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., *Kookmin Univ.

요 약

본 연구는 RAG(Retrieval-Augmented Generation) 기법을 멀티모달(Large Multimodal Model, LMM) 환경으로 확장하여 재현율(Recall)을 개선하는 전략을 제시한다. 최근 텍스트 기반 RAG 기법을 통해 환각(Hallucination)을 억제하고 모델의 정확도를 향상시키는 연구가 활발히 진행되고 있으나, 이미지를 입력으로 받는 멀티모달 환경에서의 RAG 기법 효과는 충분히 검증되지 않았다. 이에 본 연구는 LMM 을 활용한 딥페이크 탐지 과정에 RAG 기법을 도입하여 탐지 성능 변화를 관찰하고자 하였다. 연구에서는 LMM 이 실제 사진으로 분류한 이미지들을 CLIP(Contrastive Language-Image Pre-training) 기반 유사도 검색을 통해 조작 단서 키워드와 매칭한 후, 이를 체크리스트 형태로 활용하여 오분류를 최소화하는 방안을 제안하였다. 실험 결과, 정확도가 약 2~3%정도 하락하는 대신 재현율이 최대 6.4%까지 향상되는 트레이드오프가 관찰되며, 파라미터 규모가 큰 모델일수록 개선 폭이 더욱 두드려져 기존 탐지 과정에서 누락된 위조 사례도 추가로 포착할 수 있었다. 이를 통해 본 연구에서 제안하는 RAG 파이프라인이 딥페이크 탐지 성능을 효과적으로 향상시킬 수 있음을 확인하였다. 나아가 본 연구는 멀티모달 기반 딥페이크 탐지의 새로운 해법을 제시함으로써, LMM 생태계 전반에서 RAG 기법의 활용 가능성을 확대하고, 향후 의료 영상 분석, 자율주행, 보안 감시 등 다양한 분야에서 신뢰성과 안정성을 갖춘 멀티모달 정보 처리 모델 설계의 토대를 마련할 수 있을 것으로 기대된다.

Ⅰ. 서 론

1. 연구 배경

최근 인공지능 기술의 발전으로 딥페이크(Deepfake) 콘텐츠 생성이 더욱 정교해지고 있다. 이에 따라 영상, 이미지, 오디오 등 다양한 멀티모달 데이터를 기반으로 한 조작 콘텐츠가 손쉽게 유포되고 있으며, 이는 뉴스, 소셜 미디어, 보안 인증, 법적 증거물 등 신뢰성 있는 정보 전달이 필수적인 분야에 심각한 위협을 초래하고 있다. 기존 탐지 기법들은 머신러닝 및 딥러닝을 중심으로 발전하였으나, 새롭게 등장하는 합성·위조 콘텐츠들을 완벽하게 식별하기에는 한계가 존재한다. 더군다나 이러한 접근 방식은 전문지식과 복잡한 환경 설정을 필요로 하여, 일반 사용자가 접근하기 어렵다는 문제점도 지적되고 있다[1].

이러한 상황에서 최근 LMM 은 뛰어난 접근성과 사용성으로 전문지식이 부족한 사용자도 딥페이크 탐지 결과를 쉽고 효율적으로 얻을 수 있다는 강점을 보유하고 있어 새로운 딥페이크 탐지 방법으로 연구되고 있다. 뿐만 아니라 단순히 위조 여부를 판별하는 데 그치지 않고, 판단 근거를 인간이 이해할 수 있는 방식으로 설명함으로써 투명성을 크게 개선할 수 있다는 점도 주요 장점이다. 그러나 LMM 기반 딥페이크 탐지는 영상, 이미지, 오디오 등 다차원 데이터를 통합적으로 처리해야 하는 기술적 난관과 정보 환각(hallucination) 문제로 인해 초기 단계에 머물러 있다. [2]

2. 연구의 필요성

딥페이크 탐지 분야에서 발생하는 환각 현상은 모델의 정확도와 신뢰도를 저하시키는 주요 원인 중 하나이다. 따라서 보다 높은 신뢰성을 보유한 LMM 기반 딥페이크 탐지 모델을 개발하기 위해서는 이러한 문제를 반드시 해결해야 한다.

최근 RAG 기법은 외부 지식을 검색하여 context 로 활용함으로써 이러한 환각 문제를 완화하고 정밀한 추론을 가능케 하는 전략으로 주목받고 있으나, 주로 텍스트 기반 LLM 연구로 집중되어 있다는 한계가 있다. 일부 멀티모달에 관한 연구 역시 텍스트 입력에 대한 이미지 검색을 수행하므로, 딥페이크 탐지와 같이 이미지를 입력받는 상황에서의 RAG 연구는 미흡한 상황이다.

3. 연구 방법

본 연구는 RAG 기법을 LMM 기반 딥페이크 탐지 환경에 적용하여 재현율을 개선하는 전략을 제시한다. 설계된 RAG 기법은 다음과 같은 절차로 진행된다.

먼저, 모델이 1 차적으로 딥페이크 여부를 탐지한다. 실제 사진으로 판단된 경우 사전에 구축된 RAG 파이프라인을 활용하여, 해당 사진과 관련성이 높은 딥페이크 단서 키워드를 CLIP 기반의 유사도 검색을 통해 결합한다. 이후, 프롬프트 엔지니어링을 통해 딥페이크 단서 키워드를 체크리스트로 활용하는 과정을 거친다. 이러한 접근 방식을 통해 기존 탐지 과정에서

누락된 위조 사례를 추가로 포착하고, 전반적인 탐지 성능을 향상시키고자 한다.

본 연구는 LMM 기반 딥페이크 탐지의 새로운 해법을 제시함으로써, LMM 전반에서 RAG 기법의 활용 가능성을 확장하고, 향후 이미지를 입력으로 하는 다양한 분야에서 신뢰성과 안정성을 갖춘 모델 설계를 가능하게 할 것이다.

Ⅱ. 관련 연구

2.1. RAG

RAG 는 LLM 의 성능을 강화하기 위해 외부 지식 소스(External Knowledge Sources)를 통합하는 기술로, 검색(Retriever),증강(Augmentation), 생성(Generator) 세 가지 주요 단계로 작동한다. 검색 단계에서는 먼저 사용자의 질의와 가장 관련성이 높은 정보를 외부 데이터베이스에서 선별한다. 이어서 증강 단계에서 검색된 정보를 LLM 이 활용할 수 있도록 가공하거나 가중치를 부여하는 작업을 수행한다. 마지막으로 생성 단계에서 증강된 정보를 바탕으로 사용자 질의에 적합한 최종 답변을 생성한다. 이러한 단계적 구조를 통해 RAG 는 정확하고 신뢰할 수 있는 응답을 제공한다. [3]

2.2. CLIP Model

텍스트 중심 RAG 접근법을 멀티모달 환경으로 확장하려면, 시각·음성 등 다양한 데이터를 텍스트 정보와 매끄럽게 연결할 수 있는 매개 기술이 필수적이다. 여기서 CLIP(Contrastive Language-Image Pre-training)은 약 4 억개의 대규모 이미지-텍스트 쌍을 대조 학습(Contrasitive learning)하여 양자 간 의미를 코사인 유사도로 계산하는 모델로, 이미지 내 시각적 특징을 텍스트 형태의 키워드나 설명문과 효율적으로 매핑하는 데 탁월한 성능을 보인다. 이 같은 장점은 이미지를 입력으로 받는 LMM 딥페이크 탐지 모델에 외부 지식을 결합할 수 있는 기반을 제공한다. [4] [5]

Ⅲ. 실험 설계

1. 실험 환경

1.1. 데이터셋

본 연구에서는 딥페이크 관련 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 것이다. 이 데이터셋에는 약 90,000 개의 딥페이크 이미지와 30,000 개의 실제 이미지가 포함되어 있으며, 다양한 합성 기법을 반영해 폭넓은 딥페이크 사례를 포괄한다. [6]

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. [7]

본 연구에서는 DeepFakeFace 를 대상으로 GPT-4o 모델을 활용하여 레이블 정보를 생성하고, 이를 키워드 형태로 추출해 데이터베이스(DB)에 저장하였다. 또한 OpenAI 의 CLIP(ViT-B/32) 모델로 이미지와 키워드 간 유사도를 계산한 뒤, 512 차원 임베딩 형태로 DB 에 함께 저장하였다.

1.2. 학습 모델

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
Table 1. Experimental model configuration

Model	Detail
Gemini-1.5-flash (Zero-shot)	Google [8]
Gemini-1.5-flash (fine-tuned)	2000 개의 딥페이크 이미지로 파인 튜닝
Llama-3.2-90B-Vision-Instruct	Meta [9]
Llama-3.2-11B-Vision-Instruct	Meta

표 2 는 성능 평가 과정에서의 데이터셋 구성을 나타낸다. 데이터셋은 DB 에 저장되지 않은 10%의 DeepFakeFace 와, 새로운 생성 기법(Seq-DeepFake) 이미지를 테스트 데이터셋으로 구축하였다. 이를 통해 모델의 일반화 성능과 RAG 적용에 따른 성능 변화를 검증할 수 있도록 실험 환경을 설계하였다.

표 2. 실험 데이터셋 구성
Table 2. Experimental dataset configuration

Dataset	Component
RAG DB	DeepFakeFace Image + Keyword
Test	DeepFakeFace(87.5%) + Seq-DeepFake(12.5%)

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 파이프라인의 흐름은 다음과 같다. 먼저 모델이 이미지의 특징을 바탕으로 딥페이크 여부를 판별한다. 이후 실제 사진으로 분류한 이미지 리스트를 CLIP 유사도 계산을 통해 텍스트로 구조화된 단서 키워드와 결합한다. 이를 체크리스트(외부 지식)형태로 참조하여 다시 한번 탐지를 수행함으로써, 정확도와 재현율(Recall)사이의 균형점을 조정한다.

해당 RAG 는 DB 에 딥페이크 단서만을 저장하고 검색하도록 설계하였다. 이는 일반적인 RAG 방식을 적용할 경우, CLIP 의 탐지 성능에 지나치게 좌우되는 현상이 발생했기 때문이다. 이러한 차별화된 설계를 통해 CLIP 모델 성능에 전적으로 의존하지 않으면서도, Context 기반의 성능 향상을 기대할 수 있었다.

이와 유사한 방법으로 초기에 다수 딥페이크 단서를 체크리스트 형태의 프롬프트로 입력하는 방법이 있다. 하지만 이는 토큰 사용량이 크게 늘어나 답변의 품질과 속도가 저하될 우려가 있으며 모델이 지나치게 딥페이크로 판단하려는 편향이 발생할 수 있다. 실제로 표 3 에서 보이는 것처럼, LLaMA 모델에 탐지 관련 프롬프트를 구체적으로 제공할수록, 딥페이크로 분류하는 편향이 높아짐을 확인할 수 있었다.

표 3. 프롬프트에 따른 탐지 편향 (LLaMA-3.2-11B-Vision-Instruct-Turbo)
Table 3. Bias by Prompt (LLaMA-3.2-11B-Vision-Instruct-Turbo)

Prompt	Real	Fake
기본 프롬프트(이진 분류)	68.0%	32.0%
기본 프롬프트+판단 근거 요구+탐지 방법 제시	32.0%	68.0%

이러한 편향은 본 연구의 RAG 기법 역시 발생할 수 있으므로, 프롬프트 엔지니어링을 통해 제약 조건을 설정해 주었다. 그림 1 은 본 연구 프롬프트의 일부 내용으로, 가정, 유추 등을 금지하고 중립, 높은 가능성 등을 강조하였다.

```
### Instructions:
1. For each keyword, carefully examine the image and objectively verify whether
the described clue is present. Avoid making assumptions and clearly state if no issues are observed.
2. Maintain neutrality in your evaluation and do not assume the presence of
  manipulation unless there is strong and consistent evidence across multiple clues.
3. If anomalies are observed, provide a factual description without inferring
  intent or confirming manipulation prematurely.
4. If no clear anomalies are found, highlight natural variations or explain
why observed features may align with expected real-world variations.
5. There is a high probability that the provided list of image URLs are REAL;
  only categorize them as FAKE if there is a high probability.
```

그림 1. 편향 방지를 위한 프롬프트
Figure 1. Prompts to avoid bias

또한 RAG 기법을 사전에 미세 조정된 모델뿐만 아니라 zero-shot 상태의 LMM 모델에도 적용함으로써, 제안한 접근법이 다양한 환경에서 일관된 성능 개선을 가져오는지 검증한다. 이를 통해 실제 산업·서비스 환경에서 모델 학습 자원이 제한된 경우에도, 본 연구가 제안하는 멀티모달 RAG 접근이 충분한 가치를 지닐 수 있음을 검증하고자 하였다.

그림 2 에 제시된 예시와 같이, CLIP 은 눈동자 반사 이상, 치아 배열 불균열등 딥페이크 특유의 특징을 성공적으로 매칭하는 모습을 보였으며, 이를 통해 이미지와 유사도가 높은 키워드를 계산할 수 있다는 사실을 확인하였다.

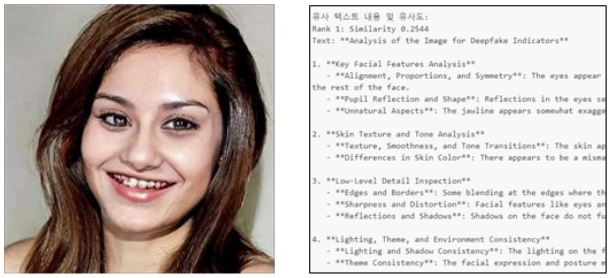

그림 2. 딥페이크 이미지 및 CLIP 로 매칭된 딥페이크 단서

Figure 2. Deepfake image and matched deepfake clues by CLIP

결론적으로, RAG 기반 파이프라인에 CLIP 을 활용한 텍스트-이미지 유사도 계산을 결합함으로써 LMM 기반 딥페이크 탐지 모델은 외부 지식을 보다 체계적으로 활용할 수 있게 된다. 이는 모델의 신뢰성과 실용성을 높이는 동시에, 실제 서비스 환경에서 발생할 수 있는 다양한 위조 사례를 놓치지 않는 안정적 딥페이크 탐지 시스템 구축에도 기여한다.

Ⅳ. 실험 결과

1. 성능 평가 결과

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 4 는 기존 LMM 모델의 딥페이크 탐지 성능을 정리한 결과이다. 파라미터 수가 많은 모델일수록 전반적으로 높은 정확도와 재현율을 보이는 경향이 확인되었으나, 범용적 사용을 목적으로 설계된 일반 LMM 모델은 파인튜닝된 모델에 비해 상대적으로 낮은 성능을 나타내는 특징이 관찰된다. 또한, Gemini-1.5-flash 모델의 경우 zero-shot 설정에서 Accuracy 55.6, F1 20.0, Recall 11.0 으로 활용이 어려울 정도로 낮은 성능을 보였으며, 내부 실험 결과 RAG 를 적용하더라도 유의미한 성능 개선이 관찰되지 않았다. 이에 따라 이후 RAG 적용 비교에서는 Gemini-1.5-flash 모델의 zero-shot 결과를 제외하였다. 대신 Gemini-1.5-flash 모델의 탐지 성능 변화 양상을 관찰하기 위해 추가로 파인튜닝을 수행하였고, 이를 통해 정확도와 재현율이 크게 향상됨을 확인하였다.

표 4. 기존 LMM 모델 딥페이크 탐지 성능
Table 4. Original LMM model deepfake detection performance

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 5 는 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 5. RAG 적용 후 딥페이크 탐지 성능
Table 5. Deepfake detection performance after RAG

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

2. 실험 결과 분석

RAG 기법을 적용한 결과, 일부 지표 간 상호 보완적인 변화가 관찰되었다. 예를 들어, Gemini-1.5-flash 모델은 파인튜닝 후 RAG 를 적용했을 때 Recall 이 6.0% 증가하여 탐지 성능이 크게 향상되었지만 F1 Score 가 0.8% 하락하고 Accuracy 도 2.5% 낮아졌다. 이는 Recall 향상을 위해 정밀도(Precision)를 일부 희생해야 하는 전형적인 트레이드오프 현상을 보여준다. 한편, Llama-3.2-11B-Vision-Instruct-Turbo 모델은 Recall 이 3.0% 상승했지만 F1 Score 와 Accuracy 가 각각 0.7%, 3.0% 하락하여, 상대적으로 파라미터 수가 적은 모델에서는 RAG 적용 효과가 제한적일 수 있음을 보여주었다.

가장 주목할 만한 개선은 Llama-3.2-90B-Vision-Instruct-Turbo 모델에서 관찰되었다. 이 모델은 Recall 을 6.4% 높여 탐지 민감도를 크게 끌어올리는 동시에, F1 Score 도 0.8% 증가시키는 결과를 보였다. 다만, Accuracy 는 1.4% 감소하였는데, 이는 탐지율 증대를 위한 정밀도 손실이 어느 정도 불가피함을 의미한다.

결과적으로, 본 연구에서 제시한 RAG 파이프라인은 LMM 딥페이크 탐지 모델에 외부 지식을 결합한 검증 단계를 통해 재현율을 중심으로 딥페이크 탐지 성능을 효과적으로 향상시킬 수 있음을 보여주었다. 특히 대규모 파라미터 모델에서 이러한 개선 폭이 더 두드러지며, zero-shot 환경에서도 일정 수준 이상의 성능 개선을 기대할 수 있다는 점에서 본 접근법의 실용적 가치를 확인할 수 있었다.

Ⅴ. 결론

본 연구는 이미지를 입력받는 LMM 환경의 대표적 사례인 딥페이크 탐지 분야의 RAG 연구 필요성에 주목하여, CLIP 모델의 유사도 계산을 결합한 RAG 접근법을 제안하였다. 특히 검색된 단서를 체크리스트 형태로 활용하는 파이프라인을 구축하여 재현율을 약 3~6.4% 개선하였다. 이는 추가적인 위조 사례를 놓치지 않는다는 점에서, 실제 서비스 환경 단계에서 위조 콘텐츠 검출의 안정성을 높일 수 있다. 동시에, 학술 연구 측면에서도 RAG 와 CLIP 을 활용한 모듈형 탐지 파이프라인 사례로 인용 가치가 높다.

향후에는 외부 지식 베이스의 품질 및 최신성을 유지하고, 음성 · 영상 · 센서 데이터 확장 시 발생할 수 있는 처리량 문제를 해결하기 위한 후속 연구가 필요할 것이다. 이러한 시도는 의료 영상 분석이나 자율주행 환경, 객체 탐지 등, 멀티모달 분석 전반에 걸쳐 유용하게 적용될 수 있을 것으로 기대된다.

참 고 문 헌

[1] S. Jia, R. Lyu, K. Zhao, Y. Chen, Z. Yan, Y. Ju, C. Hu, X. Li, B. Wu, 그리고 S. Lyu, "Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics," Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4324- 4333, 2024.

[2] Y. Li, X. Liu, X. Wang, B. S. Lee, S. Wang, A. Rocha, 그리고 W. Lin, "FakeBench: Probing Explainable Fake Image Detection via Large Multimodal Models," unpublished, 2024.

[3] 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.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, 그리고 I. Sutskever, "Learning transferable visual models from natural language supervision," Proc. Int. Conf. Machine Learning (ICML), vol. 139, pp. 8748- 8763, Jul. 2021.

[5] 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.

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

[7] 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.

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

[9] 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.