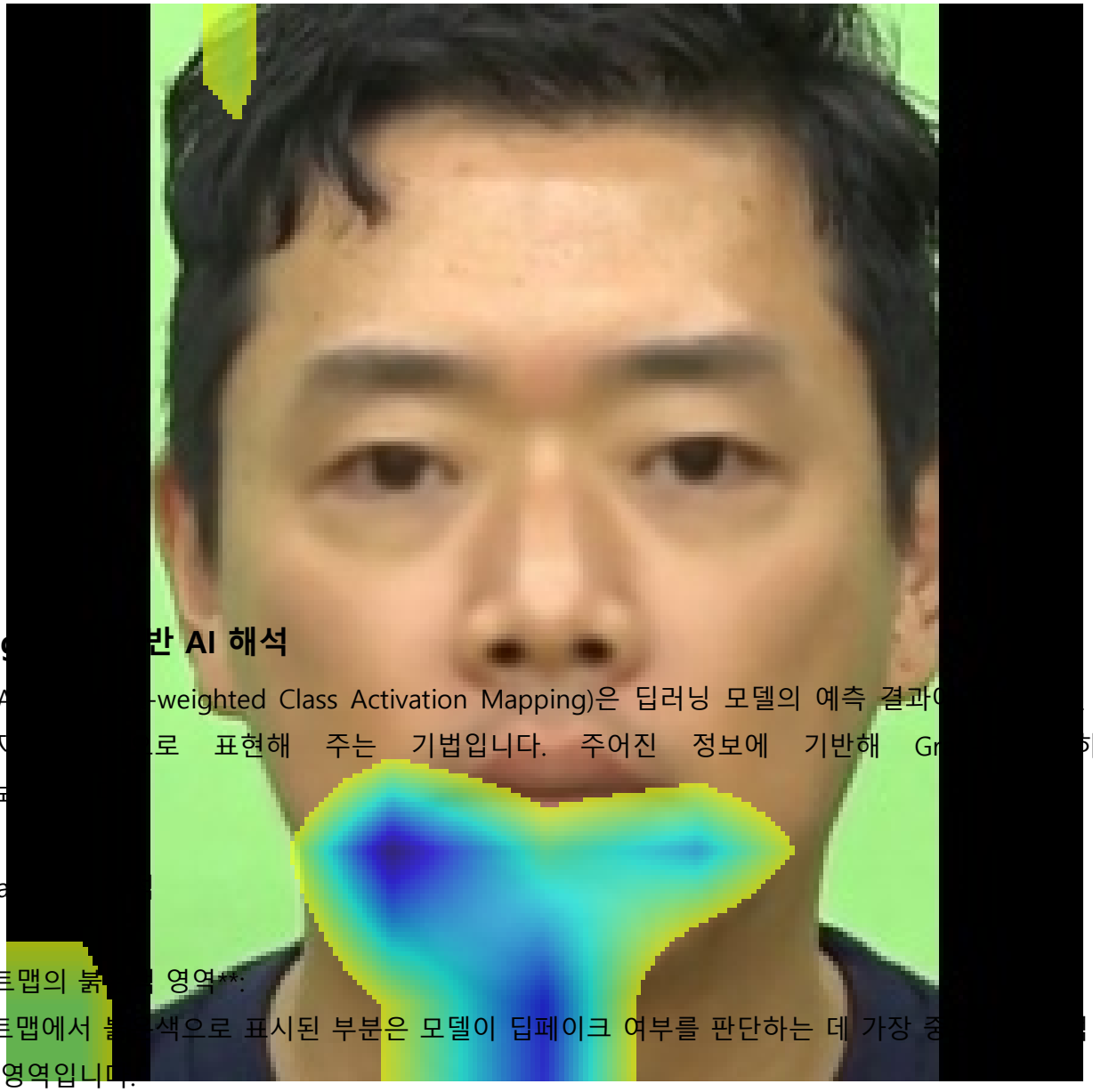
딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-06 17:31:46
- 예측 결과: Real (14.04%)
- 딥페이크 확률: 14.20%

2. Grad-CAM 시각화



3. Lang 반 AI 해석

Grad-CA -weighted Class Activation Mapping)은 딥러닝 모델의 예측 결과 부분이 중요한지 로 표현해 주는 기법입니다. 주어진 정보에 기반해 Gr 히트맵을 해석해보

Gra

1. **히트맵의 붉 영역**:
   - 히트맵에서 붉 색으로 표시된 부분은 모델이 딥페이크 여부를 판단하는 데 가장 중 요소로 인식한 영역입니다.
   - 일반적으로 딥페이크 이미지에서 이러한 붉은 영역은 얼굴의 특정 부위, 예를 들어 눈, 입술, 그리고 피부

질감이 고르게 나타나지 않은 부분에서 발견될 수 있습니다.

2. **합성 흔적**:
   - 딥페이크 기술은 종종 합성 처리에서 발생하는 경계 불명확성 또는 얼굴 윤곽선의 이상을 드러나게 할 수 있습니다. 붉은색 영역이 피부 경계에 가까우면 합성과정에서 생긴 경계 문제를 나타낼 수 있습니다.
   - 일반적으로 얼굴 주변의 경계가 부자연스럽거나 흐릿할 때, 또는 합성된 얼굴과 원본 얼굴 사이의 시각적 차이에 더욱 주목하게 됩니다.

3. **피부 질감**:
   - 모델이 붉은색 영역으로 강조한 부위가 피부 질감과 관련이 있다면, 일반적인 피부의 질감 차이를 보여줄 수 있습니다. 예를 들어, 고르지 못한 피부의 질감이나 비정상적인 주름, 또는 매끄러운 표면 등이 이러한 부위에서 확인될 수 있습니다.

4. **조명 왜곡**:
   - 조명이 딥페이크 이미지와 원본 이미지 간의 깊은 차이를 만들어낼 수 있습니다. 반사나 음영의 불일치가 붉은 영역에 나타나면, 조명 효과가 자연스럽지 못해 인식의 단서로 작용할 수 있습니다. 보통 합성된 이미지에서는 조명 일관성이 부족한 경우가 많기 때문입니다.

전문가의 관점에서의 신뢰도 및 한계점

- **신뢰도**:
  - Grad-CAM으로 나타난 시각적 정보는 모델이 어떤 패턴을 학습했는지를 이해하는 데 유용합니다. 예를 들어, 특정 부위가 딥페이크 판단의 주요 요소로 작용한 경우, 이는 그 부위에서 찾을 수 있는 비정상적인 특징이 있음을 시사합니다.
  - 여러 히트맵을 함께 분석하거나 동일한 모델의 연속적인 평가를 통해 더 높은 일관성을 갖는 판단이 가능할 수 있습니다.

- **한계점**:
  - Grad-CAM은 시각적 해석에 매우 의존적이기 때문에, 모호하거나 불확실한 이미지에서는 잘못된 판단을 초래할 수 있습니다. 시각적 근거만으로 신뢰성을 평가하기에는 부족할 수 있습니다.
  - 또한, 딥페이크 기술이 발전하면서 점점 더 정교해짐에 따라, 기존의 시각적 특성이 더 이상 유효하지 않을 수 있습니다. 즉, 새로운 종류의 딥페이크는 기존 모델의 판단 기준에서 벗어날 수 있습니다.

심층적인 결과

- **다양한 분석 기법 병합**:
  - Grad-CAM의 결과를 다른 탐지 기법과 병합(예: 외부 전처리, 다양한 딥러닝 모델의 조합)하면 성능 향상 및 신뢰도 있는 결과를 얻는 데 유리할 수 있습니다.

- **모델의 훈련 데이터**:
  - 사용한 훈련 데이터의 질과 다양성이 모델 성능에 크게 영향을 미칩니다. 특정 인종이나 환경에서 아주 잘 작동하더라도 다양한 조건에서 테스트할 때 신뢰성이 떨어질 수 있습니다.

이와 같은 분석을 바탕으로 모델의 판단 과정을 이해하고 더 나은 탐지 기법을 개발하는 데 기여할 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.