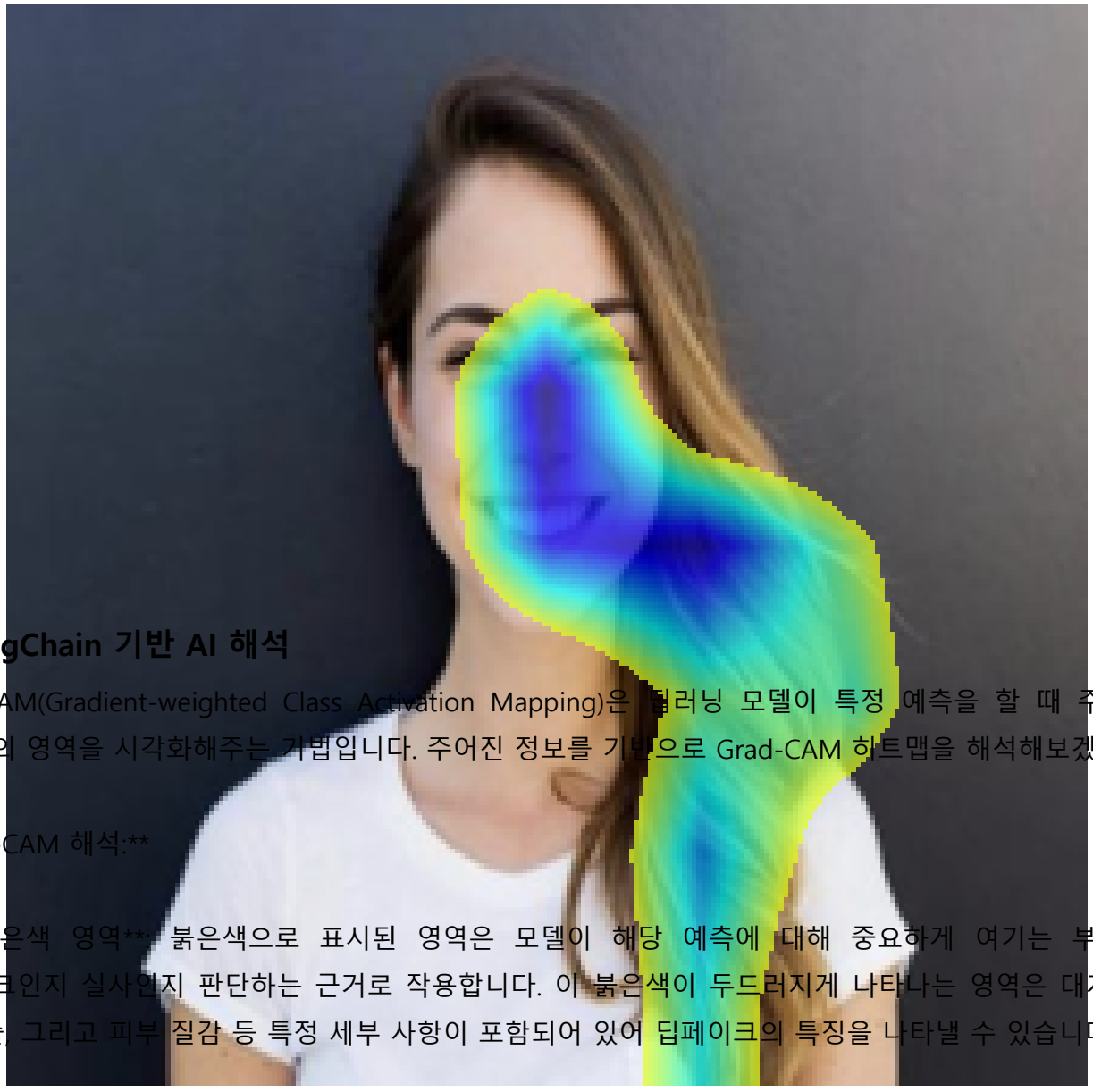
딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 외국인 전용 모델
- 분석 일시: 2025-11-06 15:21:41
- 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 예측을 할 때 주목하는 이미지의 영역을 시각화해주는 기법입니다. 주어진 정보를 기반으로 Grad-CAM 히트맵을 해석해보겠습니다.

Grad-CAM 해석:

1. **붉은색 영역**: 붉은색으로 표시된 영역은 모델이 해당 예측에 대해 중요하게 여기는 부분으로, 딥페이크인지 실사인지 판단하는 근거로 작용합니다. 이 붉은색이 두드러지게 나타나는 영역은 대개 얼굴, 눈, 입술, 그리고 피부 질감 등 특정 세부 사항이 포함되어 있어 딥페이크의 특징을 나타낼 수 있습니다.

2. **합성 흔적**: 딥페이크에서는 종종 경계가 뚜렷하게 나타나는 경우가 많습니다. 붉은색 영역이 경계가

흐리거나 날카로운 경우, 이는 인위적인 합성 작업에서 발생할 수 있는 흔적을 나타낼 수 있습니다. 여기서 모델이 주목하는 영역이 이러한 불일치를 반영했을 가능성이 있습니다.

3. **피부 질감**: 피부 질감의 일관성이 떨어지거나, 자연스러운 표현과 비교해 부자연스럽게 보이는 부분이 모델의 분석 초점이었을 수 있습니다. 딥페이크는 종종 피부가 지나치게 매끄럽거나 인위적으로 처리된 경우가 많기 때문에 이를 모델이 탐지했을 가능성이 있습니다.

4. **조명 왜곡**: 얼굴의 조명은 자연스러운 사진에서 일관된 경향성을 보입니다. 만약 특정 지역(예: 머리카락, 얼굴의 어느 쪽)이 비정상적으로 조명이 비치는 경우, 모양이 왜곡되어 보일 수 있습니다. 붉은 색 영역이 그런 비정상적인 조명 효과가 나타나는 곳이라면, 모델은 이를 딥페이크의 징후로 해석했을 것입니다.

신뢰도와 한계점:

- **신뢰도**: Grad-CAM을 통해 시각화된 정보는 모델의 주목 지점이기 때문에 그 자체로는 깊은 통찰을 제공합니다. 하지만 예측의 신뢰도는 모델이 훈련된 데이터의 품질과 다양성에 크게 의존합니다. 좋은 성능을 가진 모델이라 할지라도, 적절히 훈련되지 않은 경우 거짓으로 예측하는 경우가 발생할 수 있습니다.

- **한계점**: Grad-CAM은 모델이 왜 특정 결정을 내렸는지를 시각적으로 표현하지만, 이는 완전한 해석을 제공하지 못할 수 있습니다. 또한, 신뢰하지 못 할 사소한 두 가지 요소가 혼합된 경우(예: 현실 이미지가 엔지니어링 작업과 결합된 경우), 그 복잡성 때문에 모델의 판단을 왜곡할 수 있습니다.

심층 결과:

- **Feature Importance**: Grad-CAM만으로도 모델이 어떤 기능에 중점을 두고 있는지를 시각적으로 보여줄 수 있지만, 실제로 컬러 스케일이 다른 특정 픽셀 영역의 중요성을 나타내지 않기 때문에 단일 카테고리에 국한된 정보가 아닙니다.

- **모델의 전반적인 해석**: 이러한 추세는 이미지를 통해 감지하는 특성과 부족한 정보를 통해 나타나는 추세를 모두 포함하게 됩니다. 딥페이크와 상대적으로 유사한 전통적인 포트레이트의 모습이 포함된다면, 이를 정당화하기 위해 더욱 복잡한 추가 분석이 필요할 수 있습니다.

결론적으로, Grad-CAM을 통해 시각적으로 모델이 중요하게 여기는 요소를 이해할 수 있지만, 단일 매트릭스의 해석에 대해서는 주의가 필요합니다. 데이터의 품질, 다양한 딥페이크 종류, 그리고 인간 전문가의 시각에도 절대적인 기준이 없다는 점을 인식해야 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.