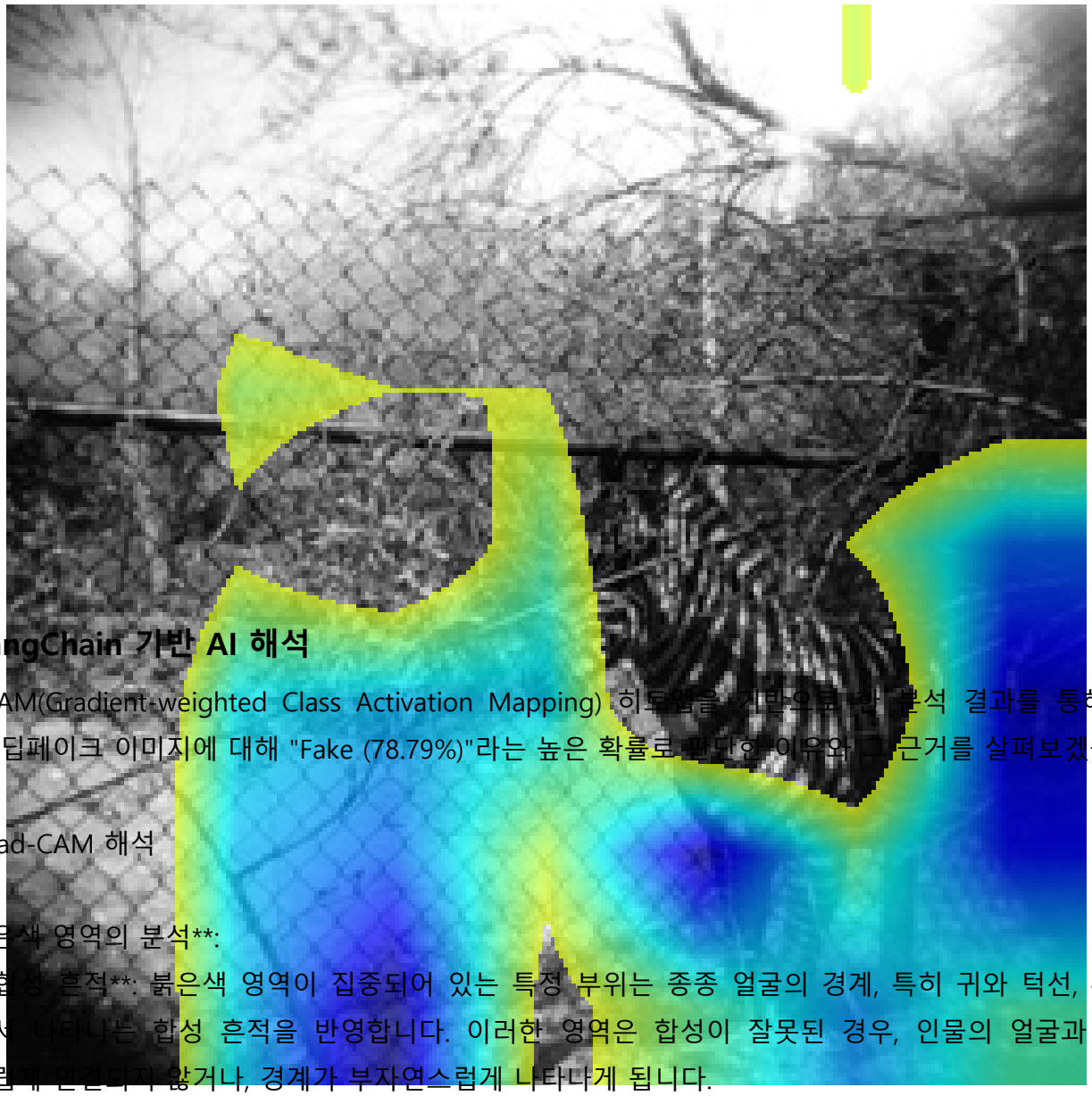
딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-10 00:20:32
- 예측 결과: Fake (78.79%)
- 딥페이크 확률: 34.00%

2. Grad-CAM 시각화



3. LangChain 기반 AI 해석

Grad-CAM(Gradient-weighted Class Activation Mapping) 히트맵을 기반으로 한 분석 결과를 통해, 해당 모델이 딥페이크 이미지에 대해 "Fake (78.79%)"라는 높은 확률로 판단한 이유와 그 근거를 살펴보겠습니다.

Grad-CAM 해석

1. **붉은색 영역의 분석**:
    - **합성 흔적**: 붉은색 영역이 집중되어 있는 특정 부위는 종종 얼굴의 경계, 특히 귀와 턱선, 혹은 눈 주위에서 나타나는 합성 흔적을 반영합니다. 이러한 영역은 합성이 잘못된 경우, 인물의 얼굴과 배경이 자연스럽게 연결되지 않거나, 경계가 부자연스럽게 나타나게 됩니다.

    - **피부 질감**: 딥페이크 모델은 종종 피부 질감을 잘 재현하지 못합니다. 이로 인해 붉은색 영역이

피부의 특정 부위에 집중된다면, 해당 영역에서 인공적인 질감, 고르지 못한 색상, 또는 과도한 매끈함이 보일 수 있습니다. 내추럴한 피부에서 찾아볼 수 있는 미세한 결점이나 주름이 отсутств되거나 단순화되면, 그 부분이 딥페이크로 인식될 확률을 높입니다.

- **조명 왜곡**: 조명이 불균형하게 나타나는 경우도 딥페이크를 탐지하는 중요한 지표입니다. 붉은색 히트맵이 특정 얼굴 부위(예: 이마, 턱 등)에서 비정상적인 조명 패턴을 나타낸다면, 그 부분이 왜곡된 색조나 그림자와 관련이 있을 수 있습니다. 특히 인물의 조명과 배경의 조명에서의 불일치는 딥페이크의 전형적인 특징입니다.

2. **확률적 결과 해석**:
- 모델이 "딥페이크 확률 34.00%"를 제시한 것은 딥페이크로 의심되는 요소들이 있지만, 모델이 모든 특징을 종합적으로 고려했을 때 100% 확신할 수는 없다는 것을 나타냅니다. 이는 모델이 신뢰할 수 있는 많은 정보가 있었음에도 불구하고, 완벽히 판별하지는 못했다는 것을 보여줍니다.

신뢰도와 한계점

- **신뢰도**: 모델의 예측이 78.79% 확신한다는 것은 관련 데이터셋에서 훈련된 경험을 바탕으로 상당히 신뢰할 수 있는 결과입니다. 특히, 피부 질감이나 조명과 같은 특정 특성과 관련된 시각적 기준이 일관되게 나타나면 모델의 신뢰도가 높아집니다.

- **한계점**:
  - **일반화된 특성**: 특정 이미지에서 나타나는 결과가 모든 경우에 해당하지 않을 수 있으며, 과적합(overfitting)의 위험이 있습니다.
  - **다양한 딥페이크 기법**: 딥페이크 기술은 날로 발전하고 있어, 최신 기법이나 변형된 방법에 대해서는 탐지 모델이 효과적이지 않을 수 있습니다.
  - **해석의 주관성**: Grad-CAM 기반의 분석은 시각적 해석에 의존하고 있어, 특정 영역이 주관적으로 해석될 수 있습니다. 서로 다른 전문가들이 같은 히트맵을 보고 다르게 해석할 가능성도 존재합니다.

심층 결과

- **추가 분석**: 딥페이크 탐지는 단지 시각적 요소에 국한되지 않고, 소리의 비동기성, 영상의 프레임 간 불일치, 얼굴의 움직임 패턴 불일치 등을 포함해야 합니다. 모델이 이런 심층 비교를 한다면, 보다 정교한 딥페이크 탐지가 가능할 것입니다.

결과적으로, 이 모델은 매우 효과적일 수 있지만, 다양한 외부 변수와 딥페이크 기술 발전의 속도에 따라 지속적으로 업데이트되고 검증되어야 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.