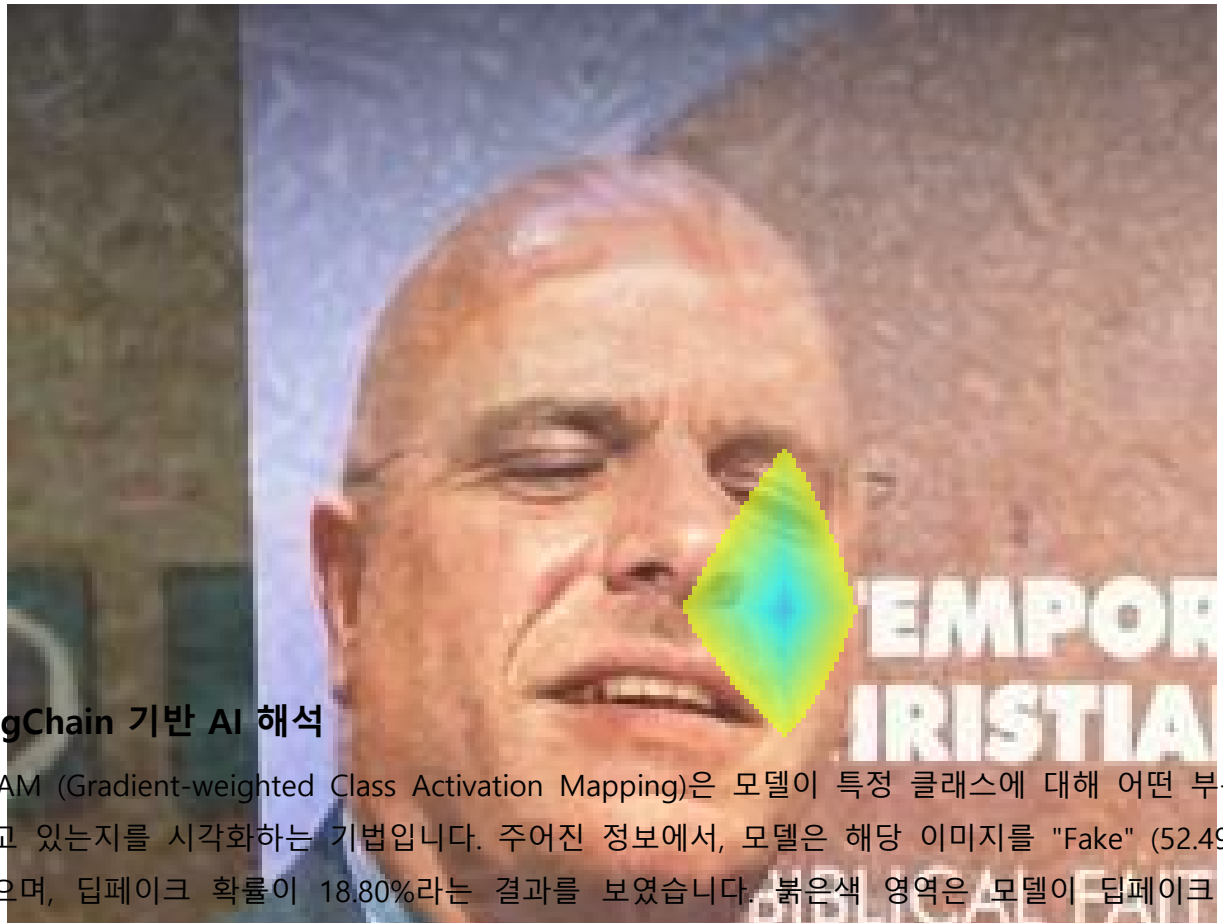
딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 외국인 전용 모델
- 분석 일시: 2025-11-07 09:57:13
- 예측 결과: Fake (52.49%)
- 딥페이크 확률: 18.80%

2. Grad-CAM 시각화



3. LangChain 기반 AI 해석

Grad-CAM (Gradient-weighted Class Activation Mapping)은 모델이 특정 클래스에 대해 어떤 부분에 더 집중하고 있는지를 시각화하는 기법입니다. 주어진 정보에서, 모델은 해당 이미지를 "Fake" (52.49%)으로 분류했으며, 딥페이크 확률이 18.80%라는 결과를 보였습니다. 붉은색 영역은 모델이 딥페이크 판단의 근거로 삼은 부분입니다. 이를 통해 드러나는 기술적 분석은 다음과 같습니다.

시각적 근거 분석

1. **합성 흔적**:
   - 붉은색 영역에서 모델이 중요하게 판단한 섹션은 주로 얼굴의 경계선, 턱선, 그리고 눈 주위로 집중되고 있을 가능성이 높습니다. 딥페이크에서 자주 발생하는 합성 흔적은 경계가 부자연스럽거나 매끄럽지 못한

경우가 많습니다. 이러한 영역에서 모델은 인공적으로 처리된 흔적을 인식했을 수 있습니다.

2. **피부 질감**:
   - 딥페이크 기술에서는 종종 피부 질감이 자연스러운 것과 다르게 나타나는 문제가 발생합니다. 붉은색 영역이 피부와 직접적인 연관이 있다면, 모델은 고해상도 이미지에서 티가 나는 텍스처 차이를 인식했을 가능성이 있습니다. 즉, 피부의 세밀한 주름, 결점, 또는 고르지 않은 질감이 AI의 생성과 다르게 나타날 수 있습니다.

3. **조명 왜곡**:
   - 인공적으로 합성된 이미지에서 조명은 종종 비일관적이거나 비자연스러울 수 있습니다. 만약 붉은색 영역이 조명의 방향이나 색상에 관한 부분이라면, 이러한 불일치가 모델이 딥페이크로 판단한 근거가 되었을 것입니다. 이는 일반적으로 인물의 일부가 잘 조명되지 않거나 그림자가 잘못 촘촘히 형성될 때 나타날 수 있습니다.

신뢰도와 한계점

1. **신뢰도**:
   - Grad-CAM을 사용한 시각적 해석은 모델이 왜 이러한 결과를 내렸는지를 인간 전문가가 이해하는 데 도움을 줍니다. 특히 특정 이미지에서 모델이 주목한 상세 영역을 시각적으로 확인할 수 있기 때문이다. 이는 딥페이크 탐지에 대한 설명 가능성을 부여하며, 모델의 신뢰성을 어느 정도 높이는 요소이기도 합니다. 모델이 특정한 의미 있는 구조를 인식하고 있을 수 있기 때문에 이것이 딥페이크를 판단하는 데 도움이 될 수 있습니다.

2. **한계점**:
   - Grad-CAM은 모델의 예측에 대해 직관적인 이해를 제공하지만, 특정 영역이 항상 신뢰할 수 있는지는 의문이 남습니다. 예를 들어, 모델이 특정 영역에서 잘못된 해석을 하거나, 훈련 데이터에서의 편향성에 의해 오염된 정보를 기반으로 판단할 수 있습니다. 또한, Grad-CAM은 인식된 특징에 대한 상대적 중요도를 보여줄 뿐이며, 이러한 특징들이 실제로 딥페이크와 어떻게 연관되는지를 보장하지 않습니다.

심층 결과

- 만약 해당 모델이 학습할 때 다양한 딥페이크 데이터셋에 과적합되었다면, 일반적인 얼굴 특징이나 의도하지 않은 세부사항에 대해 민감한 반응을 할 수 있습니다. 예를 들어, 얼굴의 특정 특징이나 구조적 비율이 비정상적인 경우에도 "Fake" 판단을 내릴 수 있습니다. 이러한 경우에는 진짜와 가짜 딥페이크의 차별성을 더욱 심층적으로 분석해야 할 필요가 있습니다.

이러한 분석을 통해, Grad-CAM의 시각적 결과와 함께 실제 이미지에서의 딥페이크 탐지 가능성을 더욱 발전시킬 수 있는 기초 자료를 마련할 수 있습니다.

4. 결론 및 권장 조치

본 분석은 Grad-CAM 시각 주목도를 중심으로 진행되었습니다.
AI의 결과는 참고용으로 사용해야 하며, 법적 판단이나 공식 증거로 사용되지 않습니다.
결과의 신뢰도를 높이기 위해 다양한 이미지 소스로 교차 검증을 권장합니다.