AI 생성 이미지를 탐지하기 위한 시스템에 관한 연구

이호수, 홍준호, 류지선, 윤선아, 김용강

국립공주대학교

lhs38434126@gmail.com, nolja30@naver.com, wltjsdl4061@naver.com, ycy238@smail.kongju.ac.kr, ygkim@kongju.ac.kr

A Study on the AI-Generated Image Detection Systems

Lee Hosu, Hong Junho, Ryu Jiseon, Yoon Seona, and Yonggang Kim

Kongju National University

요 약

본 연구는 고도화되는 AI 생성 이미지 확산에 대응하고, 단일 탐지 모델의 낮은 일반화 성능을 극복하기 위해 얼굴 탐지 기반의 적응형 앙상블 탐지 시스템을 제안한다. 제안하는 시스템은 Flask-FastAPI 기반의 3-Tier 마이크로서비스 아키텍처로 구현되어 실시간 서비스 환경을 지원한다. 핵심 탐지 알고리즘은 OpenCV의 Haarcascade를 활용한 전처리 단계를 통해 이원화된다. 얼굴 영역이 탐지된 경우, 픽셀 미세 패턴 성능이 우수한 GenDet(0.3)과 ResNet(0.7)을 결합한 가중 앙상블을 수행하며, 얼굴이 탐지되지 않은 경우 GenDet 단일 모델을 사용하여 배경 및 사물 이미지에 유연하게 대응한다. 실험 결과, 제안된 적응형 시스템은 다양한 이미지 도메인에서 안정적인 성능을 보였으며, 특히 ROC-AUC 및 재현율(Recall) 지표에서 유의미한 향상을 달성하였다. 본 연구는 지능형 모델 선택 로직을 갖춘 시스템을 실제 서비스 환경에 배포 가능한 형태로 구축하여 AI 이미지 탐지 기술의 실용화와 신뢰도 제고에 기여한다.

Ⅰ. 서 론

최근 딥러닝 기반 이미지 생성 모델의 발전으로, 사람의 눈으로는 실제 사진과 거의 구분하기 어려운 고품질 합성 이미지가 대량으로 생성되고 있다. 이러한 기술은 창작과 콘텐츠 제작에 새로운 가능성을 제공하는 한편, 딥페이크 사기, 조작된 인물 사진을 이용한 여론 조작, 저작권 및 윤리 문제 등 다양한 위험을 발생시키며 디지털 이미지에 대한 신뢰를 약화시키고 있다. 특히 얼굴 이미지는 개인 식별, 신뢰 형성, 감정 전달 등에 직접적으로 연결되기 때문에, 생성 · 조작된 얼굴 이미지를 정확히 탐지하는 기술은 사회적 · 보안적 측면에서 중요한 과제가 되고 있다.

기존 연구들은 주로 ResNet과 같은 CNN 기반 분류기를 이용해 픽셀 공간에서 실제 이미지와 합성 이미지를 직접 구분하는 방식을 사용해 왔다 [2]. 이러한 방법은 학습에 사용된 특정 생성 모델과 도메인에 대해서는 높은 정확도를 보이지만, 다른 생성 모델이나 새로운 데이터 분포에 대해서는 성능이 급격히 저하되는 일반화 한계가 있다. 한편, 최근에는 CLIP 기반 표현[3]과 Teacher-Student 구조를 활용하여, 특징 공간에서 실제와 가짜 간의 차이를 학습함으로써 보다 강건한 생성 이미지 탐지를 시도하는 연구도 등장하고 있다[1].

그러나 얼굴과 비얼굴 장면이 섞여 있는 실사용 환경을 전제로 할 때, 단일 구조만으로 모든 경우를 안정적으로 처리하는 데에는 여전히 제약이 존재한다. 본 연구에서는 이러한 문제를 완화하기 위해, 입력 이미지에 얼굴이 존재하는지 여부에 따라 탐지 전략을 다르게 적용하는 하이브리드 생성 이미지 탐지 시스템을 제안한다. 먼저 OpenCV의 Haarcascade 기반 얼굴 검출기를 사용하여 이미지에서 얼굴 영역을 탐지한다. 얼굴이 검출된 경우에는 얼굴 영역을 중심으로 학습된 ResNet 기반 CNN 분류기와, CLIP 특징을 사용하는 GenDet 계열 분류기를 함께 적용하고 앙상블하여 최종 판정을 내린다. 반대로 얼굴이 검출되지 않는 일반 장면 이미지에 대해서는 GenDet 계열 분류기만을 이용해 판별을 수행한다.

Ⅱ. 제안하는 AI 모델

본 연구는 AI 생성 이미지 탐지의 일반화 성능 향상을 목표로 GenDet의 Teacher-Student 기반 학습 전략을 참고하되[1], 입력 표현을 CLIP 비전 특징으로 확장한 변형 구조를 제안한다.

2.1 CLIP 기반 특징 표현

입력 이미지는 전처리 후 CLIP 비전 모델을 통해 고정 차원의 특징 벡터로 변환된다[3]. 구체적으로, 이미지를 224×224로 리사이즈하고 정규화한 뒤, TFCLIPVisionModel의 pooler_output을 사용하여 768차원 특징을 추출한다. 이 특징 벡터는 Teacher, Student, Augmenter의 공통 입력으로 사용되며, CLIP 가중치는 학습 과정에서 업데이트하지 않는다.

2.2 Teacher-Student 및 Augmenter 구조

Teacher와 Student는 동일한 경량 Transformer 기반 헤드 구조를 공유하며, CLIP 특징을 잠재 공간으로 투영하여 Fake 확률을 산출한다. Teacher는 사전 학습을 통해 기준 판별기로 기능하며, Student는 Teacher의 출력을 모방(Real)하거나 분리(Fake)되도록 학습한다. Augmenter는 Fake 특징을 교란하여 Teacher와 Student의 출력을 유사하게 만드는 방향으로 학습되며, Student는 이러한 "더 어려운 Fake"에 대해서도 Teacher와의 출력을 분리하도록 경쟁적인 적대 학습을 수행한다.

2.3 최종 판정

최종적으로 Teacher와 Student의 출력 차이(Discrepancy)를 입력으로 하는 분류기(Classifier)를 학습한다. 추론 시에는 이미지 자체의 직접 분류보다 두 모델의 반응 차이를 기반으로 판단을 내림으로써, 도메인 변화에 대한 일반화 가능성을 높인다.

Ⅲ. 제안하는 AI 생성 이미지 탐지 시스템

3.1 시스템 아키텍처 및 백엔드 구현

본 시스템은 확장성과 유지보수성을 위해 3-Tier 구조(Frontend-Backend-AI Server)로 설계되었다.

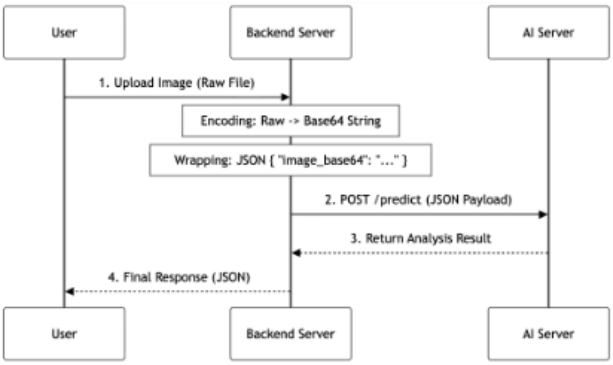

그림 1 데이터 플로우

본 시스템의 전체적인 데이터 처리 및 통신 흐름은 [그림 1]과 같다. 사용자가 업로드한 Raw 이미지는 백엔드 서버에서 Base64 문자열로 인코딩된 후, JSON 페이로드에 담겨 AI 서버로 전송된다

3.2 프론트엔드 및 UX 최적화

사용자 경험을 위해 클라이언트 측에서 10MB 이하의 용량 제한과 포맷 검증을 선행하여 서버 부하를 줄였다. 또한 fetch API 기반의 비동기 통신을 통해 분석 중 로딩 상태(Spinner)를 시각화하고, 결과는 단순 텍스트가 아닌 "AI 사용됨(0.7 이상)", "애매함", "AI 사용 적음" 등의 직관적인 구간 정보와 프로그레스 바로 제공한다.

3.3 적응형 앙상블 방법론

본 연구의 핵심은 이미지 내 얼굴 존재 여부에 따라 추론 전략을 달리하는 것이다.

- 얼굴 영역 탐지: OpenCV의 Haarcascade Classifier를 사용하여 전처리 단계에서 효율적으로 얼굴을 검출한다.
- Case A (얼굴 탐지 시): 인물 딥페이크 가능성이 높으므로, 미세 픽셀 탐지에 강한 ResNet(가중치 0.7)과 의미적 특징을 보는 GenDet(가중치 0.3)을 앙상블하여 최종 점수를 산출한다.
- Case B (얼굴 미탐지 시): 풍경이나 사물 이미지 등 ResNet이 오탐을 일으킬 수 있는 영역에 대해서는 일반화 성능이 뛰어난 GenDet 단일 모델만을 사용한다.

$$S_{final} - \begin{cases} 0.7 \times S_{ResNet} + 0.3 \times S_{Gendet} & \text{if } Case\ A \\ 1.0 \times S_{GenDet} & \text{if } Case\ B \end{cases}$$

Ⅳ. 실험 및 결과

4.1 실험 환경 및 데이터셋

실험은 NVIDIA GeForce RTX 4070 GPU 환경에서 수행되었다. 데이터셋은 인물 영역 검증을 위한 FFHQ, Fake Face Images와 비인물 영역 검증을 위한 Stable Diffusion v1.4 데이터를 사용하여 적응형 알고리즘의 유효성을 검증하였다.

4.2 성능 평가 결과

얼굴 데이터셋에 대한 평가:

얼굴이 포함된 데이터($N = 21,196$)에 대해 세 가지 전략을 비교하였다.

전략 1.

$$Score_{final} = \max(Score_{ResNet}, Score_{GenDet})$$

Accuracy	Precision	Recall	F1-Score
0.73	0.81	0.71	0.71

표 1 전략 1 성능 측정 결과

전략 2.

$$Score_{final} = \frac{(Score_{ResNet} + Score_{GenDet})}{2}$$

Accuracy	Precision	Recall	F1-Score
0.87	0.88	0.87	0.87

표 2 전략 2 성능 측정 결과

전략 3.

$$Score_{final} = 0.7 \times Score_{ResNet} + 0.3 \times Score_{GenDet}$$

Accuracy	Precision	Recall	F1-Score
0.98	0.98	0.98	0.98

표 3 전략 3 성능 측정 결과

단순 결합 방식(Max, Average)은 각각 오탐 증가와 신호 희석(Dilution) 문제로 최적의 성능을 보이지 못했다. 반면, 제안 전략은 얼굴 탐지 시 특화 모델(ResNet)에 가중치(0.7)를 부여해 도메인 전문성을 극대화함으로써, 0.98의 정확도와 F1-Score를 달성하였다. 이는 얼굴 유무라는 사전 정보의 반영이 고정밀 탐지의 핵심임을 시사한다.

사물 및 풍경 데이터셋에 대한 평가

$$Score_{final} = Score_{GenDet}$$

Accuracy	Precision	Recall	F1-Score
0.99	0.99	0.99	0.99

표 4

Ⅴ. 결론

본 연구는 고성능 AI 탐지 모델을 개발하는 단계를 넘어, 이를 실제 사용자가 활용 가능한 웹 서비스로 구현하고 기술적 난제들을 해결하는 데 초점을 맞추었다. 모델링 측면에서는 얼굴 유무에 따른 적응형 앙상블 전략을 통해 단일 모델의 한계를 극복하고 다양한 도메인에서 높은 정확도를 확보하였다. 시스템 구현 측면에서는 3-Tier 아키텍처, Mock API, Base64 최적화 등을 통해 안정적이고 빠른 응답 속도를 가진 시스템을 구축하였다. 결과적으로 본 연구는 지능형 탐지 알고리즘과 견고한 웹 엔지니어링 기술의 결합을 통해 AI 생성 이미지 문제에 대한 실용적인 대응책을 제시하였다. 향후 연구에서는 최신 Diffusion 모델에 대한 데이터셋 확충과 비디오 스트리밍 처리를 위한 경량화 기술 연구를 진행할 계획이다.

ACKNOWLEDGMENT

본 결과물은 2025년도 교육부 및 충청남도의 재원으로 충남RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과임(2025-RISE-12-003).

참 고 문 헌

[1] Zhu, Mingjian, et al. "Gendet: Towards good generalizations for AI-generated image detection" *arXiv preprint arXiv:2312.08880*, 2023.

[2] He, K., Zhang, X., Ren, S., & Sun, J., "Deep residual learning for image recognition" in *IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[3] Radford, A., et al., "Learning transferable visual models from natural language supervision" in *International conference on machine learning*, pp. 8748-8763, 2021.