**[학회 발표 대본]**

**1. 인사 및 연구 소개**

안녕하십니까. 저는 CLIP의 유사도 기반 멀티모달 RAG 검증 기법에 대해 발표할 엔투솔루션 김은오 연구원입니다.

본 연구는 기존 RAG 기법이 텍스트 중심으로 이루어지던 한계를 극복하고, 이미지 입력을 활용하는 LMM 환경에서도 적용할 수 있도록 확장하였습니다. 이를 통해 딥페이크 탐지 성능을 개선하는 새로운 전략을 제시합니다.

발표 목차는 다음과 같습니다.

**2. 연구 배경 및 필요성**

최근 인공지능 기술의 발전으로 인해 딥페이크 콘텐츠가 더욱 정교해지고 있으며, 이에 따른 사회적, 보안적 위협이 심각해지고 있습니다. 이미 피싱, 성범죄 등으로 악용되는 사례가 발생하고 있으며, 앞으로 신뢰성이 중요한 뉴스, 보안 인증, 법적 증거물 등의 분야에도 큰 위협으로 작용할 것입니다.

기존 딥페이크 탐지 기법은 머신러닝 및 딥러닝 모델을 중심으로 발전해왔지만, 전문적인 지식과 복잡한 환경 설정이 필요하여 일반 사용자가 쉽게 접근하기 어렵다는 단점이 존재합니다.

한편 LMM은 손쉬운 접근성과 자연어 형식의 답변을 주요 강점으로 새로운 딥페이크 탐지 방법으로 주목받고 있습니다. 이를 통해 사용자는 단순한 이진 분류 결과가 아닌, 보다 직관적인 판단 근거를 제공받을 수 있습니다.

하지만 LMM 기반 탐지는 대형 언어 모델로 파생된 만큼 여전히 정보 환각(Hallucination) 문제라는 한계가 존재합니다. LLM은 이러한 환각 문제를 RAG로 극복했으나, 멀티모달에서의 RAG연구는 아직 초기단계이므로 본 연구는 CLIP 기반의 멀티모달 RAG 기법을 고안하였습니다.

**3. 관련 연구**

다음은 관련 연구입니다.

RAG는 대규모 언어 모델이 외부 지식 소스를 활용해 응답의 정확성과 신뢰성을 높이는 기술입니다. 먼저, 사용자의 질의와 관련된 정보를 외부 데이터베이스에서 검색한 뒤, 이를 모델의 응답 생성 과정에 반영합니다. 따라서, 검색 단계에서 얼마나 적절한 정보를 가져오느냐가 RAG의 성능을 결정짓는 핵심 요소입니다.

멀티모달 환경에서도 이러한 RAG의 검색 단계를 수행하기 위해 CLIP 모델이 사용됩니다. CLIP은 이미지와 텍스트를 함께 학습하는 대조 학습(Contrastive Learning) 방식을 사용하여, 이미지와 텍스트 간의 의미적 유사도를 높은 정확도로 측정할 수 있습니다. 이를 통해, 기존 텍스트 기반 RAG 모델과 달리 이미지 정보를 보다 정밀하게 검색하고 활용할 수 있습니다.

**4. 실험 설계**

본 연구의 실험 설계를 설명드리겠습니다.

**먼저 데이터셋** 입니다. 본 연구에서는 DeepFakeFace와 Seq-DeepFake 두 가지 데이터셋을 사용하였습니다. DeepFakeFace는 약 90,000개의 딥페이크 이미지와 30,000개의 실제 이미지를 포함하고 있으며, 3 가지의 다양한 합성 기법을 반영해 폭넓은 딥페이크 사례를 포괄한다는 특징이 있습니다. 본 연구에서는 해당 데이터 셋을 GPT-4o 모델을 활용하여 레이블링 작업을 수행하였으며, 90%의 데이터는 RAG DB에 저장, 나머지 10%를 테스트 데이터셋으로 활용하였습니다.   
또한 새로운 생성 기법에 대한 탐지 능력도 함께 살펴보기 위해 Seq-DeepFake 데이터셋을 테스트용으로 함께 활용하였습니다. 해당 데이터셋은 연속적인 얼굴 조작을 포함하여 보다 복합적인 위조 양상을 반영하고 있다는 특징이 있습니다.

**다음은 실험 모델 입니다.** 네 가지 모델을 사용하여 RAG 적용 전후의 탐지 성능을 비교하였습니다. 실험에는 Google의 Gemini-1.5-flash 모델과 이를 DeepFakeFace 데이터셋으로 파인튜닝한 버전, Meta의 Llama-3.2 계열 모델 등이 포함되었습니다.

**3) RAG 파이프라인**

본 연구에서 제안하는 RAG 파이프라인의 주요 과정은 다음과 같습니다:

1. 모델이 1차적으로 딥페이크 여부를 탐지합니다.
2. 실제 사진으로 분류된 이미지를 CLIP 기반 유사도 검색을 통해 딥페이크 단서 키워드와 결합합니다.
3. 이를 체크리스트 형태로 활용하여 추가 탐지를 수행한 후 최종 결과를 도출합니다..

이러한 파이프라인을 통해, 추가적인 딥페이크 사례를 포착할 수 있게됩니다.

**5. 개선 조치**

이 사진은 실제 딥페이크 합성 이미지를 CLIP 모델을 활용하여 유사한 단서리스트와 매칭한 결과입니다. 시선 및 눈 정렬의 불일치, 입술 및 치아의 부자연스러움 등 관련성 높은 단서들이 성공적으로 매칭된 것을 확인할 수 있었습니다.

다만 이 과정에서 딥페이크 관련 키워드가 프롬프트에 많이 포함될수록 탐지 모델이 편향된 결과를 내는 경향이 관찰되었습니다. 표 3번을 보시면 판단 근거를 요구하거나 탐지 방법에 대한 자세한 설명을 포함한 경우, 기본 프롬프트 대비 두 배 이상 높은 확률로 딥페이크로 탐지하는 현상을 확인할 수 있습니다

따라서, RAG로 검색된 Context가 모델의 탐지 결과를 과도하게 왜곡하는 것을 방지하기 위해, 검색된 단서를 체크리스트로 활용하고 탐지 신뢰도를 보정하는 방식을 적용하였습니다. 이를 통해, 탐지 모델이 검색된 단서를 보조 정보로 활용하면서도, 과도한 탐지 편향이 발생하지 않도록 조정하였습니다.

**6. 실험 결과 및 분석**

**1) 결과 요약**

실험 결과는 다음과 같습니다. RAG 기법을 적용한 모델은 기존 대비 3~6.4% 높은 재현율을 보였으며, 이는 누락되었던 딥페이크를 추가로 포착할 수 있음을 시사합니다. 다만 정확도(Accuracy)가 약 1.4~3% 까지 감소하는 경향을 보였으며, 이는 탐지율을 높이는 과정에서 일부 정밀도가 희생되는 트레이드오프 현상으로 분석됩니다.

해당 결과를 통해 튜닝된 모델뿐만 아니라, Zero-shot 환경에서도 RAG 적용 시 유의미한 재현율 향상이 나타났습니다. 특히, 대규모 모델에서 성능 향상 폭이 더욱 두드러져, 향후 등장할 고성능 LMM 모델에서도 해당 RAG 기법이 높은 활용 가능성을 가질 것으로 예상됩니다

**7. 결론 및 향후 연구 방향**

본 연구에서는 RAG 기법을 LMM 기반 딥페이크 탐지에 적용하여 탐지 성능을 효과적으로 개선하는 방법론을 제시하였습니다. 특히, CLIP 기반 유사도 계산을 활용하여 외부 지식을 보다 체계적으로 결합하는 새로운 파이프라인을 구축하였으며, 이를 통해 재현율을 중심으로 탐지 성능을 강화하였습니다.

이는 특히 재현율이 중요한 분야에서 높은 활용 가능성을 보일 것으로 기대됩니다.

향후 연구에서는 데이터셋의 다양성을 확대하고, 다양한 모델 구조를 실험하여 일반화 성능을 보다 정밀하게 검증할 필요가 있습니다. 또한, 음성 및 영상 데이터로의 확장, 실시간 처리 속도 개선 등이 향후 주요 과제가 될 것입니다..

이상으로 발표를 마치겠습니다. 감사합니다. 질문 있으시면 부탁드립니다.