ResNet50 DeepFake Detector: Unmasking Reality

“ResNet50 DeepFake Detector: Unmasking Reality,” Indian Journal Of Science And Technology, vol. 17, no. 13. Indian Society for Education and Environment, pp. 1263–1271, 25-Mar-2024.

작성자

소프트웨어학과 32202841 유석

Overview

Introduction - What is deepfake detection?

출처 - youtube Midwell

Key Research Content

Purpose

공공 안전을 강화하기 위해 비디오 스트림에서 이상 현상을 자동으로 식별할 수 있는 도구에 대한 수요증가

Video Anomaly Detection

Image Transformation

Deep Learning

Research Objectives

- 딥페이크 영상의 탐지 정확도를 높이고 최적의 결과를 제공

- 미디어 처리의 시간 복잡성을 줄이면서 모델의 정확도를 향상

*Nvidia Tesla T4 GPU에서 32번의 epoch 진행

Previous deepfake detection research

(2019~2020) : 딥러닝, 고전적 머신러닝, 통계적 기법, 블록체인 기반 기법

- 매우 광범위한 딥페이크 탐지 연구의 진행

ResNet50과 LSTM을 결합한 하이브리드 아키텍처 / 파이썬으로 구현

- 인공지능과 머신러닝을 활용한 딥페이크 생성 및 탐지

Digital forensics capability analyzer

- 디지털 포렌식 도구의 발전, 현재 상태, 미래 전망을 심도 있게 탐구
- 맞춤형 디지털 포렌식 셀을 구축하고자 할 때 비용을 최소화하고 필수 정보를 제공

Limitations of previous research

모델 정확성 부족

- 기존 연구는 이미지 인식에 뛰어난 ResNet-50과 순차적 데이터 처리에 특화된 LSTM을 결합한 비디오 이상 감지에 관한 포괄적인 연구가 부족함

▶ ResNet50과 LSTM의 결합 모델을 사용하여 측정해 보다 높은 정확도를 달성

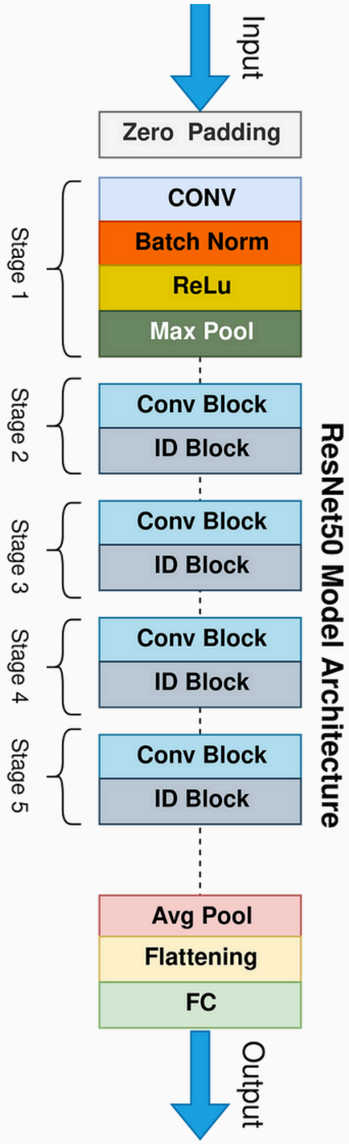
방법론 명확성 부족

- 기존 Digital Forensics Capability Analyzer를 개발하는 데 사용된 도구의 설계, 구현, 검증 과정에 대한 자세한 설명이 없어서 독자들이 도구가 어떻게 개발되고 테스트되었는지 이해하기 어려움

▶ 다양한 데이터셋에서 훈련을 거쳐 높은 정확도를 유지하고, 전체적인 방법론을 제시

How it works?

ResNet50
+
LSTM



ResNet50

아키텍처는 ResNet50을 기반으로 하며, 50개의 레이어와 32x4 차원을 갖고 있습니다.

Input

비디오 프레임을 입력으로 받아, 정상화 및 데이터 증강과 같은 전처리 과정을 진행합니다

50 Layers

전처리된 프레임에서 계층적 특징을 추출합니다.

Residual block

딥페이크 콘텐츠가 포함된 레이블이 있는 대규모 데이터셋으로 학습합니다.

Output

consensus decision-making이 적용되어 비디오에 딥페이크가 포함되어 있는지 최종적으로 판단합니다.

ResNet50

ResNet

- Residual Network의 약자로, 딥러닝 모델 중 하나
- 이미지 인식, 객체 탐지와 같은 컴퓨터 비전 작업에 주로 사용

ResNet에는 여러 가지 버전이 있으며, 각 버전은 레이어 수와 네트워크 깊이에 따라 구분

- ResNet-34, ResNet-50, ResNet-101, ResNet-152
- 상위 버전으로 갈수록 더 높은 수준의 패턴 인식이 가능하지만 계산 비용이 큼
- 다양한 복잡도의 문제에 맞게 사용

Features of ResNet50

50 Layers

총 50개의 층으로 이루어져 있으며, 이를 통해 매우 복잡한 패턴을 학습할 수 있음

Residual Conection

딥 뉴럴 네트워크 학습에서의 깊은 신경망에서 흔히 발생하는 기울기 소실문제를 해결

Image Scraping

이미지에서 특징을 추출해 다양한 이미지 분류 및 객체 인식 작업에서 뛰어난 성능 보유

LSTM(Long Short-Term Memory)

RNN(Recurrent Neural Network)

- 이전의 정보를 기억하고 다음 계산에 활용할 수 있는 구조
- 시퀀스 데이터나 시간에 따른 데이터 학습에 특화된 인공 신경망
- 시간이 길어질수록 기울기 소실(vanishing gradient) 문제로 인해 과거 정보를 잃어버릴 수 있음

LSTM

- 순환 신경망(RNN)의 한 종류
- 시퀀스 데이터나 시간에 따른 데이터(텍스트, 음성, 주식 가격 등)를 처리하고 학습하는데 효과적
- 자연어 처리, 음성 인식, 시계열 예측 등에 많이 사용

Features of LSTM

내부 메모리 셀

메모리 셀을 통해 중요한 정보를 장기간 저장하고, 필요한 정보를 새로운 입력 데이터와 조합해 업데이트

게이트 메커니즘

입력 데이터가 얼마나 메모리 셀에 저장될지,
어느 정도를 다음 단계로 전달할지,
출력으로 어느정도를 전송할지

기울기 소실 해결

각 게이트는 시퀀스의 중요한 정보를 유지하며, 필요없는 정보는 망각하는 메커니즘을 통해 기울기 소실과 폭주 문제를 줄임

Performance Metrics(1)

Accuracy

- 전체 데이터 중에서 정확히 분류된 비율

*TP = True Positive, TN = True Negative, F = False

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

- 긍정으로 분류된 사례 중에서 실제로 긍정인 사례의 비율을 나타냅니다.

$$Precision = \frac{TP}{TP + FP}$$

Specificity

- 실제 부정 사례 중에서 정확히 부정으로 분류된 비율
- Recall의 반대 개념

$$Specificity = \frac{TN}{TN + FP}$$

Performance Metrics(2)

Recall

- 실제 긍정 사례 중에서 정확히 긍정으로 분류된 비율

$$Recall = \frac{TP}{TP + FN}$$

F1 - score

- Recall과 Precision간의 균형을 잡는 지표로, FP와 FN의 비용이 다를때 유용함

$$F1 - score = 2 * \frac{FN}{TP + FN}$$

AUC

- ROC 곡선에서 양성 및 음성 사례를 얼마나 잘 구분하는지 나타냄, 곡선 아래 영역
- 값이 높을수록 모델의 효율성과 구분 능력이 뛰어남

$$AUC = \sum \frac{TPR[i] + tpr[i+1]}{2} * FPR[i+1] - FPR[i]$$

*TPR = True Positive Rate, FPR = False Positive Rate

Performance Results

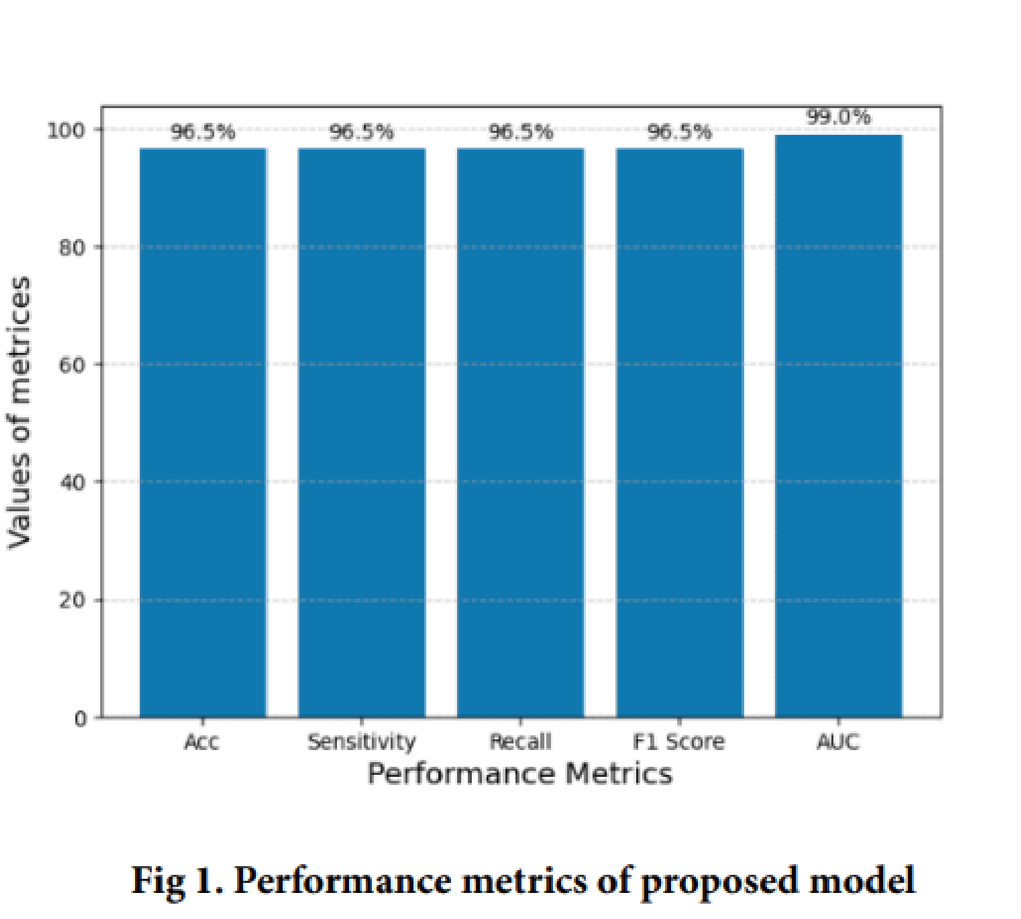


Fig 1. Performance metrics of proposed model

LSTM+ResNet50 비디오 이상 탐지 시스템의 효율성 지표

Used performance matrics

- Accuracy, Sensitivity, Recall, F1 Score, AUC
- MAP, TPR, FPR

Used Data

- CelebDf 데이터셋을 사용하여 훈련된 다양한 데이터 유형에 대해 높은 정확성과 응답성을 보여줌

Training of ResNet50 model

1. 실제 및 딥페이크 콘텐츠를 포함한 대규모 라벨링된 이미지 또는 비디오 프레임 인풋
2. 모델의 가중치 초기화 및 배치 단위로 네트워크에 훈련 데이터 제공
3. 네트워크는 각 훈련 예제에 대한 예측값 생성
4. 예측값과 실제 레이블 간의 차이를 교차 엔트로피와 같은 손실 함수로 계산
5. Backpropagation을 사용하여 손실 함수의 기울기 측정, 손실이 최소화가 되도록 가중치 업데이트
6. 최적화 알고리즘은 확률적 경사하강법, Adam과 같은 변형을 통해 가중치를 약간씩 조정
7. 여러번의 epochs 동안 위의 과정을 반복

ResNet50 model WorkFlows

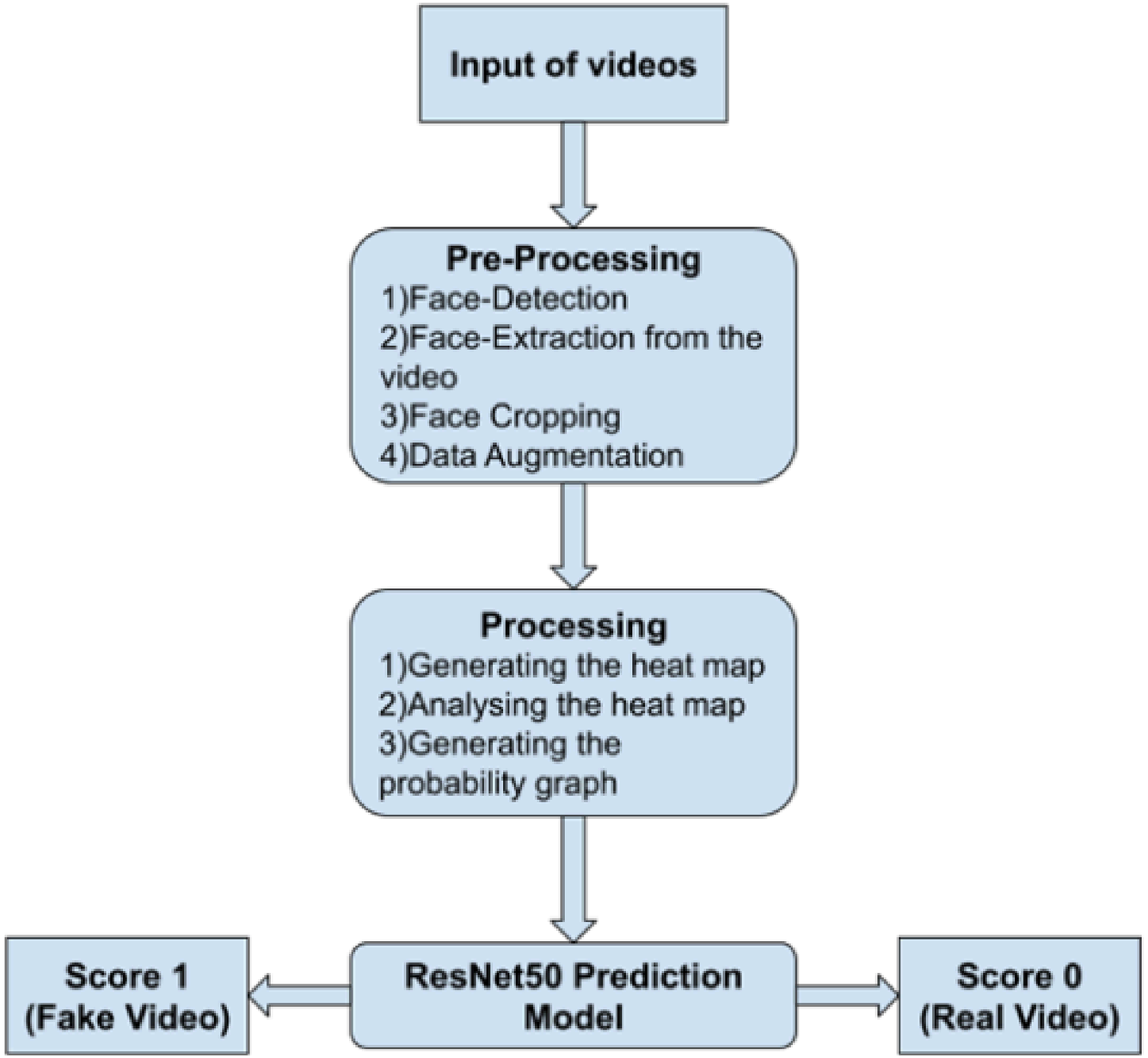


Fig 2. Flowchart of the proposed model

Traing & Used Dataset

CelebDf + DFDC dataset

- 전 세계적으로 인정받는 현실적인 데이터셋을 활용함으로써 모델은 최적의 결과를 내도록 훈련됨

- 적은 수의 epochs를 사용하는 효율적인 모델 훈련을 도움

  *epochs: 40, acc: 97%

Used Dataset

- 개정된 딥페이크 생성 방법을 사용하여 795개의 합성 비디오와 408개의 실제 비디오를 사용

  *테스트 비디오 158개/ 유튜브 영상 250개

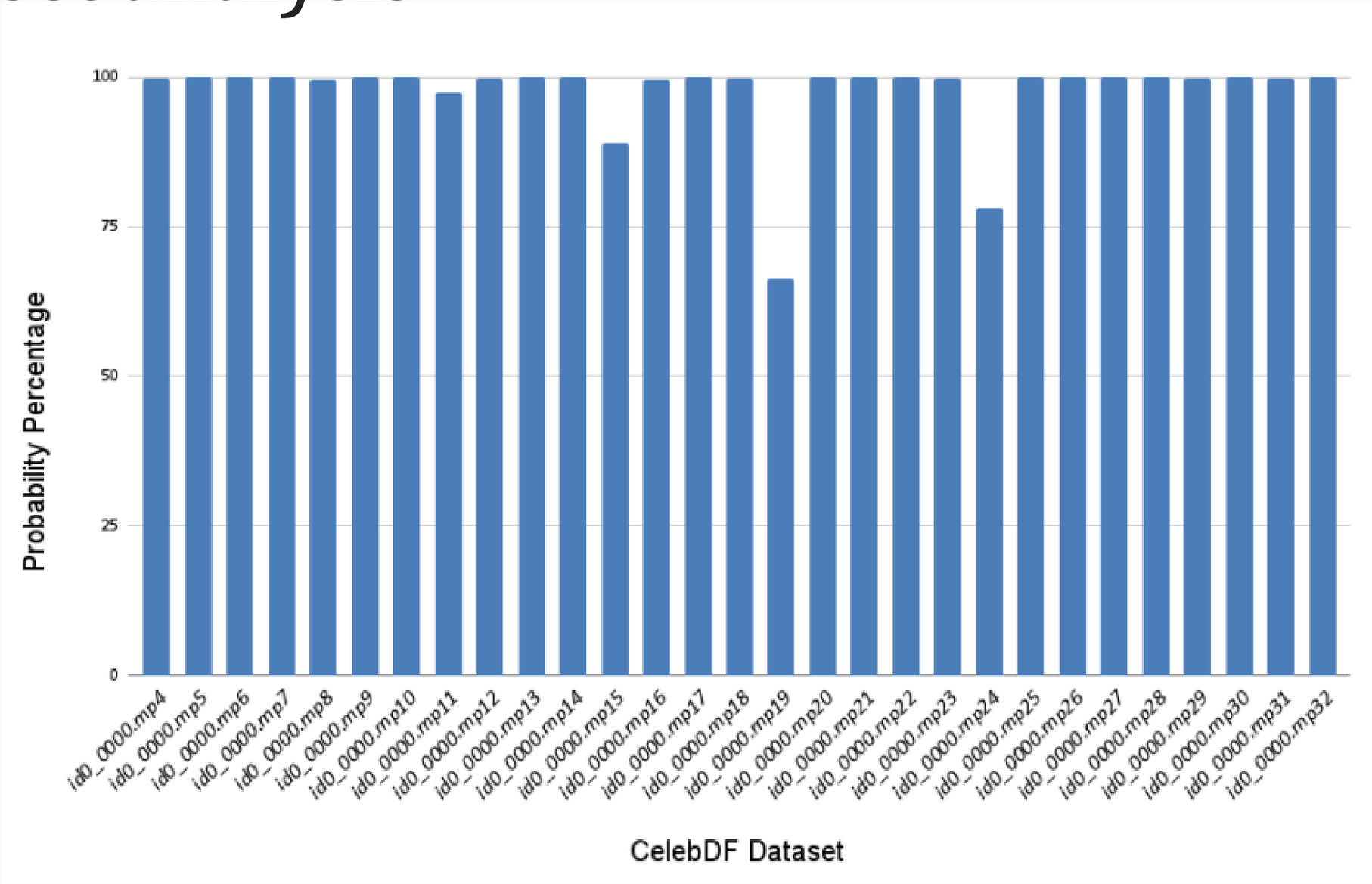
- 초당 30프레임의 속도로 13초 길이의 영상을 사용

Dataset Example

a b

Fig 3. (a) Video frame from real video (Left Side) (b) Video frame from Fake video (Right Side)

Dataset analysis



Potential gaps from previous research

model	top-1 err.	top-5 err.
VGG-16 [41]	28.07	9.33
GoogLeNet [44]	-	9.15
PReLU-net [13]	24.27	7.38
plain-34	28.54	10.02
ResNet-34 A	25.03	7.76
ResNet-34 B	24.52	7.46
ResNet-34 C	24.19	7.40
ResNet-50	22.85	6.71
ResNet-101	21.75	6.05
ResNet-152	**21.43**	**5.71**

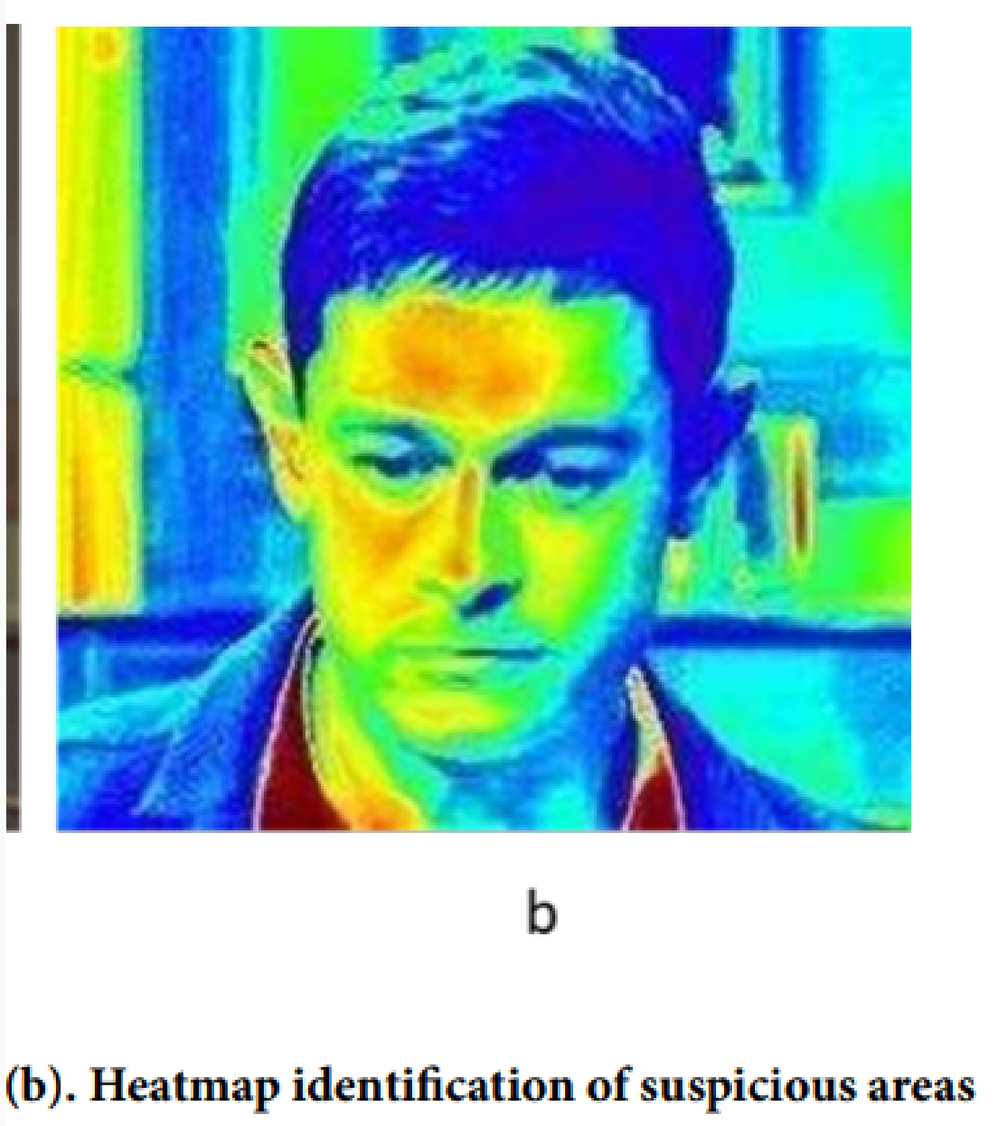
Fig 5. ResNet50 error percentage of referred model

- 오류율 22% -> 최대 5%
- ResNet50모델에 CNN 아키텍처를 적용, surface scraping 기법 활용

Heatmap

Heatmap - 프레임에서 중요한 영역을 표시함

- 비전문가도 시각적으로 딥페이크 여부를 판별하는 데 유용한 세부사항을 쉽게 확인할 수 있음.
- 모든 픽셀이 동일하게 처리되는 것은 아님, 눈 입 등 특정 영역이 더 많은 정보를 제공할 수 있음.
- 따뜻한 색상일수록 더 중요한 영역을 나타냄
- 모델의 학습과정에서 히트맵 이상징후를 딥페이크와 연관시킴

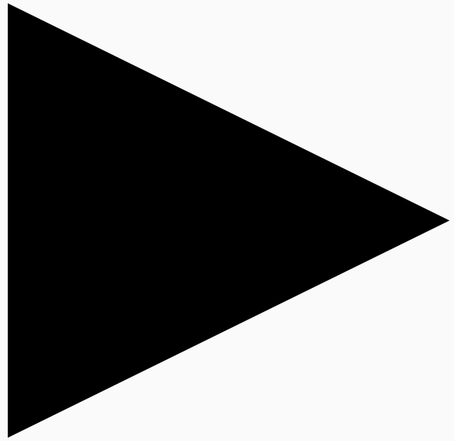


(b). Heatmap identification of suspicious areas

Heatmap Generation

Backpropagating

출력 클래스(예: 실제 또는 딥페이크)에 대한 입력 이미지의 기울기를 역전파하여 생성



Grad-CAM

기울기를 히트맵으로 시각화하는 기법

Heatmap

Relation between results & epochs

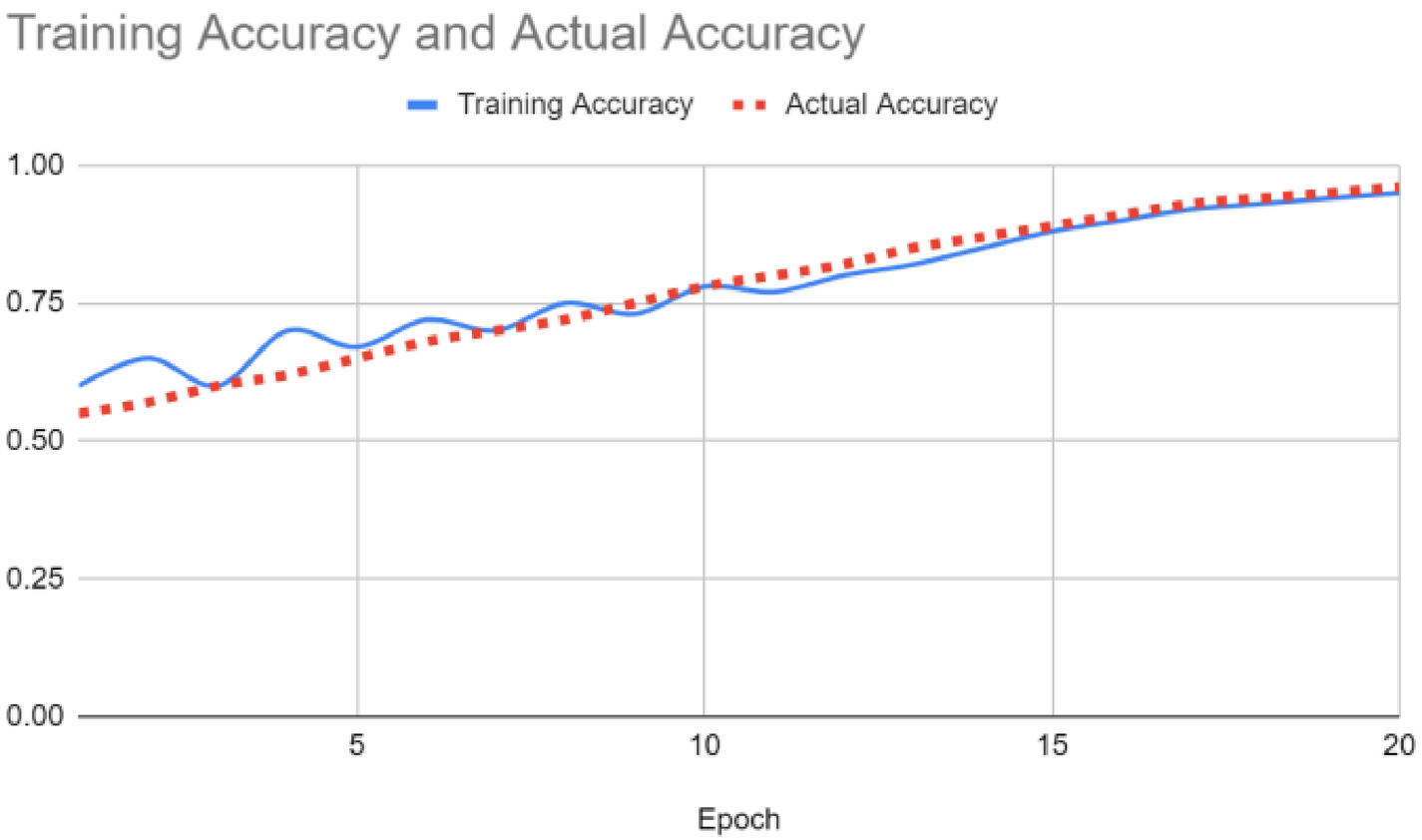


Fig 7. Epoch to accuracy graph of the model

Conclusion

- Combine of ResNet50 model + image scraping technique
- Reduced error rates, Increased detection accuracy by 97%
- Reduction in false positives by 25%
- Shows practical applicability and effectiveness in realworld scenarios, where rapid and accurate deepfake detection is paramount.

Limitations

- Remains room for further enhancement and refinement
- Computational time can be improved in the future via several additional features.

Q & A