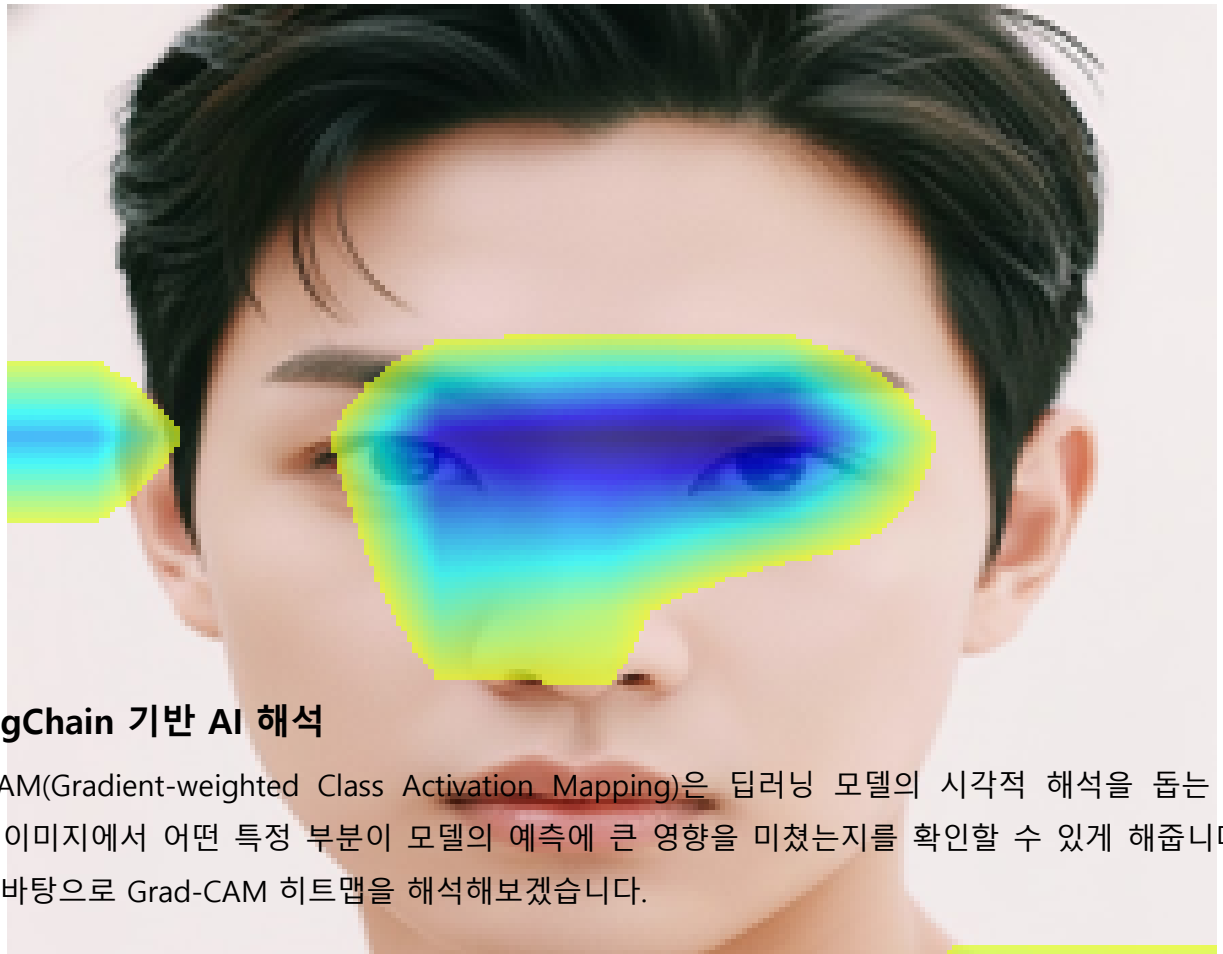
딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-06 14:54:51
- 예측 결과: Fake (52.61%)
- 딥페이크 확률: 16.20%

2. Grad-CAM 시각화



3. LangChain 기반 AI 해석

Grad-CAM(Gradient-weighted Class Activation Mapping)은 딥러닝 모델의 시각적 해석을 돕는 기술로, 주어진 이미지에서 어떤 특정 부분이 모델의 예측에 큰 영향을 미쳤는지를 확인할 수 있게 해줍니다. 위의 정보를 바탕으로 Grad-CAM 히트맵을 해석해보겠습니다.

1. Grad-CAM 해석
- **붉은색 영역**: 붉은색 영역은 모델이 Fake(위조)로 예측하는 데 가장 큰 영향을 미친 부분입니다. 이 구역에서 관찰되는 특징들은 다음과 같습니다.
  - **합성 흔적**: 붉은색 영역에 해당하는 부위에서는 경계가 뚜렷한 영역이 있을 가능성이 높습니다. 이는 원본 이미지와 합성된 부분의 자연스러운 전이 없이 뚜렷하게 구분되어 보일 수 있습니다. 합성의 흔적으로는 턱선, 헤어라인, 귀 등의 경계가 자연스럽지 못하거나 인위적으로 수정된 요소가 관찰될 수

있습니다.
  - **피부 질감**: 딥페이크 기술로 생성된 이미지의 피부 질감은 종종 부자연스럽고, 너무 매끄럽거나 단조로운 경향을 보일 수 있습니다. 따라서 붉은색 표시가 피부 질감의 비정상적인 부분을 강조하기도 합니다.
  - **조명 왜곡**: 조명이 인공적으로 조정된 경우, 고유한 조명 패턴과 그림자가 나타나지 않거나, 씬의 다른 부분과 비해 조명 강도가 불균형한 경우가 많습니다. 이로 인해 붉은 영역이 조명의 이상함을 나타낼 수 있습니다. 조명에 관련된 흐림이나 강렬한 특징이 이 부분에서 강조될 수 있습니다.

2. 신뢰도 및 한계점
- **신뢰도**: 모델이 Fake(52.61%)로 분류한 결과와 관련하여, Grad-CAM 히트맵의 붉은 영역이 모델이 왜 이렇게 판단했는지에 대한 시각적 근거를 제공하고 있습니다. 특히, 붉은 영역의 특징이 잘 정의되어 있고, 신뢰할 수 있는 지표 (조명 왜곡, 피부 질감, 합성 흔적 등)가 확인된 경우 모델 판단에 대한 신뢰도를 높일 수 있습니다. 그러나 확률(16.20%)이 매우 낮아, 이 예측이 상당히 불확실하다는 점도 고려해야 합니다.

- **한계점**:
  - Grad-CAM은 고해상도의 이미지를 다룰 때 특정 단서의 위력을 강조할 수 없기 때문에 세밀한 분석이 부족할 수 있습니다. 이로 인해 합성의 아주 미세한 단서들이 간과될 수 있습니다.
  - 또한, 모델이 붉은색으로 강조한 부분이 꼭 Fake와 관련이 있는 부분인지, 아니면 다른 요인에 의해 발생한 자연스러운 이미지의 특성인지에 대한 해석의 여지가 남아 있습니다. 즉, Grad-CAM이 강조하는 부분이 진정한 원인의 지표가 아닐 수 있습니다.
  - 더불어, 인간 전문가가 판단할 때 모델의 결과가 100% 신뢰할 수 없기에 추가적인 검증 과정이나 다른 딥페이크 탐지 모델과의 결과를 비교하는 것이 중요합니다.

3. 추가적인 심층 결과
- **심리적 요소**: 딥페이크 이미지에서는 감정 표현에 대한 왜곡도 발생할 수 있습니다. 예를 들어, 표정이 비정상적으로 나타나거나 자연스러운 움직임이 결여된 경우가 많습니다. 이러한 요소는 감정 인식과 관련하여 추가적인 시각적 단서를 모델이 활용했을지에 대한 가능성도 존재합니다.
- **유사인물 모델**: 딥페이크 이미지가 특정 인물 모사의 경우, 해당 인물의 일반적인 이미지와 비교하여 중요한 차이점을 강조할 수 있습니다. 따라서 배열된 인물의 유사성과 그 촬영 타이밍이 특별한 경우에 더욱 중요한 시각적 차이가 발생할 수 있습니다.

결론적으로, Grad-CAM을 활용한 분석은 딥페이크 탐지에 있어 매우 유용한 도구가 될 수 있지만, 이러한 시각적 해석이 항상 신뢰할 수 있는 판단으로 이어질 수는 없으며, 다양한 요소와 그 한계를 신중하게 고려해야 합니다.

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.