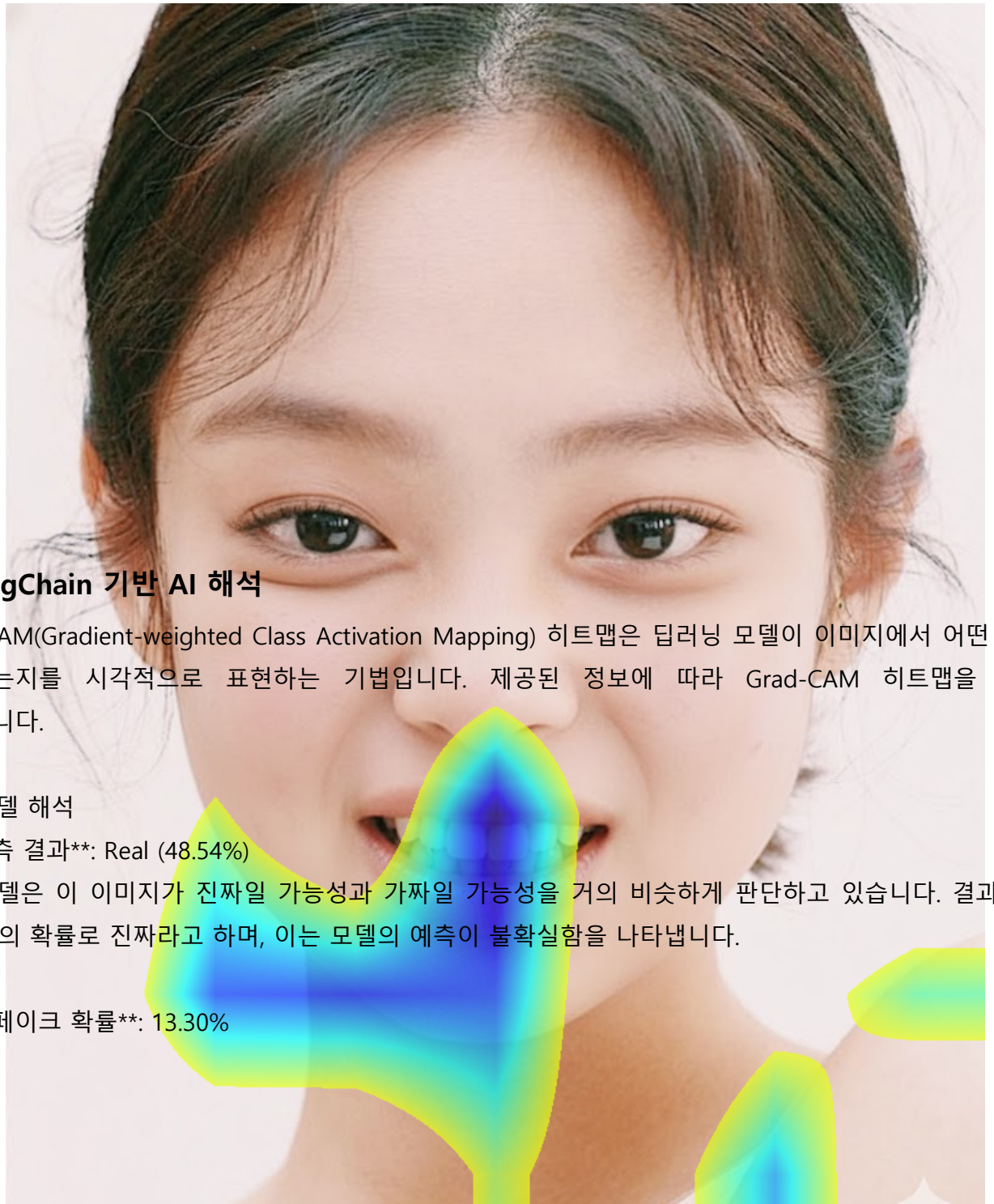
딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-06 16:43:14
- 예측 결과: Real (48.54%)
- 딥페이크 확률: 13.30%

2. Grad-CAM 시각화



3. LangChain 기반 AI 해석

Grad-CAM(Gradient-weighted Class Activation Mapping) 히트맵은 딥러닝 모델이 이미지에서 어떤 부분에 주목했는지를 시각적으로 표현하는 기법입니다. 제공된 정보에 따라 Grad-CAM 히트맵을 해석해 보겠습니다.

모델 해석

1. **예측 결과**: Real (48.54%)
   - 모델은 이 이미지가 진짜일 가능성과 가짜일 가능성을 거의 비슷하게 판단하고 있습니다. 결과적으로 48.54%의 확률로 진짜라고 하며, 이는 모델의 예측이 불확실함을 나타냅니다.

2. **딥페이크 확률**: 13.30%

- 딥페이크로 판단할 확률이 낮은 것으로 보아, 모델은 이미지의 진정성을 유지한다고 인식하지만, 여전히 일부 영역에서는 딥페이크의 가능성을 엿보고 있을 수 있습니다.

Grad-CAM 히트맵 해석
- **붉은색 영역**: 모델이 딥페이크 판단의 근거로 본 부분입니다. 이러한 영역에서는 다음과 같은 시각적 특징이 존재할 수 있습니다:

1. **합성 흔적**:
   - 붉은색 영역에서 경계가 분명하지 않거나 불규칙한 형태가 발견된다면, 이는 합성이 이루어진 부위임을 나타낼 수 있습니다. 예를 들어, 얼굴 주위의 세부 특징(머리카락이나 배경) 경계가 부드럽지 않고 인위적으로 보일 수 있습니다.

2. **피부 질감**:
   - 피부 질감의 불균형이나 비정상적인 광택이 붉은 영역에서 나타나는 경우, 이는 합성으로 인해 피부가 비정상적으로 처리되었음을 시사할 수 있습니다. 진짜 피부는 자연스럽고 다양한 질감이 있을 수 있지만, 딥페이크 이미지에서는 이러한 질감이 왜곡될 수 있습니다.

3. **조명 왜곡**:
   - 이미지의 조명이 불균형하게 분포되어 있다면, 이는 합성의 신호일 수 있습니다. 적절한 조명이 각기 다른 부위에 고르게 퍼져야 하지만, 붉은색 히트맵이 강조된 부분에서 조명의 방향이나 강도가 비대칭적이라면 딥페이크의 징후로 해석될 수 있습니다.

신뢰도 및 한계점
- **신뢰도**:
  - Grad-CAM은 모델이 주목하는 영역을 시각화하는 유용한 도구로, 특정 지역에서의 딥페이크 판단 근거를 시각적으로 잘 드러냅니다. 전문가가 이러한 히트맵을 바탕으로 추가적인 검토를 할 수 있는 기회를 제공합니다.

- **한계점**:
  - Grad-CAM은 모델이 주목한 시각적 신호를 해석하는데 도움이 되지만, 이 히트맵이 반드시 명확한 딥페이크의 징후라고 해석되지는 않습니다. 모델의 불확실성(48.54%의 Real 확률)도 보여주듯이, 단일 히트맵만으로 결론을 내리기엔 부족할 수 있습니다.
  - 또한, 고급 합성 기술이 발전함에 따라 진짜와 구별하기 어려운 패턴이 생기기도 하여, 히트맵에서 탐지된 징후가 반드시 딥페이크와 관련 있지 않을 가능성도 있습니다.

심층 결과
- **추가적인 분석**:
  - 심층 결과로는 모델이 추출한 다양한 특성들 (예: 색상, 윤곽, 형태 등)의 상관 관계를 이해하는 것도

중요합니다. 모델이 어떤 종류의 패턴을 잘 학습하는지 분석하여, 추가적인 해법이나 보완적인 방법론 개발에 도움을 줄 수 있습니다.

- 또, 다양한 얼굴 표현(예: 미소, 찡그리기 등)이나 환경적 변수를 고려하여 더 많은 데이터를 활용한 피드백을 제공하는것도 심층적 리포트에서 추가적으로 다뤄야 할 요소입니다.

결론

Grad-CAM 히트맵의 시각적 정보를 통해 모델이 딥페이크로 판단한 근거를 이해할 수 있지만, 이는 전문가의 해석과 경험을 바탕으로 더욱 정교화되어야 합니다. 다양한 변수를 고려한 종합적인 접근을 통해 보다 정확한 판별이 가능할 것입니다.

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.