CLIP 유사도 기반 Multimodal RAG 검증 기법

김은오, 이창근, 윤수연*
㈜엔투솔루션, *국민대학교

eunoh.kim@ntoday.kr, changgeun.lee@ntoday.kr, *1104py@kookmin.ac.kr

CLIP similarity-based Multimodal RAG verification method

Eun-Oh Kim, Chang-Geun Lee, Soo-Yeon Yoon*
N2Soultion Co.,Ltd., *Kookmin Univ.

요 약

본 연구는 RAG(Retrieval-Augmented Generation) 기법을 멀티모달(Large Multimodal Model, LMM) 환경으로 확장하여 딥페이크 탐지의 재현율(Recall)을 개선하는 전략을 제시한다. 실험 과정에서 실제 사진으로 분류된 이미지를 CLIP(Contrastive Language-Image Pre-training) 유사도 검색을 통해 조작 단서 키워드와 매칭하고, 이를 체크리스트 형태로 활용하여 오분류를 최소화하였다. 그 결과 정확도가 약 2~3%정도 하락하는 대신, 재현율이 최대 6.4%까지 향상되는 트레이드오프가 관찰되었다. 이는 RAG 파이프라인이 누락된 위조 사례들을 추가로 포착할 수 있음을 시사한다. 본 연구는 향후 LMM 생태계 전반의 RAG 기법 활용 가능성을 확대하고, 의료 영상 분석, 자율주행, 보안 감시 등 다양한 분야에서 신뢰성과 안정성을 갖춘 멀티모달 정보 처리 모델 설계의 기반이 될 것으로 기대된다.

Ⅰ. 서 론

최근 딥페이크 탐지 분야에서 주목받고 있는 LMM 은, 전문지식이 부족한 사용자도 결과를 쉽고 효율적으로 얻을 수 있을 뿐 아니라, 위조 여부와 함께 판단 근거까지 제시해주는 새로운 탐지 방법으로 연구되고 있다. 하지만 다차원 데이터를 통합적으로 처리하는 데 따른 기술적 난관과 정보 환각(hallucination) 문제로 인해 아직 초기 단계에 머물러 있다.

최근 RAG 기법은 외부 지식을 검색하여 문맥(context)으로 활용함으로써 이러한 환각 문제를 완화하고 정밀한 추론을 가능케 하는 전략으로 주목받고 있다. 하지만 주로 텍스트 기반 LLM 연구로 집중되어 멀티모달에서의 활용은 충분히 이뤄지지 않고 있다[1].

Ⅱ. 실험 설계

1. 데이터셋

본 연구에서는 실험을 위해 두 가지 데이터셋을 사용하였다. 첫 번째는 DeepFakeFace 로서, IMDB-WIKI 를 기반으로 Stable Diffusion v1.5, Stable Diffusion Inpainting, InsightFace 등의 기법을 적용해 구축된 데이터셋이다. 이처럼 다양한 합성 기법을 반영해 폭넓은 딥페이크 사례를 포괄한다. [2]

두 번째 데이터셋은 Seq-DeepFake 로, 단일 변환(Face-Swap 등)에 그치지 않고 여러 단계의 연속적 얼굴 조작을 수행한 이미지를 제공한다. 이는 보다 복합적인 위조 양상을 반영함으로써, 모델의 일반화 능력을 보다 엄밀하게 검증할 수 있다. [3]

본 연구에서는 DeepFakeFace 데이터셋을 GPT-4o 모델로 레이블을 생성한 뒤, 이를 키워드 형태로 추출하여 데이터베이스(DB)에 저장하였다. 해당 이미지와 키워드는 OpenAI 의 CLIP[4] 모델을 사용해 512 차원으로 임베딩하여 동일한 DB 에 저장하였다.

테스트 세트는 DB 에 저장되지 않은 10%의 DeepFakeFace 와 Seq-DeepFake 로 구성하여 모델의 일반화 성능과 RAG 적용 효과를 검증할 수 있도록 설계했다.

2. 실험 방법

본 연구의 실험은 LMM 기반 딥페이크 탐지 모델에 RAG 기법을 적용함으로써 발생하는 성능 변화를 관찰하는 것을 목표로 한다. 그러나 기존 LMM 은 이미지나 영상을 base64 형태로 처리하기 때문에 외부 지식과 직접적으로 대조하거나 검색하기에 제약이 있었다. 이를 극복하기 위해, 본 연구에서는 CLIP 의 유사도 계산을 활용하는 외부 파이프라인을 별도로 구성하여 RAG 기법을 적용하였다.

RAG 파이프라인은 다음과 같이 구성된다. 먼저 모델이 이미지 특징을 바탕으로 딥페이크 여부를 판별한 뒤 실제 사진으로 분류할 경우 CLIP 유사도 계산을 통해 단서 키워드와 결합한다. 이후 체크리스트 형태로 단서를 참조하여 다시 탐지를 수행하도록 설계되었다. 이때 DB 는 딥페이크 단서만을 저장·검색되도록 구성했는데, 이는 실제 사진 단서를 함께 저장할 경우 CLIP 유사도 계산에 지나치게 좌우되는 문제가 발생하기 때문이다. 이러한 방법을 통해 CLIP 모델에 전적으로 의존하지 않으면서도, 맥락(Context) 기반의 성능 향상을 기대할 수 있는 차별화된 설계를 구축했다.

실험에 활용한 딥페이크 탐지 모델은 표 1 과 같이 총 네 가지로 구성된다. 다양한 모델 구성을 통해 RAG 적용 전 후의 딥페이크 탐지 성능 변화를 종합적으로 분석한다.

표 1. 실험 모델 구성

Model	Detail
Gemini-1.5-flash (Zero-shot)	Google [5]
Gemini-1.5-flash (fine-tuned)	2000 개의 딥페이크 이미지로 파인 튜닝
Llama-3.2-90B-Vision-Instruct	Meta [6]
Llama-3.2-11B-Vision-Instruct	Meta

Ⅲ. 실험 및 성능 평가 분석

본 연구에서는 먼저 모델의 zero-shot 성능을 평가한 뒤, RAG 적용 전후 성능 차이를 관찰하였다. 정확도(Accuracy), F1 Score, Recall 을 주요 지표로 삼았으며, 특히 Recall 은 잠재적 위·변조 콘텐츠를 놓치지 않는 모델의 민감도 측면에서 핵심 지표로 활용하였다.

표 2 는 기존 LMM 모델의 딥페이크 탐지 성능을 정리한 결과이다. Gemini-1.5-flash 모델의 경우 zero-shot 에서 Accuracy 55.6, F1 20.0, Recall 11.0 으로 낮은 성능을 보였기에, 파인튜닝을 통해 성능 향상된 버전으로 실험을 진행하였다.

표 2. 기존 LMM 모델 탐지 성능

Model	Accuracy	F1-score	Recall
Gemini-1.5-flash(zero-shot)	55.6	20.0	11.0
Gemini-1.5-flash(Fine-tuning)	**82.5**	**81.4**	**77.0**
Llama-3.2-11B-Vision-Instruct-Turbo	59.0	57.29	55.0
Llama-3.2-90B-Vision-Instruct-Turbo	74.7	70.9	71.4

표 3 은 RAG 기법을 적용한 뒤의 딥페이크 탐지 성능이다. 전반적으로 성능이 개선된 것으로 나타났으며, Gemini-1.5-flash(Fine-tuning) 모델뿐 아니라 Llama-3.2 계열 모델에서도 일정 수준 이상의 성능 향상을 확인하였다.

표 3. RAG 적용 후 탐지 성능

Model	Accuracy	F1-score	Recall
Gemini-1.5-flash(Fine-tuning)	**80.0**	**80.6**	**83.0**
Llama-3.2-11B-Vision-Instruct-Turbo	56.0	56.9	58.0
Llama-3.2-90B-Vision-Instruct-Turbo	73.3	71.5	77.8

결과에 따르면, RAG 기법 적용 시 대부분 모델에서 재현율이 유의미하게 향상되었으나 정확도나 F1 Score 가 일부 하락하는 트레이드오프가 발생했다. Gemini-1.5-flash(Fine-tuning)의 경우 Recall 이 6.0% 상승했으나 F1 과 Accuracy 가 각각 0.8%, 2.5% 하락했고, Llama-3.2-11B-Vision-Instruct-Turbo 도 Recall 이 3.0% 오르는 대신 F1 과 Accuracy 는 각각 0.7%, 3.0% 내려갔다. 반면 Llama-3.2-90B-Vision-Instruct-Turbo 는 Recall 이 6.4% 상승하면서 F1 도 0.8% 함께 증가하여 대규모 파라미터 모델에서 RAG 효과가 더 두드러짐을 보여주었다. 결국, RAG 파이프라인은 LMM 기반 딥페이크 탐지 모델에 외부 지식을 결합함으로써 재현율을 높이고, zero-shot 환경에서도 일정 수준 이상의 성능 개선이 가능하다는 실용적 가치를 제시한다.

Ⅳ. 결론

본 연구는 LMM 기반 딥페이크 탐지에서 RAG 접근법을 제안하여, CLIP 모델의 유사도 계산과 체크리스트 기반 파이프라인으로 재현율을 약 3~6.4% 개선하였다. 이를 통해 추가 위조 사례를 놓치지 않고 안정적으로 검출할 수 있으며, 학술적으로도 RAG·CLIP 을 결합한 모듈형 탐지 파이프라인 사례로 활용도가 높다. 향후에는 외부 지식 베이스의 품질·최신성을 유지하고, 음성·영상·센서 데이터로 확장하는 과정에서 발생할 수 있는 처리량 문제를 해결하기 위한 후속 연구가 필요하다. 이러한 시도는 의료 영상 분석, 자율주행, 객체 탐지 등 멀티모달 분석 전반에서 신뢰도와 안정성을 높이는 데 유용할 것으로 기대된다.

참 고 문 헌

[1]. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, 그리고 D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459- 9474, 2020.

[2]. H. Song, S. Huang, Y. Dong, 그리고 W.-W. Tu, "Robustness and generalizability of Deepfake Detection: A study with Diffusion Models," unpublished, 2023.

[3]. R. Shao, T. Wu, 그리고 Z. Liu, "Detecting and recovering sequential deepfake manipulation," Proc. European Conf. Computer Vision (ECCV), Cham: Springer Nature Switzerland, pp. 712-728, Oct. 2022.

[4]. G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, 그리고 L. Schmidt, "OpenCLIP (Version v0.1)," Zenodo Software Repository, DOI: 10.5281/zenodo.5143773, 2021.

[5]. G. G. Team, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024. [Online]. Available: https://goo.gl/GeminiV1-5

[6]. J. Chi, U. Karn, H. Zhan, E. Smith, J. Rando, Y. Zhang, 그리고 M. Pasupuleti, "Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations," arXiv preprint, arXiv:2411.10414, 2024.