

Final project Report

01. Introduction

02. Data

03. Method

04. Results & Discussion

05. Additional Experiments

06. Conclusion

01. INTRODUCTION

주제 : GAN이 만든 생성형 이미지와 실제 인물 이미지 분류

최근 몇 년간, 인공지능 기술의 발전으로 인해 이미지 생성 기술이 놀라운 수준에 도달하였다. 특히 ‘딥페이크’나 ‘가상 모델’처럼, 실제 인물과 거의 구분되지 않는 이미지를 만들어내는 기술이 다양한 미디어와 산업에 빠르게 확산되고 있다.

이처럼 GAN을 기반으로 한 생성형 이미지가 범람하면서, 사람들은 이제 화면 속 인물이 진짜 사람인지, 혹은 인공지능이 만들어낸 가짜 이미지인지 쉽게 구분하기 어려운 상황에 이르렀다.



이에 따라 ‘진짜’와 ‘가짜’를 식별하는 기술의 중요성도 커지고 있다. 이러한 흐름에서, 본 프로젝트는 주어진 얼굴 이미지가 실제 인물인지 ai 생성 인물인지 판별하는 분류 모델을 설계하고 성능을 비교 분석하는 것이 목표이다.

본 프로젝트는 총 3가지 실험으로 구성되어 있으며, 각각의 실험은 생성형 이미지 탐지의 정밀도 향상과 실용성을 고려하여 다음과 같은 목적을 가진다.

1. 기본 분류 모델 개발: 주어진 얼굴 이미지가 실제 인물인지, 혹은 GAN이 생성한 인물인지를 이진 분류하는 모델을 구축해 Classification Metrics로 평가한다.
2. 입력 이미지 정보의 경량화 실험 - 추가실험: 전체 얼굴 이미지뿐만 아니라, 눈, 코, 입 등 일부 영역만을 입력으로 사용할 경우에도 유사한 분류 성능을 낼 수 있는지를 검증한다. 이를 통해 모델 입력의 경량화 가능성과 주요 특징 부위의 중요성을 분석한다.
3. 실제 이미지 내 보정 여부에 따른 분류 성능 분석 - 추가실험: 실제 인물 사진 중에서도 포토샵 등 가공된 이미지에 대해 모델이 여전히 real로 정확히 분류할 수 있는지를 실험한다. 이 과정에서, ‘진짜’의 정의를 시각적으로 확장하고 분류기의 일반화 능력을 평가한다.

02. DATA

Describe Data & EDA

data는 Kaggle의 [140k Real and Fake Faces]를 사용하였다. 해당 데이터는 Training data: fake 50000장, real 50000장으로 구성되어 있고, Valid data는 fake 10000장, real 10000장으로 구성되어 있고, Test data는 fake 10000장, real 10000장으로 구성되어 있다.



왼쪽은 fake, 오른쪽은 real이다. 라벨링 정보 파일 이름을 통해 fake(0), real(1)으로 이진 분류되어 있고, 입력 형식은 RGB 컬러 이미지 256 * 256이다.

Data Processing

• normalize

모델의 학습 안정성과 효율성을 높이기 위해 입력 데이터의 분포를 정규화하는 작업을 진행하였다.

이를 위해 본 프로젝트에서는 전체 학습 데이터를 대상으로 이미지의 평균과 표준편차를 계산하였다.

정규화 파라미터를 구하는 과정에서는 우선 transforms.ToTensor()만 적용하여 데이터를 로드하고, 후에 전체 데이터를 배치 단위로 순회하며 각 배치에서 계산된 평균과 표준편차를 누적하여 전체 데이터셋의 최종 평균 및 표준편차를 산출하였다. 이렇게 얻어진 값은 추후 Normalize()를 적용하여 학습 데이터를 정규화하는 데 활용했다.

• **image augmentation**

본 프로젝트에서는 데이터의 다양성을 확보하고 과적합을 방지하고자 데이터 증강을 하였다.

- RandomCrop(224), RandomHorizontalFlip(), ColorJitter(brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1), GaussianBlur(kernel_size=3)



test set에 대해서는 증강을 적용하지 않고, 고정된 전처리만 적용하였다. CenterCrop(224)를 적용하여 이미지 중심을 기준으로 224x224 크기로 잘라낸 후 ToTensor()를 적용하여 텐서 형태로 변환하였다.

• **image labeling**

본 프로젝트에서는 PyTorch의 ImageFolder 클래스를 사용하여 이미지 데이터셋을 구성하였다. 이때 클래스 라벨은 ImageFolder 내부 로직에 정수형 라벨이 자동 할당된다. 데이터 디렉토리 하위에 fake/, real/ 폴더가 존재하므로, 알파벳 순으로 'fake' < 'real'에 따라 {'fake': 0, 'real': 1}로 라벨 매핑된다.

실험에서 사용한 데이터는 데이터셋 내에서 train은 true 2000, false 2000장에 true와 false 각각 500장씩 증강해서 2500 + 2500인 5000장을 사용하였고, test은 true 500, false 500으로 1000장을 사용하였다.

03. METHOD

MODEL

본 프로젝트에서는 모델 학습 시 다음과 같은 고정된 기본 설정을 사용하였다:

- lr scheduler는 CosineAnnealingLR(patience=5)를 적용하였고, 손실 함수는 BCEWithLogitsLoss를 사용하였고, batch size는 64, Optimizer는 Adam로 고정하였다.

모델 fine-tuning 방식에 따라 다양한 학습 전략을 실험하였다. 구체적으로는 다음과 같은 방법들을 적용했다.

1. Fully Connected (FC) 층만 학습
2. FC + Batch Normalization(BN) 층 학습
3. 출력층의 노드 개수만 변경 (기존 학습된 가중치는 고정)
4. 상위 일부 블록만 학습
5. 전체 네트워크를 fine-tuning
 - 각 전략별로 lr은 [0.005, 0.001, 0.0005, 0.0003, 0.0001, 0.00005, 0.00001] 범위로, 이후 최적의 fine-tuning 방식과 해당 방식에서의 최적 학습률 조합을 바탕으로, 정규화 계수 [0.00001, 0.00005, 0.0001, 0.0005] 범위로 설정해 실험하였다.

ResNet50

본 프로젝트에서는 전이학습 기반 모델로 ResNet-50으로 실험을 진행하였다. 다양한 fine-tuning 전략을 적용하고, 학습률을 조정하며 성능을 비교한 결과, 전체 네트워크 학습 방식에서 학습률 0.0001에서 Accuracy 94%로 가장 우수한 성능을 보였다. 그 외의 방식에서는 FC만 학습 시 Accuracy 80%, FC+BN 학습 시 90%, 출력층 노드 개수만 변경한 경우 49%, 상위 블럭 4개만 학습한 경우 90% 결과가 도출되었다. 이후 성능 향상을 위해 기존의 단일 FC층 대신, FC층을 여러 개 쌓은 구조를 적용해봤지만, 성능이 저하되었다. 따라서 최종 모델에서는 CNN 결과를 바로 이진 분류 노드로 연결하는 단일 FC층 구조를 유지하였다. 이후 정규화 값을 조정하며 성능을 추가로 개선해, 정규화 값이 0.0001에서 가장 높은 Accuracy 95%를 기록하여, 이를 최종 모델의 하이퍼파라미터로 선정하였다.

EfficientNetB0

본 프로젝트에서는 EfficientNet-B0를 기반으로 실험을 진행하였다. 우선 사전 학습된 가중치를 고정한 채 FC 층만 학습하면서 학습률을 조정한 결과, 학습률 0.0005에서 Accuracy 77.5%의 성능을 확인했다. FC + BN 구조에서는 학습률 0.01에서 Accuracy 90.2%로 향상되었다. 상위 4개 block만 학습한 부분 fine-tuning에서는 학습률 0.007에서 Accuracy 90.8%를 달성하였다. 전 층을 모두 fine-tuning 하였을 때는 학습률 0.001에서 Accuracy 93.9%로 성능을 간신히하였다. 이후에 정규화 값을 조정하여 성능을 추가로 개선하였고, 0.00001에서 Accuracy 94.5%를 기록하여 최고 성능을 달성하였다. 이를 최종 모델의 하이퍼파라미터로 선정하였다.

MobileNetV2

학습률을 조정하며 성능을 비교한 결과, 전체 fine tuning 방식에서 학습률 0.0005에서 Accuracy 92.90%로 가장 우수한 성능을 보였다. 그리고 각 방식에서 최선의 Accuracy에 대해도 말해보자면, 출력층 노드 개수만 변경했을 때는 Accuracy가 50.8%로 거의 학습되지 않은 결과를 볼 수 있었고, fc+bn에서는 학습률 0.005에서 Accuracy가 76.80%가 나왔고, fc만 학습한 경우에는 학습률 0.005에서 Accuracy 77.79%가 나왔고, layer 3,4 학습에서는 학습률 0.0001에서 Accuracy 85.20%가 나왔고 layer4만 학습에서는 0.0005에서 Accuracy가 89.50%가 나왔다. 그래서 최종적으로 전체 fine tuning 방식에서 학습률 0.0005를 최종 하이퍼파라미터로 선정하였다.

Vision Transformer

본 프로젝트에서는 이미지 분류에서 Transformer 구조를 적용한 Vision Transformer(ViT) 모델을 사용하여 실험을 진행하였다. 사전학습된 ViT Base 모델을 기반으로 출력층에 단일 선형 계층(FC) 또는 Batch Normalization(BN)을 결합한 구조를 적용하여 비교 실험을 수행하였다. 학습률과 정규화 계수(weight decay)를 조정하며 성능을 비교한 결과, FC+BN 구조에 학습률 0.00001과 weight decay 0.00005를 적용한 조합에서 가장 우수한 성능을 보였다.

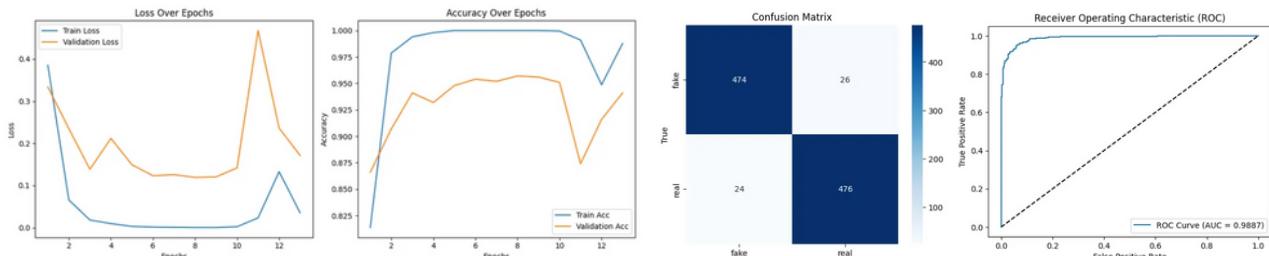
04. RESULTS & DISCUSSION

MODEL RESULTS

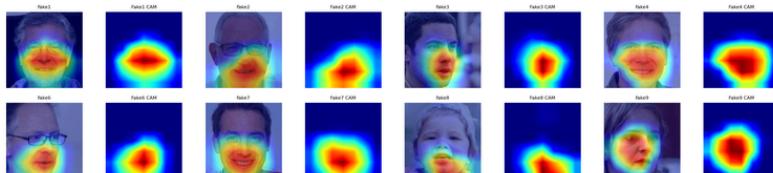
ResNet50

최종적으로 선정된 모델은 다음과 같은 구성과 파라미터를 기반으로 한다:

- Backbone 모델: ResNet-50
- Fine-tuning 방식: 전체 네트워크 학습 - linear : 출력층의 노드 개수를 맞추기 위한 단일 선형층 구조, Learning Rate: 0.0001, Weight Decay: 0.0001



최종 테스트 결과, 본 모델은 Accuracy 95%, Recall 95.2%, F1-score 95.0%, Specificity 94.8%를 기록하였다. 이러한 결과는 모델이 양쪽 클래스(real/fake) 모두에 대해 균형 잡힌 분류 성능을 보였음을 의미한다. 또한 ROC 곡선의 AUC 값은 0.9882로 매우 높게 나타났는데, 이는 다양한 분류 기준에서도 높은 구분 능력을 유지한다는 의미이다. 이러한 결과는 본 모델이 GAN이 생성한 얼굴 이미지와 실제 인물 이미지를 효과적으로 구분할 수 있는 높은 신뢰도의 이진 분류기임을 보여준다.

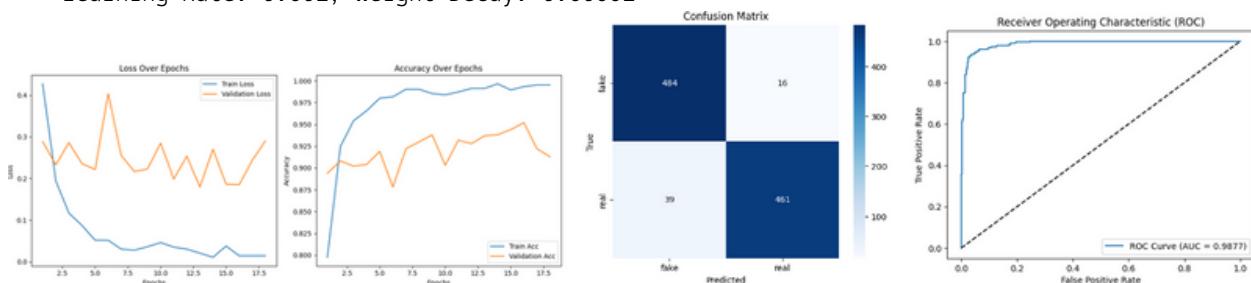


또한 Grad-CAM을 활용해 모델이 집중하는 영역을 시각화한 결과, 주로 하관 부위에 주목하고 있음을 확인할 수 있었다.

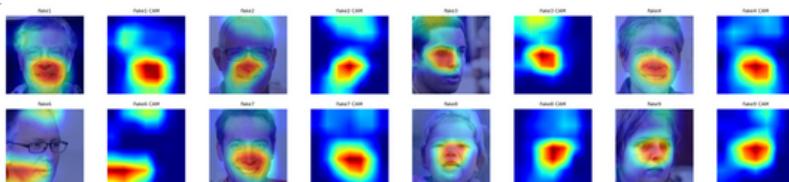
EfficientNetB0

최종적으로 선정된 모델은 다음과 같은 구성과 파라미터를 기반으로 한다:

- Backbone 모델: EfficientNet-B0
- Fine-tuning 방식: 전체 네트워크 학습 - linear : 출력층의 노드 개수를 맞추기 위한 단일 선형층 구조, Learning Rate: 0.001, Weight Decay: 0.00001



최종 테스트 결과, 본 모델은 Accuracy 94.5%, recall 92.2%, f1-score 94.3%, Specificity 96.8%를 기록하였다. 이러한 결과는 모델이 양쪽 클래스(real/fake) 모두에 대해 균형 잡힌 분류 성능을 보였음을 의미한다. 정확도 이외에도 다양한 분류 기준에서 높은 구분 능력을 유지함을 알 수 있다.

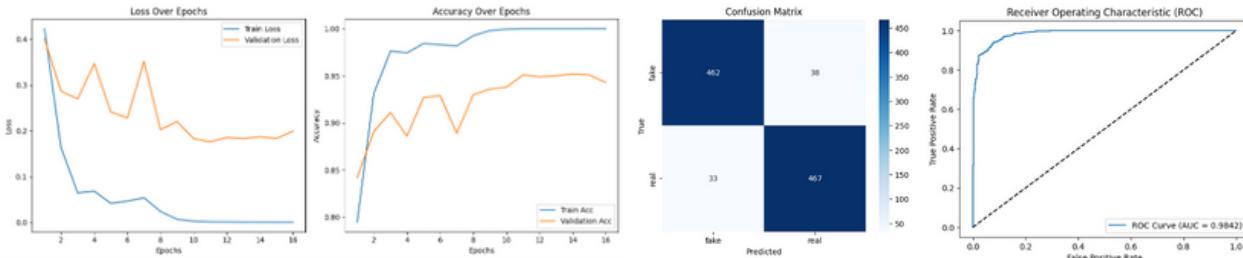


모델이 주로 하관에 강하게 주목하고 있음을 확인할 수 있다. 이는 GAN 합성 이미지에서 해당 영역의 텍스쳐와 색 번짐이 실제 얼굴보다 더 나타난다고 판단할 수 있다.

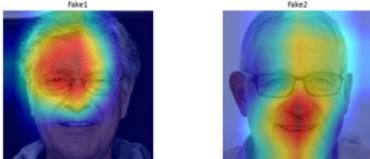
MobileNetV2

최종적으로 선정된 모델은 다음과 같은 구성과 파라미터를 기반으로 한다:

- Backbone 모델 : MobileNetV2
- Fine-tuning 방식 : 전체 네트워크 학습, learning rate : 0.0005



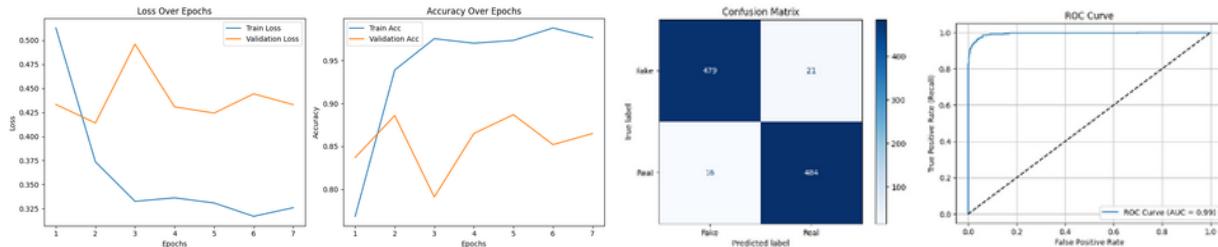
최종 테스트 결과, 본 모델은 Accuracy 92.90%, Recall 93.40%, F1-score 92.94%, Specificity 92.40%를 기록하였다. 이러한 결과는 모델이 두 클래스 모두에 대해 균형잡힌 분류를 잘 했다는 것을 알 수 있었다. ROC curve의 AUC값은 0.9842로 아주 높게 나타났다. 이는 한쪽으로 편향되게 분류하지 않는다는 것을 의미한다.



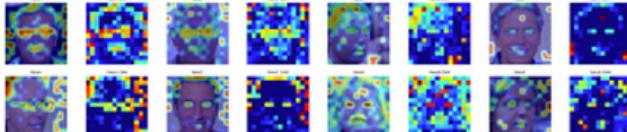
또한 Grad-CAM을 활용해서 집중되는 영역을 시각화한 결과, 처음에 예측했던 눈쪽, 코쪽, 하관 쪽에 집중한 것을 잘 알 수 있었다.

Vision Transformer

최종적으로 FC+BN 구조, 학습률 0.00001, 정규화 계수 0.00005를 최적의 하이퍼파라미터로 설정하였다.



최종 테스트 결과, 본 모델은 Accuracy 96.3%, Recall 96.8%, F1 Score 96.3%, Specificity 95.8%를 기록하였다. 이러한 결과는 Vision Transformer 기반 모델이 real/fake 양쪽 클래스 모두에 대해 매우 균형 잡힌 분류 성능을 보였음을 시사한다.



모델이 Fake 이미지 분류 시 입술의 외곽, 눈썹 등의 시각적 특징에 주목하는 것을 확인할 수 있었고, 이는 모델이 실제 의미 있는 패턴을 학습했음을 시사한다.

DISCUSSION

본 실험에서는 다양한 CNN 기반 모델(ResNet50, EfficientNet-B0, MobileNetV2)과 Transformer 기반 모델 ViT를 활용하여 real/fake 얼굴 이미지 이진 분류 성능을 비교하였다. 각 모델에 대해 다양한 fine-tuning 전략을 적용하고, 하이퍼파라미터 튜닝을 통해 최적의 성능을 도출하여 다음과 같은 관찰을 했다.

1. 결과 분석:

ViT가 Accuracy 96.3%, F1-score 96.3%로 가장 뛰어난 성능을 기록하였다. 이는 전역적인 self-attention 구조가 전체 얼굴의 시각적 일관성 및 비정상적 특징 간의 관계를 포착하는 데 효과적이었기 때문으로 해석된다. 반면 MobileNetV2는 경량 구조로 인해 표현력이 제한되어 Accuracy 92.9%로 가장 낮은 성능을 보였다.

2. 시사점:

CNN 계열 모델들도 대체로 높은 정확도를 기록했지만, Grad-CAM 결과에 따르면 ResNet50과 EfficientNet은 하관 부위에 집중하며 분류 기준을 설정한 반면, ViT는 눈썹, 입술 등 보다 세밀한 시각 특징에 주목했다. 이는 모델마다 다른 특징을 가지고 분류하고 있다는 것을 시사한다.

3. 모델 평가 및 타당성 검증:

AUC, F1-score, Specificity 등 다양한 지표에서 균형 잡힌 성능을 보였으며, 특히 ViT는 모든 지표에서 최고 수준을 기록하여 학습 데이터 분포 내에서는 가장 신뢰도 높은 분류기로 확인되었다.

05. ADDITIONAL EXPERIMENTS 1

주제 : 입력 이미지 정보의 경량화 실험

전체 얼굴 이미지뿐만 아니라, 눈, 코, 입 등 일부 영역만을 입력으로 사용할 경우에도 유사한 분류 성능을 낼 수 있는지를 검증한다. 이를 통해 모델 입력의 경량화 가능성과 주요 특징 부위의 중요성을 분석한다.

Data Processing

1. MEDIAPIPE 기반 랜드마크 추출

모든 얼굴 이미지는 MEDIAPIPE의 FACE MESH 모델을 이용하여 468개의 얼굴 랜드마크 좌표를 추출하였다.

- 눈 (EYES): 총 21개 랜드마크 (좌우 주요 눈 주변 점 포함)
- 코 (NOSE): 1번부터 19번까지의 중심 코 영역
- 입 (MOUTH): 78~87번, 308~317번 (상·하 입술 영역)
 - 복합 부위: EYES + NOSE, NOSE + MOUTH

2. 부위별 크롭 및 영역 조정

선택된 랜드마크 인덱스를 기반으로 해당 부위의 경계 좌표를 계산하고, 이를 기준으로 이미지를 크롭하였다. 이 과정에서 약간의 확장을 적용하여 약간의 위치 오차나 얼굴 움직임에도 안정적으로 특징이 포함되도록 조정하였다. 구체적으로 가로 영역은 대부분 원래 크기의 약 1.3배로 확장하였으며, 세로 영역은 부위별 특성에 맞게 각각 다른 비율로 조정하였다.

3. 크롭 이미지 후처리 및 표준화

크롭 후 이미지의 가로세로 비율이 일정하지 않기 때문에, 후처리 과정으로 정사각형 패딩을 적용하였다. 최종적으로 120X120 크기로 리사이즈하여 통일된 입력 크기를 구성하였다.



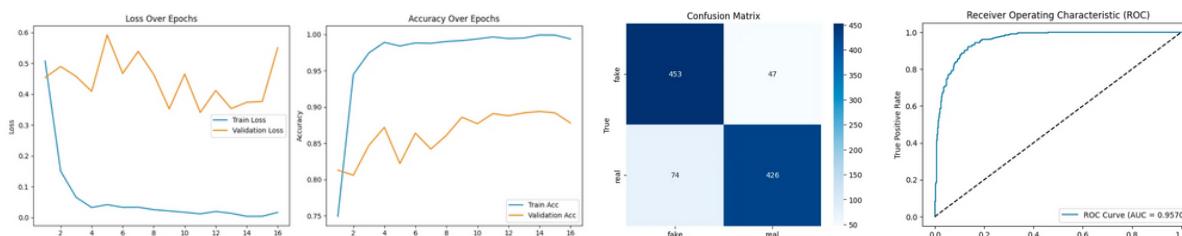
MODEL RESULTS

ResNet50

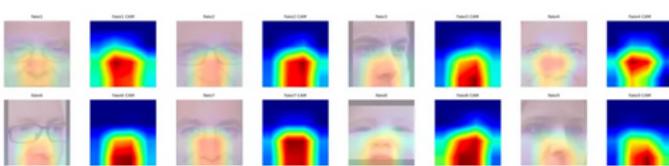
본 실험에서 가장 성능이 좋았던, 모델의 구성과 파라미터들을 사용해 학습했다.

Accuracy 기준으로 eyes: 73%, nose: 78%, mouth: 74%, eyes_nose: 83%, nose_mouth: 79%로 eyes_nose 성능이 좋았다. 하지만, 과적합 추이가 보여 정규화 값만 바꿔가며 다시 실험했다. 정규화 값은 [0.00001, 0.00005, 0.0001, 0.0005] 범위로 설정해 실험하였다.

실험 결과, weight_decay=0.00005로 Accuracy 87% 결과가 도출되었다.



eyes_nose에서 Accuracy 87.9%, Recall 85.2%, F1-score 87.5%, Specificity 90.6%를 기록하며, 전반적으로 균형 잡힌 분류 성능을 보인 것으로 평가된다.

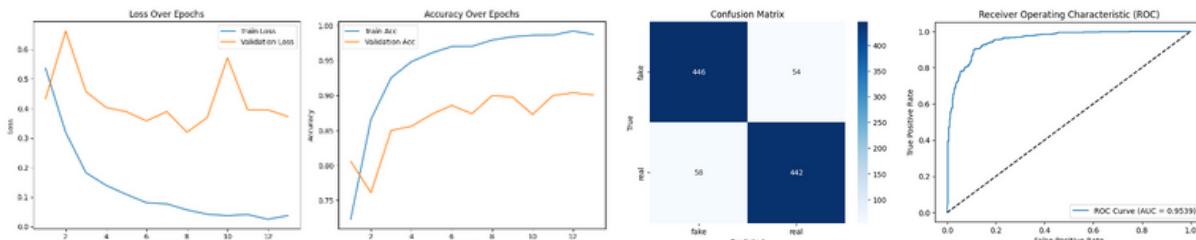


눈과 코가 포함된 이미지 실험에서는 모델이 주로 코 부위를 중심으로 판단하고 있음을 확인할 수 있었다.

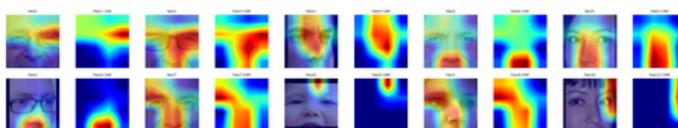
EfficientNetB0

본 실험에서 가장 성능이 좋았던 모델의 구성과 파라미터들을 사용해 학습하였다.

Accuracy 기준으로 eyes: 75.7%, nose: 85.1%, mouth: 84.5%, eyes_nose: 88.8%, nose_mouth: 85.2%로 eyes_nose의 성능이 좋았다.

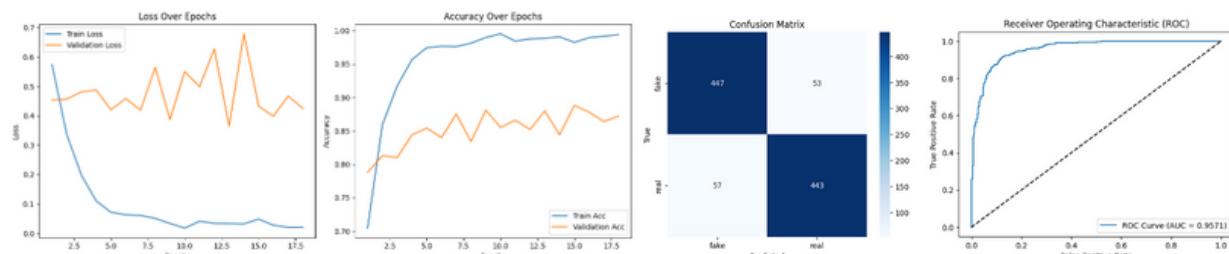


eyes_nose에서 Accuracy 88.8%, Recall 88.4%, F1-score 88.7%, Specificity 89.2%를 기록하며, 전반적으로 균형 잡힌 분류 성능을 보인 것으로 평가된다.



MobileNetV2

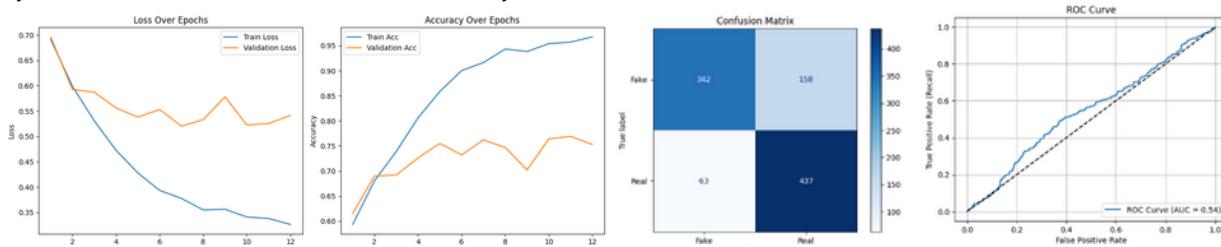
본 실험에서 가장 성능이 좋았던 하이퍼파라미터를 사용하여 학습하였다. Accuracy 기준으로, eyes: 76.60%, nose: 84.10%, mouth: 82.10%, eyes_nose: 87.90%, nose_mouth: 89.00%로 nose_mouth 성능이 가장 높게 나왔다.



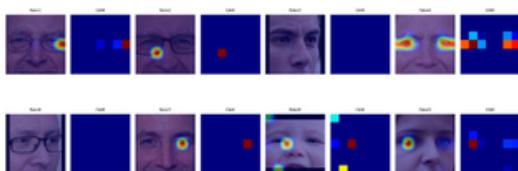
Accuracy 89.00%, Recall 88.4%, F1-score 88.96%, Specificity 89.40%를 기록하며, 전반적으로 균형잡힌 성능을 보인 것으로 평가된다.

Vision Transformer

본 실험에서는 가장 우수한 성능을 보였던 ViT 모델 구성과 동일한 구조 및 하이퍼파라미터를 유지한 상태로 실험을 수행하였다. Accuracy 기준으로 eyes 영역은 77%, nose는 76%, mouth는 73%, nose_mouth는 74%, eyes_nose는 76%로 측정되었으며, 이 중 eyes영역이 가장 높은 성능을 보였다.



Accuracy 77.9%, Recall 68.4%, Specificity 87.4%, F1 Score 75.5%를 기록하며 real은 잘 잡는데 fake는 일부 놓치는 편향이 있음을 알 수 있다.



DISCUSSION

해당 실험은 전체 얼굴 이미지가 아닌, 눈, 코, 입 등의 국소 영역만을 입력으로 활용하더라도 실질적인 분류 성능을 유지할 수 있는지를 평가하고자 하였다. 다양한 CNN 기반 모델 및 Transformer 기반 모델을 활용하여, 각 부분별 입력 조합에 대한 비교 실험을 수행하였다. 그 결과는 다음과 같은 의미 있는 시사점을 도출하였다.

1. 결과 분석:

전체 이미지 대신 eyes_nose, nose_mouth 등의 복합 부위를 입력으로 사용한 경우, CNN 기반 모델은 87~89%의 높은 정확도를 유지하였다. 특히 MobileNetV2는 nose_mouth 조합에서 Accuracy 89.0%로 모든 모델 중 최고 성능을 기록했는데, 이는 입술과 코 주변이 GAN 위조 특징이 가장 뚜렷하게 드러나는 국소 영역이며, MobileNetV2의 depthwise separable convolution을 활용해 로컬 특징을 빠르게 추출할 수 있는 구조이기 때문에, 제한된 입력 조건에서도 강한 성능을 발휘한 것으로 해석된다. 반면, ViT는 eyes 영역 입력에서 Accuracy 77%로 가장 낮은 성능을 보였으며, 이는 구조적으로 local inductive bias가 없고, 입력 패치 수가 줄어들면 self-attention이 효과적으로 작동하지 못하는 한계 때문으로 해석된다.

2. 시사점:

Grad-CAM 시각화 결과, 대부분의 모델이 입 주변, 턱 등 하관 영역에 주목하는 경향을 보였으나, 실험적으로는 오히려 눈과 코 조합을 입력으로 사용했을 때 가장 높은 분류 성능이 나타났다. 이는 GAN 이미지에서 눈과 코 주변의 텍스처나 비현실적인 그림자 분포가 모델 판단에 더 효과적인 단서로 작용할 수 있음을 시사한다. 또한 모델마다 주목하는 영역이 상이하다는 걸 관찰할 수 있다.

- ResNet50은 깊은 계층 구조를 활용해 얼굴 전체의 복합적인 국소 특징에 반응하며, 주로 하관과 코 주변을 고르게 참고했다. MobileNetV2는 가볍고 지역적인 convolution 구조에 최적화되어 있어 입술이나 코 경계와 같은 로컬 anomaly에 민감하게 반응했다. 반면, ViT는 입술, 눈썹, 눈꼬리 등 미세한 시각 패턴에 집중했으며, 이는 self-attention 구조가 전역적 관계보다는 세부적인 선·윤곽의 배열에 주목하고 있음을 보여준다.

3. 모델 평가 및 타당성 검증:

기존 전체 이미지(240×240)를 국소 부위 중심으로 잘라낸 120×120 크기의 입력만을 사용했음에도, CNN 기반 모델에서 87~89%의 높은 정확도를 달성하였다. 이는 연산량 감소와 함께 실시간 추론 최적화 측면에서 매우 유의미한 결과이며, 실제 응용 환경에서 효율적이고 경량화된 얼굴 진위 판별 시스템의 구현 가능성을 강하게 시사한다.

05. ADDITIONAL EXPERIMENTS 2

주제 : 실제 이미지 내 보정 여부에 따른 분류 성능 분석

실제 인물 사진 중에서도 포토샵 등 가공된 이미지에 대해 모델이 여전히 REAL로 정확히 분류할 수 있는지를 실험한다. 이 과정에서, ‘진짜’의 정의를 시각적으로 확장하고 분류기의 일반화 능력을 평가할 수 있다.

Data Processing

주어진 Kaggle 데이터에서 test의 real 이미지의 100장을 직접 포토샵해 데이터로 사용했다. 해당 실험은 테스트 목적이므로, 본 실험에서 사용한 normalize와 224 크기 resize 방식만 적용하여 전처리를 수행하였다.



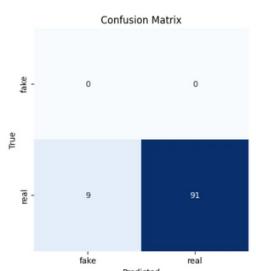
<보정 전>

<보정 후>

MODEL RESULTS

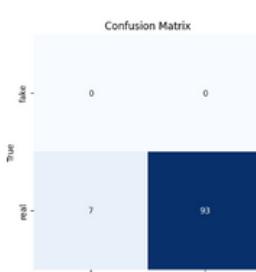
본 실험에서 가장 성능이 좋았던, 모델의 가중치를 저장한 모델 파일을 불러와 테스트를 진행했다.

RESENT50



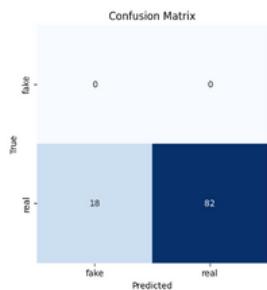
Accuracy 91.0, Recall 91.0%, F1-score 95.29%를 기록하며, 보정된 사진도 real로 분류했음을 볼 수 있다.

EfficientNetB0

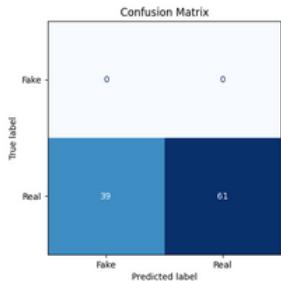


Accuracy 93%, Recall 93%, F1-score 96.3%를 기록하며, 보정된 사진도 real로 분류했음을 볼 수 있다.

MobilenetV2



Vision Transformer



DISCUSSION

해당 실험은 기존의 학습에 사용된 이미지가 GAN 기반 생성 이미지였던 것과 달리, 실제 인물 사진을 대상으로 하여 해당 모델이 보정된 이미지 역시 정확히 분류할 수 있는지를 평가하고자 하였다. 이는 모델의 실사용 가능성과 함께, ‘진짜(real)’에 대한 정의를 시각적으로 확장하는 데 목적이 있다.

1. 결과 분석:

EfficientNetB0는 Accuracy 93%로 보정된 실제 real 이미지를 가장 안정적으로 분류하였다. EfficientNet의 구조는 resolution, depth, width를 균형 있게 확장하는 compound scaling 전략을 사용하기 때문에, 색상 변화나 윤곽선 보정 등 다양한 시각적 변화에 robust하게 대응했다고 생각한다. 반면 ViT는 Accuracy 61%로 성능이 크게 하락하였다. ViT는 attention weight가 전체 이미지 분포의 일관성에 기반하므로, 보정으로 인해 전역 시각 특성이 바뀌면 성능이 급감하는 경향이 확인되었다.

2. 시사점:

Grad-CAM 결과에 따르면 EfficientNetB0는 얼굴 전체에서 일관된 하관 중심의 특징을 참고하여 판단한 반면, ViT는 눈썹, 눈가, 입술 주변 등 미세한 윤곽 및 텍스처 변화에 주목하였다. 하지만 보정 과정에서 색조, 피부 톤, 선명도 등의 시각적 특성이 달라지자, ViT는 기존 학습과의 전역 패턴 불일치로 인해 판단 기준을 상실한 반면, EfficientNet은 다양한 해상도와 시각 변화에 강건한 구조적 특성 덕분에 높은 성능을 유지할 수 있었다. 이러한 결과는, 동일한 입력이라도 모델 구조에 따라 해석 방식이 달라지고, 그에 따라 실제 환경에서의 판별 정확도에도 큰 차이가 생길 수 있음을 의미한다.

3. 모델 평가 및 타당성 검증:

해당 실험은 실제 사진에서의 ‘real’ 정의가 더 다양하고 가변적일 수 있음을 보여주며, 모델이 어느 정도 이를 시각적 특징 기반으로 포괄적으로 인식하고 있음을 확인한 데 의의가 있다. CNN 기반 모델들은 보정 이미지에서도 90% 이상 정확도를 유지하며 학습 데이터 외의 입력에 대해 높은 일반화 성능을 입증하였다. 입력 이미지의 세부적인 조명, 보정 유무, 피부 질감의 변화 등에도 불구하고 안정적인 판단을 수행했다는 점에서, 실 환경에서의 활용 가능성을 뒷받침한다.

06. CONCLUSION

본실험 : GAN이 만든 생성형 이미지와 실제 인물 이미지 분류

1. 주요 실험 요약

다양한 CNN 및 Transformer 기반 모델을 활용하여, GAN으로 생성된 얼굴 이미지와 실제 인물 이미지를 이진 분류하는 모델을 학습하였다. ViT는 본실험에서 Accuracy 96.3%, Recall 96.8%, F1-score 96.3%로 가장 높은 성능을 기록했다. 이는 ViT의 전역적 self-attention 구조가 전체 얼굴의 일관된 패턴을 포착하는 데 효과적이었기 때문으로 해석된다. GAN 이미지의 경우, 눈썹·입술·피부 윤곽 등에서 국소적으로는 자연스러워 보일 수 있지만 전체적으로는 미묘한 일관성 결여를 가지는데, ViT는 이러한 전역적 이상 패턴을 patch 간 관계를 학습함으로써 정밀하게 탐지할 수 있었다. 반면, MobileNetV2는 가장 낮은 정확도(92.9%)를 기록하였다. 이는 해당 모델이 경량화를 위해 depthwise separable convolution을 채택해 채널 간 정보 융합이 약하고 표현력이 제한되었기 때문으로 보인다. GAN 이미지처럼 고해상도 시각 정보의 정밀한 해석이 요구되는 문제에서는, MobileNet 구조의 한계가 상대적으로 더 크게 작용했을 가능성이 높다.

2. 한계

본 실험은 단일한 유형의 GAN으로 생성된 이미지만을 대상으로 모델을 학습하고 평가하였기 때문에, 다양한 GAN 구조에서 생성된 이미지에 대한 일반화 성능을 확인하지 못한 한계가 존재한다. 최근에는 StyleGAN 기반 생성 모델 등 보다 정교한 이미지를 생성할 수 있는 다양한 GAN 계열 모델들이 지속적으로 등장하고 있으며, 이러한 다양한 생성 방식에 대응할 수 있는 모델의 견고성을 평가하기 위해서는 보다 폭넓은 생성 이미지 데이터셋을 포함한 추가적인 실험이 필요하다.

추가 실험 1 : 입력 이미지 정보의 경량화 실험

1. 주요 실험 요약

얼굴 전체가 아닌 눈, 코, 입 등 일부 부위만 입력으로 활용해도 어느 정도의 분류 성능을 확보할 수 있는지를 실험하였다. Grad-CAM과 랜드마크 기반 크롭을 통해 다양한 부위 조합을 테스트한 결과, CNN 모델들이 ViT보다 전반적으로 더 나은 성능을 보였다. 특히 MobileNetV2는 nose+mouth 조합에서 모든 모델 중 가장 높은 성능을 기록하였는데, 이는 GAN 이미지에서 입술과 코 주변에 시각적으로 부자연스러운 위조 특징이 뚜렷하게 나타나는 경향이 있고, MobileNet이 이러한 국소 영역에서의 특징을 빠르게 추출하는 데 적합한 경량 구조이기 때문으로 해석된다. 특히 depthwise separable convolution 구조는 복잡한 전역 정보를 필요로 하지 않고도 효율적으로 로컬 특징을 포착할 수 있어, 제한된 입력 정보 환경에서도 높은 성능을 발휘한 것으로 보인다. 반면, ViT는 eyes 영역 입력에서 Accuracy 77%로 가장 낮은 성능을 기록했다. 이는 입력이 제한될 경우 self-attention의 효과적인 작동이 어려운 구조적 한계 때문으로 분석된다. 전체 문맥을 전제로 설계된 ViT 구조는 부분 입력 상황에서 그 강점을 살리지 못하고, 국소 정보만으로 위조 여부를 판단하는데 어려움을 겪는다고 판단한다.

2. 한계

특히 눈, 코, 입과 같은 부위를 크롭한 후 모델에 입력하기 위해 모든 이미지를 120×120 크기로 일괄적으로 리사이징했으나, 실제로 크롭된 이미지의 정보량은 그보다 훨씬 작아지는 경우도 많았다. 이러한 방식은 실제 입력 부위의 시각 정보를 왜곡하거나 패딩 비율을 증가시켜 모델 학습에 영향을 줄 수 있음에도, 해당 부문에 대한 정교한 보정이나 가변 입력 처리 전략을 적용하지 못한 한계가 있다.

추가 실험 2 : 실제 이미지 내 보정 여부에 따른 분류 성능 분석

1. 주요 실험 요약

실제 인물 사진을 기반으로, 색감 보정, 얼굴 보정 전후의 이미지에 대해 기존 학습된 모델이 동일하게 'real'로 분류하는지를 실험하였다. 보정된 real 이미지 분류 실험에서는 EfficientNetB0가 가장 높은 성능을 기록하였다. EfficientNet은 compound scaling 전략을 통해, 다양한 해상도 및 시각적 조건 변화에도 강건한 표현 능력을 갖는다. 이러한 구조적 특징 덕분에, 피부 톤 변화나 얼굴 윤곽 보정 등 보정 이미지에서 발생하는 색상·조명·형태의 미세한 변화에도 안정적으로 반응하며 높은 일반화 성능을 유지할 수 있었다. 반면, ViT는 Accuracy 61%로 가장 낮은 성능을 보였다. ViT는 입력 이미지를 고정된 패치로 나눈 후 전역적으로 self-attention을 수행하는 구조로, 전체 이미지의 전반적인 시각 패턴과 위치 관계에 의존한다. 하지만 보정된 real 이미지에서는 색감·윤곽·디테일이 수정되며, 결과적으로 patch 간 관계나 전체 시각 분포가 학습 시와 달라지는 문제가 발생해 시각적 노이즈나 보정에 민감하게 반응하기 때문에 성능이 급격히 저하된 것으로 해석된다.

2. 한계

테스트에 사용된 보정 이미지의 수는 제한적이었으며, 실제 이미지를 직접 편집하여 보정 데이터를 구성하는 과정에서 여러가지 제약으로 인해 다양한 유형의 보정 사례를 충분히 확보하지 못했다. 또한 사용된 보정 방식 역시 색감 보정, 윤곽 조정 등 편집 유형이 일관되지 않았기 때문에, 모델이 각 보정 유형에 대해 어떻게 반응하는지를 정량적으로 비교하거나 분석하기에는 데이터 규모와 구성 면에서 부족한 점이 있었다.

기본 분류 성능 측면에서 CNN 모델들이 주요 지표에서 95% 내외의 우수한 이진 분류 성능을 안정적으로 달성하고, 경량화 측면에서 전체 얼굴이 아닌 눈, 코, 입 등 일부 부위만으로도 높은 분류 성능을 달성하여, 실질적으로 경량화 가능성과 실시간 적용 가능성을 입증했다. 마지막으로, 일반화 측면에서 실제 보정 이미지와 같은 도메인 변화에도 높은 정확도를 유지하여, 실환경에서의 활용 가능성과 견고한 일반화 능력을 확인했다.