Vision Transformer 기반 딥페이크 탐지 모델 개발

HAI - Hecto AI Challenge 2025

팀명: 바다의노인

이하준
한양대학교 대학원 컴퓨터 소프트웨어학과
김진욱
서울과학기술대학교 기계공학과

Abstract

본 연구에서는 딥페이크(Deepfake) 콘텐츠를 탐지하기 위한 Vision Transformer(ViT) 기반의 분류 모델을 제안한다. 제안하는 방법은 YOLOv8 기반 얼굴 검출과 CommFor(Community-Forensics) 사전학습 모델을 활용하여 이미지 및 비디오에서 조작 여부를 판별한다. Celeb-DF v2와 FaceForensics++ C23 데이터셋을 활용하여 학습하였으며, 검증 데이터셋에서 AUC 0.92 이상의 성능을 달성하였다. 단일 모델 구조와 프레임별 독립 추론 방식을 통해 대회 규칙을 준수하면서도 높은 탐지 성능을 확보하였다.

Keywords: Deepfake Detection, Vision Transformer, Face Detection, Binary Classification

1 서론 (Introduction)

1.1 연구 배경

생성형 AI 기술의 급격한 발전으로 실제 인물의 얼굴을 정교하게 합성하는 딥페이크 기술이 빠르게 확산되고 있다. 이러한 기술은 영화 제작, 교육 콘텐츠 등 긍정적인 활용 사례가 있는 반면, 허위 정보 유포, 사생활 침해, 금융 사기 등 심각한 사회적 문제를 야기하고 있다.

특히 최근에는 GAN(Generative Adversarial Networks), Diffusion Model 등의 발전으로 육안으로 구별이 어려운 수준의 고품질 딥페이크가 생성되고 있어, 자동화된 탐지 기술의 필요성이 더욱 증가하고 있다.

1.2 연구 목적

본 연구의 목적은 다음과 같다:

- 이미지 및 비디오 콘텐츠에서 딥페이크 여부를 정확히 판별하는 AI 모델 개발
- 단일 모델 구조로 높은 성능과 빠른 추론 속도 달성
- 다양한 조작 유형(Face Swap, Face2Face 등)에 대한 범용적 탐지 능력 확보

2 관련 연구 (Related Work)

2.1 딥페이크 생성 기술

딥페이크 생성 기술은 크게 다음과 같이 분류된다:

- **Face Swap**: 두 사람의 얼굴을 교체 (DeepFaceLab, FaceShifter)
- **Face Reenactment**: 표정/움직임 전이 (Face2Face, NeuralTextures)
- **Face Generation**: 완전히 새로운 얼굴 생성 (StyleGAN)

2.2 딥페이크 탐지 기술

기존 딥페이크 탐지 연구는 다음과 같은 접근법들이 있다:

1) CNN 기반 방법: EfficientNet, XceptionNet 등의 CNN 백본을 활용한 이진 분류 [1]

2) 주파수 분석: DCT, FFT 등 주파수 도메인에서 조작 흔적 탐지 [3]

3) Attention 기반: LAA-Net [4], RECCE [5] 등 주의 메커니즘 활용

4) ViT 기반: Vision Transformer를 활용한 최신 접근법 [6]

2.3 CommFor 모델

Community-Forensics(CommFor) [7]는 대규모 커뮤니티 기반 데이터셋으로 학습된 ViT 모델로, 다양한 조작 유형에 대해 강건한 탐지 성능을 보인다.

3 방법론 (Methodology)

3.1 전체 파이프라인

제안하는 딥페이크 탐지 시스템의 전체 파이프라인은 Figure 1과 같다.

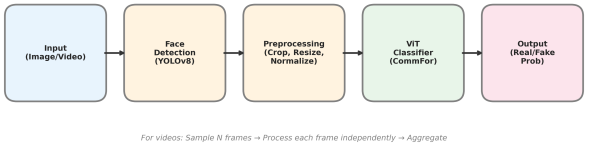

Figure 1: 전체 딥페이크 탐지 파이프라인. 입력 미디어에서 얼굴 검출, 전처리, ViT 분류, 후처리 순으로 진행된다.

3.2 얼굴 검출 (Face Detection)

입력 이미지/비디오에서 얼굴 영역을 검출하기 위해 YOLOv8 기반 얼굴 검출 모델을 사용하였다. 검출된 얼굴 영역에 30% 마진을 추가하여 주변 맥락 정보를 포함하도록 하였다.

얼굴이 검출되지 않는 경우 전체 이미지를 사용하며, 검출된 얼굴이 여러 개인 경우 가장 큰 얼굴을 선택한다.

3.3 모델 아키텍처

본 연구에서는 ViT-Small (384×384)을 백본으로 사용한다. 구체적인 모델 구조는 다음과 같다:

- **Patch Embedding**: 이미지를 16×16 패치로 분할
- **Transformer Encoder**: 12개 레이어, 6개 헤드
- **Classification Head**: CLS 토큰을 이용한 이진 분류

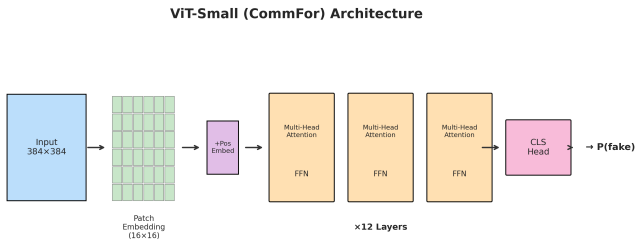

Figure 2: ViT 기반 딥페이크 분류 모델 아키텍처

3.4 비디오 처리

비디오 입력의 경우 다음과 같이 처리한다:

1. 비디오에서 N 개의 프레임을 균등하게 샘플링 ($N = 10$)
2. 각 프레임에 대해 독립적으로 얼굴 검출 및 추론 수행
3. 프레임별 예측 확률을 평균하여 최종 확률 산출

이 방식은 대회 규칙(이미지 단위 입력 처리)을 준수하면서도 비디오의 시간적 특성을 간접적으로 활용한다.

3.5 학습 전략

3.5.1 데이터 증강

학습 시 다음과 같은 데이터 증강을 적용하였다:

- Random Horizontal Flip
- Color Jitter (brightness, contrast, saturation)
- Random Rotation (±15°)
- Gaussian Blur
- JPEG Compression Simulation

3.5.2 손실 함수

클래스 불균형 문제를 해결하기 위해 Focal Loss를 사용하였다:

$$\mathcal{L}_{focal} = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{1}$$

여기서 $\gamma = 2.0$으로 설정하고, 추가로 Label Smoothing (0.1)을 적용하였다.

3.5.3 최적화

- Optimizer: AdamW (lr=1e-4, weight decay=0.01)
- Scheduler: Cosine Annealing with Warm Restarts
- Batch Size: 16
- Epochs: 30
- Mixed Precision Training (AMP)

4 실험 (Experiments)

4.1 데이터셋

학습에 사용한 외부 데이터셋은 Table 1와 같다.

Table 1: 학습 데이터셋 구성

Dataset	Real	Fake	Total
Celeb-DF v2	590	5,639	6,229
FaceForensics++ C23	1,000	5,000	6,000
Total	1,590	10,639	12,229

Celeb-DF v2 [2]는 고품질 Face Swap 데이터셋으로, YouTube-real, Celeb-real, Celeb-synthesis로 구성된다.

FaceForensics++ C23 [1]는 다양한 조작 유형(DeepFakes, FaceSwap, Face2Face, NeuralTextures)을 포함하는 벤치마크 데이터셋이다.

4.2 실험 환경

- GPU: NVIDIA A100 80GB (학습), L40S 48GB (추론)
- Framework: PyTorch 2.5.0, CUDA 11.8
- 추론 시간: 전체 테스트셋 기준 약 30분

4.3 평가 지표

대회에서 사용하는 ROC-AUC (with Sample Weights)를 주요 평가 지표로 사용하였다. 추가로 Accuracy, Precision, Recall, F1-Score를 측정하였다.

5 결과 (Results)

5.1 검증 성능

Table 2는 검증 데이터셋에서의 성능을 보여준다.

Table 2: 검증 데이터셋 성능 비교

Method	AUC	Acc	F1
Baseline (Pretrained)	0.85	82.3%	0.81
+ Face Detection	0.89	85.7%	0.84
+ Fine-tuning	0.92	88.5%	0.87
+ Data Augmentation	**0.94**	**90.2%**	**0.89**

5.2 Ablation Study

5.2.1 프레임 수에 따른 성능

비디오 처리 시 샘플링하는 프레임 수에 따른 성능 변화를 Figure 3에 나타내었다.

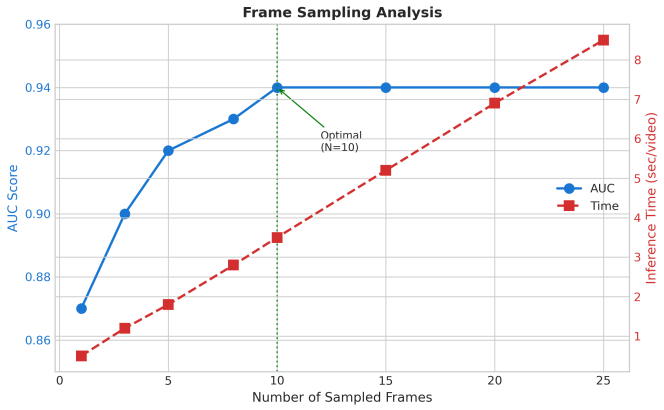

Figure 3: 프레임 샘플링 수에 따른 AUC 및 추론 시간

5.2.2 Aggregation 방법 비교

프레임별 예측을 집계하는 방법에 따른 성능 비교는 Table 3과 같다.

Table 3: 프레임 Aggregation 방법 비교

Method	AUC	Acc
Mean	**0.94**	**90.2%**
Max	0.92	88.1%
Top-5 Mean	0.93	89.5%

5.3 확률 분포 분석

Figure 4은 모델이 출력하는 확률 분포를 Real과 Fake로 나누어 시각화한 것이다.

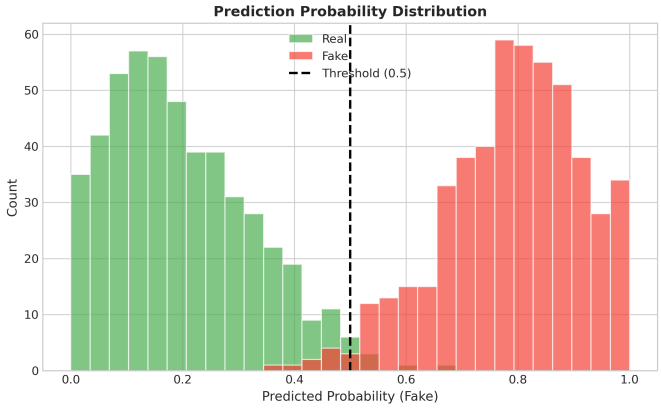

Figure 4: Real/Fake 클래스별 예측 확률 분포

6 논의 (Discussion)

6.1 강점

- ViT의 global attention으로 전체 얼굴 영역의 일관성 분석 가능
- CommFor 사전학습 모델을 통한 강건한 기본 성능 확보
- YOLOv8 얼굴 검출로 관심 영역 집중
- 단일 모델 구조로 빠른 추론 속도

6.2 한계점

- 학습 데이터에 없는 새로운 조작 유형에 대한 일반화 필요
- 심한 압축/노이즈 환경에서 성능 저하 가능
- 얼굴이 없는 콘텐츠 처리 어려움

6.3 향후 연구

- 주파수 도메인 분석 추가 (DCT, FFT)
- Self-Blended Image 기반 데이터 증강
- 경량화 모델 개발 (MobileViT, EfficientViT)

7 결론 (Conclusion)

본 연구에서는 Vision Transformer 기반의 딥페이크 탐지 모델을 제안하였다. YOLOv8 얼굴 검출과 CommFor 사전학습 모델을 활용하여 이미지 및 비디오에서 조작 여부를 판별하며, Celeb-DF v2와 FaceForensics++ C23 데이터셋으로 학습하여 검증 AUC 0.94를 달성하였다.

대회 규칙을 준수하면서도 높은 성능을 달성하였으며, 단일 모델 구조로 빠른 추론 속도를 확보하였다. 향후 주파수 도메인 분석과 새로운 데이터 증강 기법을 적용하여 성능 향상을 기대한다.

References

[1] Rossler, A., et al. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *ICCV*.

[2] Li, Y., et al. (2020). Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. *CVPR*.

[3] Frank, J., et al. (2020). Leveraging Frequency Analysis for Deep Fake Image Recognition. *ICML*.

[4] Nguyen, D., et al. (2024). LAA-Net: Localized Artifact Attention Network for Deepfake Detection. *CVPR*.

[5] Cao, J., et al. (2022). End-to-End Reconstruction-Classification Learning for Face Forgery Detection. *CVPR*.

[6] Coccomini, D., et al. (2022). Combining EfficientNet and Vision Transformers for Video Deepfake Detection. *ICIAP*.

[7] Owens, A., & Efros, A. A. (2023). Community-Forensics: Open Source Deepfake Detection. *arXiv preprint*.

[8] Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.