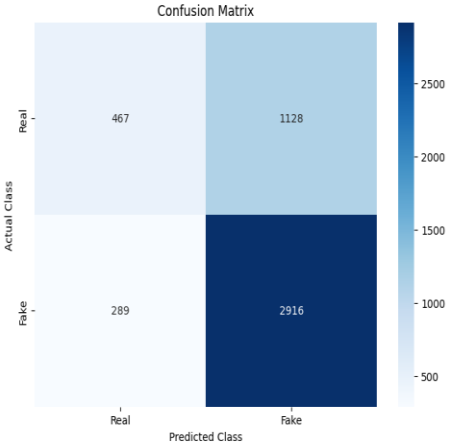
3-2. 모델 및 성능

- 자체 CNN 모델

Conv2D, MaxPooling2D, Flatten, Dense, Dropout, Output Dense 구조로 이루어짐.

성능 및 평가 요소

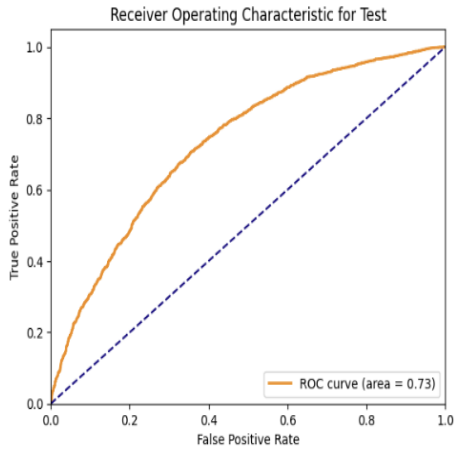


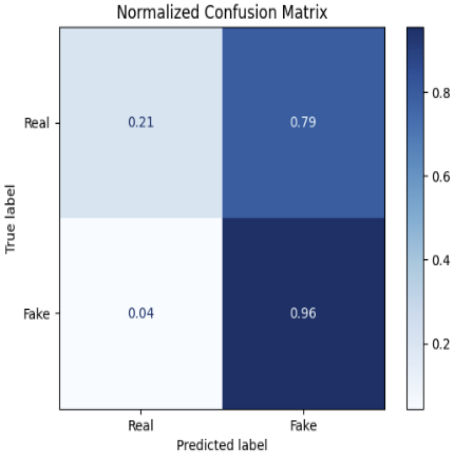


- ResNext-50

ResNext50 모델은 일반적인 ResNet 모델에 기반을 두고 있으며 ResNet 의 3x3 그룹 합성곱 계층을 병목 블록 내부의 3x3 합성곱 계층으로 대체하는 모델이다. ResNet 과 다른 점은 각 path 별로 같은 layer 구성을 갖고 있다는 점이다. 이러한 구조적 개선으로 더 깊은 네트워크 학습을 가능하게 하고, 병렬 처리와 동시성을 강화하여 이미지 분류에 높은 성능을 보인다.

성능 및 평가 요소

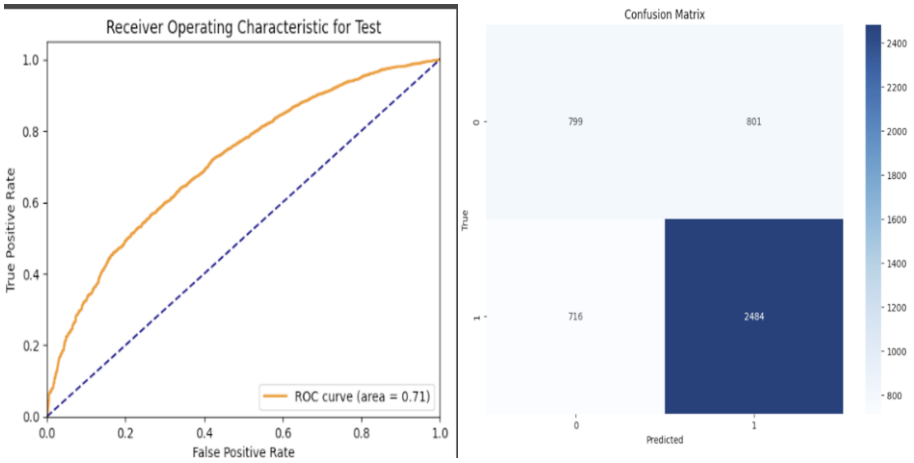




- ResNext-101

ResNext-101 모델은 ResNext-50 모델보다 더 깊은 네트워크 구조로 인해, 더 높은 성능을 발휘할 수 있다. 큰 데이터 셋과 복잡한 패턴을 학습하는 데 높은 성능을 갖고 있다.

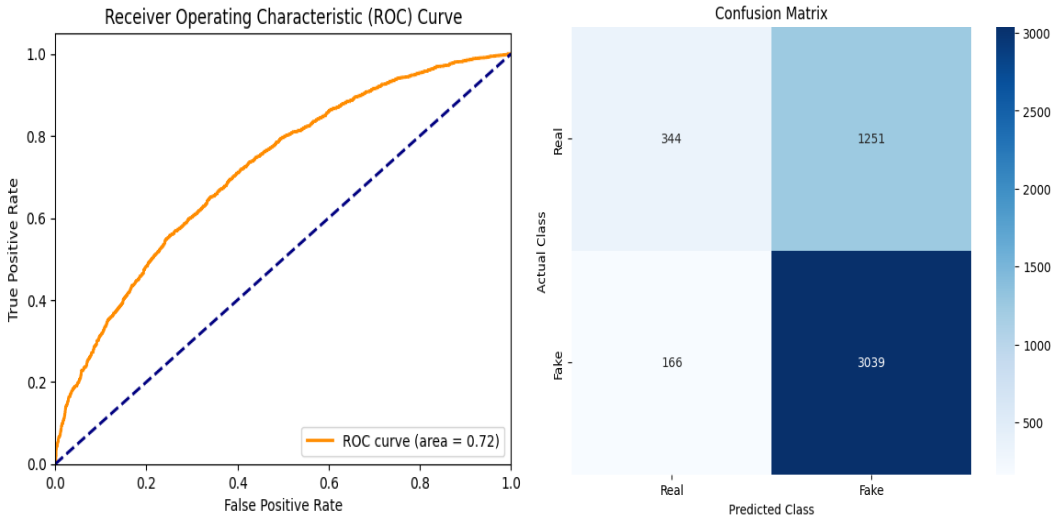
성능 및 평가 요소



- XceptionNet

XceptionNet 은 기존의 Inception 모델을 확장하여 효율성과 성능을 높인 구조로 깊이별 분리 합성곱(Depthwise Separable Convolution)을 이용한다. 즉 표준 합성곱을 두 단계로 분리한 것으로, 첫 번째 단계에서 채널별로 독립적인 깊이로 합성곱을 수행하고 두번째 단계에서는 포인트와이즈 합성곱(Pointwise Convolution)을 통해 채널 간의 정보를 통합하는 구조를 통해 효율적인 계산과 높은 표현력을 가질 수 있다.

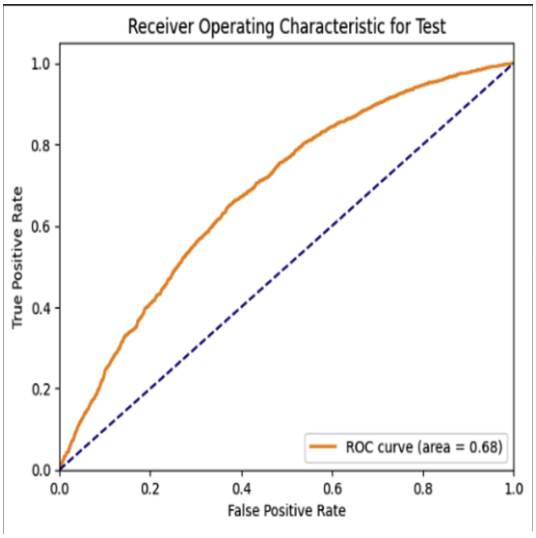
성능 및 평가 요소

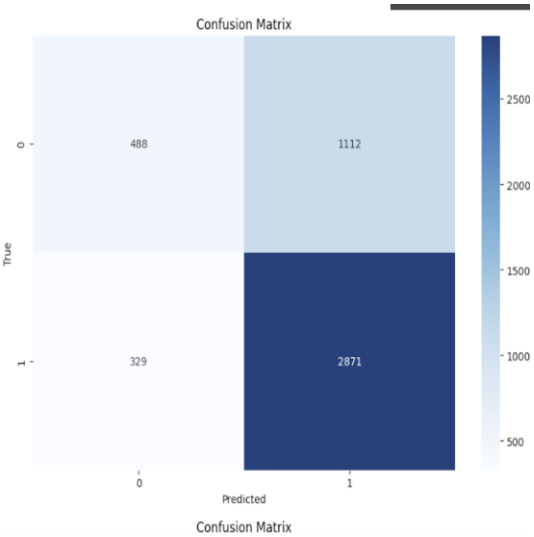


- ResNet-50

ResNet(Residual Networks)은 스킵 연결(skip connection)을 통해 정보 손실을 방지하고 기울기 소실 문제를 해결하는 잔차 블록(residual learning) 을 여러 블록 쌓아 모듈화 하여 사용하여 깊은 네트워크에서도 안정적인 학습이 가능하도록 설계된 모델이다.

성능 및 평가 요소





4. 결과 및 평가

4-1. 모델 별 지표 비교

1 번 데이터셋을 이용하여 앞서 설명한 모델들을 학습시킨 후 결과를 다시 한번 정리하면 다음과 같았다.

	정확도	Precision	Recall	F1-score
자체 CNN 모델	70.4%	0.91	0.72	0.80
ResNext-50	72.0%	0.84	0.69	0.76
ResNext_101	68.4%	0.76	0.78	0.77
XceptionNet	70.5%	0.71	0.95	0.81
ResNet-50	70.0%	0.72	0.90	0.80

[deepfake_faces 데이터셋[1] REAL:FAKE = 1:2]

전체적으로 비슷한 성능을 보였지만 정확도가 기대보다는 높지 않았다. 그 원인 중 하나로 1 번 데이터셋은 일반적인 이미지가 아닌 영상에서 캡쳐를 통해 수집한 이미지이기에 데이터 품질이 낮아 정확도가 낮게 나온 것이라는 가설을 세웠다. 따라서, 고화질의 일반 촬영 이미지인 2 번 데이터 셋만으로 다시 모델을 학습시켰고, 결과는 다음과 같았다.

	정확도	Precision	Recall	F1-score
자체 CNN 모델	78.4%	0.81	0.92	0.87
ResNext-50	84.9%	0.92	0.88	0.9
resnext_101	82.3%	0.60	0.90	0.72
XceptionNet	80.0%	0.89	0.84	0.86
ResNet-50	85.9%	0.89	0.92	0.91

[deepfake and real images 데이터셋[2] REAL:FAKE = 1:2]

전체적으로 1 번 데이터셋을 썼을 때보다 정확도와 F1 score 가 높게 나온 것을 확인할 수 있다.
추가적으로, 어떠한 데이터를 학습해도 딥 페이크 이미지를 분류할 수 있을지도 확인해보기 위해, 1 번 데이터셋을 학습시키고 2 번 데이터셋으로 평가하거나 그 반대로도 시도해보았다. 하지만 정확도가 평균적으로 40% 정도로 나와 좋지 못한 성능이 나왔다.

딥 페이크 사진을 분류하는 모델에서는 정확도도 중요하지만, 특히 'FAKE' 클래스인 이미지를 정확히 분류해내는 것이 중요하다. 따라서 정확도 뿐만 아니라 F1-score 도 함께 개선할 수 있는 조건을 고려해봐야 한다. 이에 따라 클래스 불균형을 해소하고자, F1-score 를 증가시키기 위해서는 데이터셋의 FAKE:REAL 의 비율을 1:1 로 맞춰야 한다는 가설을 세웠다. FAKE 이미지와 REAL 이미지의 사진 비율이 일정해야 Precision 과 Recall 값이 안정적으로 나올 것으로 기대했기 때문이다. 해당 가설에 준거하여 데이터 셋을 새로 구성하여 모델 학습을 진행한 결과는 다음과 같다.

	정확도	Precision	Recall	F1-score
자체 CNN 모델	55.5%	0.55	0.53	0.54
ResNext-50	72.4%	0.69	0.78	0.73
ResNext-101	63.4%	0.70	0.78	0.71
XceptionNet	63.1%	0.65	0.55	0.60
ResNet-50	71.3%	0.69	0.82	0.75

[deepfake_faces 데이터셋[1] REAL:FAKE = 1:1]

4-2. 실생활 활용

본 연구에서 직접 촬영한 사진을 실제(Real)로 간주하여 최종 모델에 적용한 결과, 대부분의 사진이 실제로 정확하게 판단되었다. 그러나 직접 촬영한 사진에 최신 딥 페이크 기술을 적용한 후 모델에 입력했을 때는 대다수의 사진을 여전히 실제로 판단하는 경향을 보였다. 이러한 결과가 나타난 이유는 학습 데이터로 사용된 데이터가 3~4 년 전의 딥 페이크 기술로 변형된 사진인 반면, 실제 팀원들의 사진에 적용된 딥 페이크 기술은 최신(올해) 기술이 적용되었기 때문이라고 분석된다.

실제로 데이터셋과 비슷한 시기에 제작된 연예인들의 딥 페이크 사진과 실제 사진을 비교한 결과, 모델은 약 60%의 정확도로 실제 사진과 딥 페이크 사진을 구별할 수 있었다. 이는 학습 데이터와 최신 딥 페이크 기술 간의 차이로 인해 분류 성능에 차이가 발생함을 시사한다.

4-3. 고찰

성능 지표들을 비교해보았을 때, 가장 복잡한 모델인 ResNext-101 이 ResNext-50 보다 성능이 떨어지는 것을 확인했다. ResNext-101 은 ResNext-50 보다 훨씬 많은 parameter 와 복잡성을 가지고 있어 더 많은 데이터를 필요로 한다. 우리의 데이터셋는 24000 장으로 비교적 작은 편이므로, 이러한 복잡한 모델은 과적합의 위험이 있다. 이로 인해 정확도가 떨어지고 새로운 데이터에 대한 일반화된 성능이 저하되었다고 판단했다. 그러므로 우리의 데이터셋에는 ResNext-101 보다 ResNext-50 의 모델이 더 적합하다는 결론을 낼 수 있었다.

딥 페이크 탐지 모델에서 주요 관심 대상은 'fake'이므로, 'fake' 이미지를 1 로 라벨링하여 positive 로 할당했다. 탐지 모델에서 가장 중요한 지표는 'fake'를 'fake'로 올바르게 판정한

True Positive (TP)이며, 가장 낮아야 할 지표는 'fake'를 'real'로 잘못 판정한 False Negative (FN)이다. 따라서 이상적인 혼동행렬은 TP > TN > FP > FN 이어야 한다.
초기 실험에서 딥 페이크 탐지 모델의 성능을 확인하기 위해 'real'과 'fake' 이미지의 비율을 1:2(8000:16000)로 설정했다. 이는 'fake'에 대한 탐지를 더 잘할 수 있을 것이라는 가정 하에 진행되었다.
그러나 1:2 의 비율로 실험한 결과, TP > FP > TN > FN 의 혼동행렬이 도출되었다. 이는 클래스 불균형으로 인한 결과로 예상되었다. 이에 데이터를 1:1(12000:12000)의 비율로 설정하고 다시 실험을 진행했다. 그 결과, TP > TN > FP > FN 의 혼동행렬을 얻을 수 있었다.
이를 통해 클래스 불균형이 탐지 모델의 성능에 미치는 영향을 확인할 수 있었으며, 적절한 클래스 비율 설정이 모델 성능 향상에 중요한 역할을 한다는 것을 알 수 있었다.
실제 모델은 딥 페이크 이미지는 딥 페이크로 잘 분류했으며, 이는 높은 Recall 값으로 확인 가능했다. 그러나 실제 사진을 딥 페이크로 분류하는 경우도 많이 발생했으며, 이는 상대적으로 낮은 Precision 값으로 확인 가능했다. 즉 단순하게 정확도가 높은 게 중요한 게 아니라 Recall 과 Precision 이 균형 잡힌 모델의 방향으로 개선하는 것이 중요하다.

5. 결론

본 연구를 통해 학습 데이터의 양, 비율 및 특성에 따라 모델의 성능이 크게 달라진다는 것을 확인할 수 있었다. 실용적인 딥 페이크 탐지 모델을 개발하기 위해서는 다양한 데이터를 풍부하게 학습시키는 것이 필수적이다. 그러나 딥 페이크 기술의 발전 속도가 매우 빠르다는 점에서 이러한 접근에는 한계가 존재한다. 따라서, 딥 페이크 기술의 진화에 맞추어 분류 모델 또한 신속하게 최신 정보를 반영하여 학습하는 것이 중요하다. 이는 딥 페이크 탐지의 효율성과 신뢰성을 높이는 데 핵심적인 요소가 될 것이다.

6. 참조문헌

[1] DAGNELIES (2020). "deepfake_faces". Kaggle. Available at: https://www.kaggle.com/datasets/dagnelies/deepfake-faces/data
[2] MANJIL KARKI (2022). "deepfake and real images". Kaggle. Available at: https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images/data