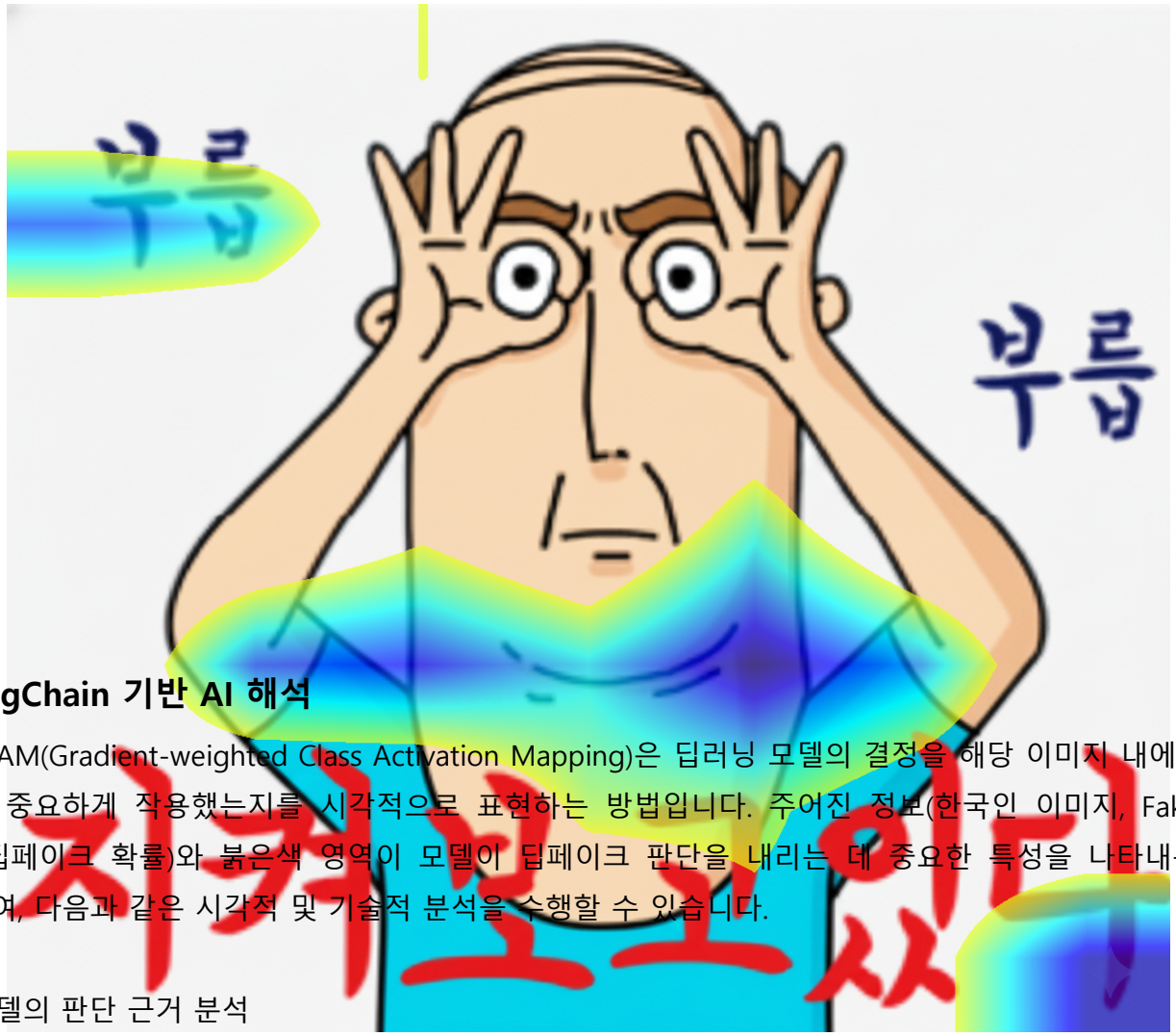
딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-06 16:39:19
- 예측 결과: Fake (60.63%)
- 딥페이크 확률: 21.50%

2. Grad-CAM 시각화



3. LangChain 기반 AI 해석

Grad-CAM(Gradient-weighted Class Activation Mapping)은 딥러닝 모델의 결정을 해당 이미지 내에서 어떤 지역이 중요하게 작용했는지를 시각적으로 표현하는 방법입니다. 주어진 정보(한국인 이미지, Fake 예측 결과, 딥페이크 확률)와 붉은색 영역이 모델이 딥페이크 판단을 내리는 데 중요한 특성을 나타내는 점에 기반하여, 다음과 같은 시각적 및 기술적 분석을 수행할 수 있습니다.

모델의 판단 근거 분석

1. **합성 흔적**:
   - 붉은색 영역에서 모델이 주목한 부분이 합성의 경계, 즉 얼굴의 턱선, 이마라인 또는 머리카락과 피부의 경계일 가능성이 높습니다. 이러한 경계는 두 개 이상의 이미지를 결합할 때 생기기 쉬운 아티팩트로,

자연스러운 피부 톤과의 끊김이 발견될 수 있습니다.
  - 합성의 흔적은 종종 인공지능이 이미지 내에서 자연스러운 전환을 만드는 데 실패했을 때 나타납니다. 이는 얼굴의 색상과 질감이 일관되지 않거나 부자연스럽게 느껴질 때 더욱 두드러집니다.

2. **피부 질감**:
  - 얼굴 피부의 질감은 일반적으로 자연스럽고 균일하지만, 딥페이크 이미지에서는 이질감 또는 과도한 부드러움(편집으로 보정된 부분)으로 인해 차이가 발생할 수 있습니다. 모델이 붉은 영역에서 이러한 질감을 주목했을 가능성이 높습니다.
  - 세부적인 피부 텍스처의 결여가 없거나 비현실적인 부분이 보일 경우, 이는 딥페이크의 특성 중 하나로 작용하고, 이를 통해 모델은 이러한 부분을 인식했을 것입니다.

3. **조명 왜곡**:
  - 조명의 불균형이나 왜곡 또한 딥페이크의 특성입니다. 자연광에서 촬영된 이미지와 인위적으로 조명된 합성 이미지는 차이가 날 수 있습니다. 특히 눈, 입, 코와 같은 주요 얼굴 부분의 조명이 일관되지 않는 경우, 모델이 그 부분을 붉은색 영역으로 강조할 가능성이 높습니다.
  - 이러한 조명 차이는 일반적으로 광고 촬영과 같은 인위적인 상황에서 발생할 수 있으며, 깊이감이 떨어지거나 그림자의 방향이 잘못되었을 때 나타납니다.

전문가의 관점에서 신뢰도와 한계점

- **신뢰도**:
  - Grad-CAM을 사용한 시각적 해석은 모델이 어떤 기준으로 판단했는지를 명확히 나타내주며, 딥페이크 탐지의 투명성을 높여주는 역할을 합니다. 비록 전체적인 예측 확률이 인간의 직관이나 다른 모델에 비해 낮을 수 있지만, 각 이미지에 대한 중요한 시각적 정보를 제공하여 결정의 신뢰성을 강화합니다.

- **한계점**:
  - Grad-CAM은 항상 정확한 원인을 설명하지 않을 수 있으며, 모델이 과거 데이터를 바탕으로 학습한 패턴에 고착될 수 있습니다. 만약 학습 데이터에 없었던 유형의 딥페이크가 등장하면, 모델이 그것을 잘 탐지하지 못할 수 있습니다.
  - 또한 모델이 시각적 데이터를 단순히 통계적 관점에서 분석하는 데 반해, 인간의 감각은 맥락과 문화적 배경에 따라 큰 차이를 보일 수 있습니다. 이로 인해 모델이 불확실한 경우에 대해 인간 전문가의 추가적인 판단이 필요할 수 있습니다.

심층 결과

- 모델의 결론에 도달할 때, 단순히 합성 흔적, 피부 질감, 조명 왜곡 외에도 다른 요인(예: 얼굴 비율, 표현의 일관성, 시선의 자연스러움 등)도 고려될 수 있습니다. 이러한 요소들은 추가적인 분석을 통해 더 깊이 있는 근거로 작용할 수 있으며, 특히 FaceSwap과 같은 고급 기술을 사용할 경우 더욱 두드러질 수 있습니다.

결론적으로, Grad-CAM을 통한 분석은 딥페이크 탐지 모델의 내적 작용과 시각적 근거를 이해하는 데 매우 유용하며, 앞으로의 발전에 있어서 인간 전문가와 AI의 협업이 더욱 중요할 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.