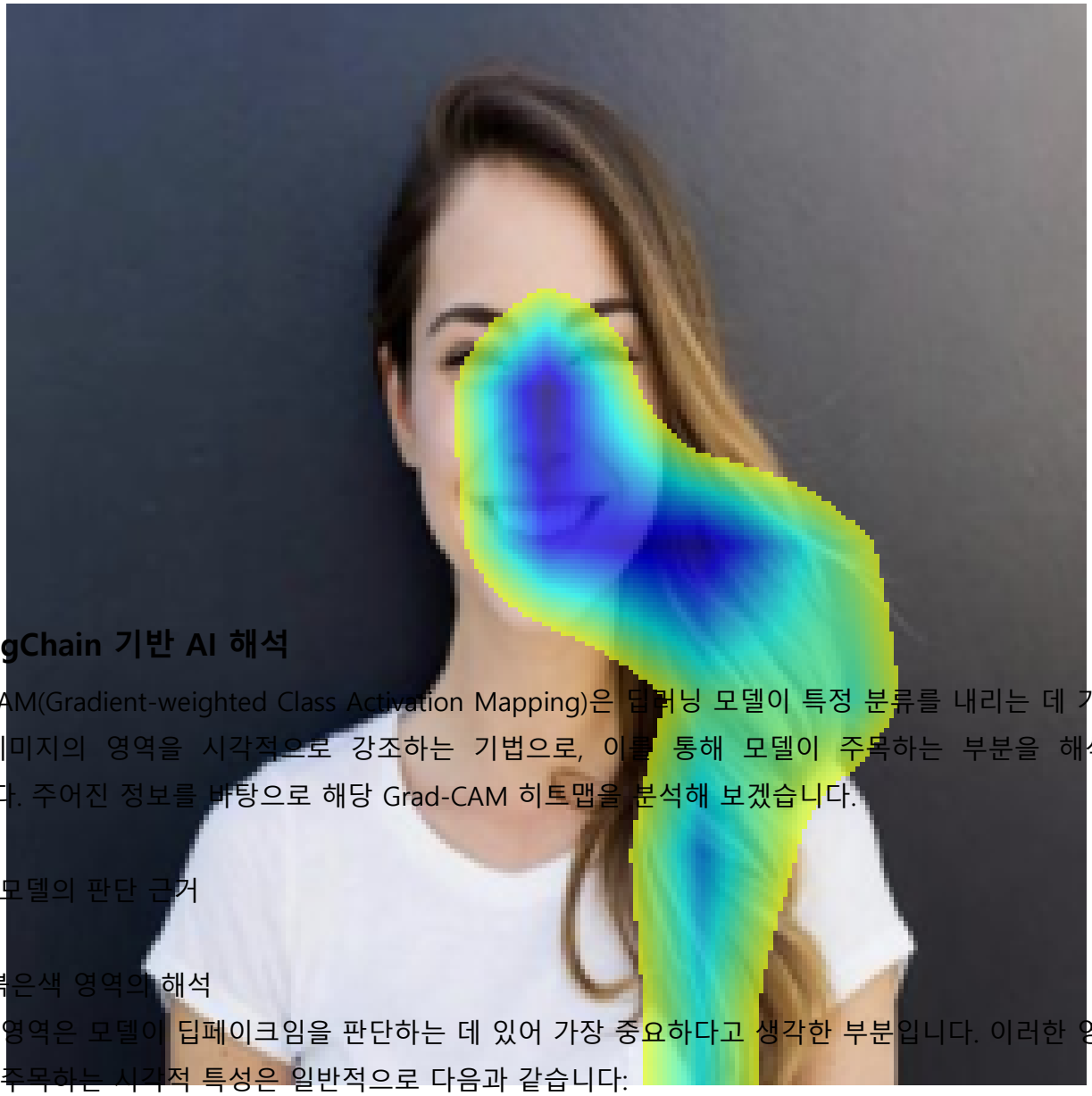
딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 외국인 전용 모델
- 분석 일시: 2025-11-06 15:21:21
- 예측 결과: Real (7.69%)
- 딥페이크 확률: 14.70%

2. Grad-CAM 시각화



3. LangChain 기반 AI 해석

Grad-CAM(Gradient-weighted Class Activation Mapping)은 딥러닝 모델이 특정 분류를 내리는 데 기여하는 입력 이미지의 영역을 시각적으로 강조하는 기법으로, 이를 통해 모델이 주목하는 부분을 해석할 수 있습니다. 주어진 정보를 바탕으로 해당 Grad-CAM 히트맵을 분석해 보겠습니다.

1. 모델의 판단 근거

붉은색 영역의 해석
붉은색 영역은 모델이 딥페이크임을 판단하는 데 있어 가장 중요하다고 생각한 부분입니다. 이러한 영역에서 모델이 주목하는 시각적 특성은 일반적으로 다음과 같습니다:

- **합성 흔적**: 딥페이크 기술은 보통 얼굴의 경계 부분에서 합성이 이루어지기 때문에, 특히 턱선, 이마, 볼 부분 등에서 비정상적인 경계가 나타날 수 있습니다. 이러한 합성과 관련된 불일치는 붉은색 영역으로 강조되었을 가능성이 높습니다.

- **피부 질감**: 자연 이미지는 사람의 피부 질감에서 다양한 변화를 보입니다. 반면에 딥페이크는 종종 너무 매끈하거나 비현실적인 질감을 가질 수 있으며, 그로 인해 모델이 특정 위치에서 비정상적인 질감을 감지했을 가능성이 있습니다.

- **조명 왜곡**: 딥페이크는 배경과의 조화가 부족할 수 있어 조명이나 그림자의 불일치가 발생할 수 있습니다. 붉은색 영역이 이러한 광원의 이상 여부를 나타내며, 이미지 내 조명과 그림자가 자연스럽지 않은 부분에 대해 모델이 조명을 근거로 판단했을 수 있습니다.

2. 신뢰도와 한계점

신뢰도
이러한 Grad-CAM 해석을 통해 모델이 특정 시각적 요소를 근거로 판단했다는 점에서 과학적 합리성과 신뢰성을 부여할 수 있습니다. 그러나 다음과 같은 고려 사항이 필요합니다:

- **데이터 편향**: 모델이 학습한 데이터셋의 품질과 다양성에 따라 다르게 작용할 수 있습니다. 특히, 모델이 딥페이크보다 실제 이미지를 상정하여 학습했을 경우, 특정 유형의 딥페이크에 대해 약한 성능을 보일 수 있습니다.

- **시각적 정성**: Grad-CAM은 시각화를 통해 해석을 제공하지만, 해석이 반드시 올바른 것은 아닙니다. 특정 영역이 강조된 이유는 그 영역에서의 시각적 특성이 아니라 모델의 특정 학습 패턴에 기인할 수 있습니다.

한계점
- **복잡한 이미지 처리**: 전문가들은 종종 여러 요소를 고려하여 판단합니다. 그에 반해 모델은 고정된 규칙에 따라 작동하기 때문에, 매우 복잡한 이미지 내 여러 요소들이 상호작용하는 경우에는 한계가 있습니다.

- **일관성 부족**: 모델이 같은 종류의 이미지를 다르게 판단할 수 있으며, 그로 인해 신뢰성에 균열이 생길 수 있습니다. 다양한 조명 조건, 배경, 요소들에 따라 이미지가 달리 해석될 수 있습니다.

3. 심층 결과
모델이 Red 영역 예측 결과 외에도 추가적으로 다른 심층 정보를 바탕으로 한 특성이 있을 수 있습니다. 특정 특징이나 패턴들, 예를 들어 눈 주위의 선명도, 얼굴 표정의 일관성, 머리카락의 움직임 등이나 모델 구조의 병목 구간에서 나타나는 가중치 분포도 딥페이크 탐지의 중요한 힌트를 제공할 수 있습니다. 이러한

특성들은 전문가에게 서로 다른 판단 근거를 제공하여, 전체 이미지에 대한 평가를 더욱 복합적으로 만들어 줄 수 있습니다.

결론적으로, Grad-CAM 히트맵의 붉은색 영역은 모델이 딥페이크를 감지하는 데 주요한 기준이 되는 시각적 요소를 나타내지만, 전문가의 판단과 함께 다각적인 검토가 필요합니다.

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.