딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-05 14:59:02
- 예측 결과: 이 이미지는 Real로 분류되었으며, 예측신뢰도는 98.11%입니다.
- 딥페이크 확률: 23.03%

2. Grad-CAM 시각화

3. LangChain 기반 AI 해석

Grad-CAM(Gradient-weighted Class Activation Mapping)은 딥러닝 모델이 어떤 이미지의 특정 부분에서 어떤 특징에 주목했는지 시각적으로 표현할 수 있는 기술입니다. 여기서 제공된 정보를 바탕으로 Grad-CAM 히트맵을 해석해 보겠습니다.

해석:

1. **예측 결과 분석**:
   - 모델은 이미지를 98.11%의 높은 신뢰도로 실제(real) 이미지로 분류했습니다. 그러나 딥페이크 확률이 23.03%라는 점에서, 모델이 이 이미지에서 일부 딥페이크 특성을 포착했음을 암시합니다.
   - 두 가지 결과 간의 차이는 모델이 이미지의 일부 특징이 실제일 가능성을 높게 평가하면서도, 특정

위치나 패턴에서 합성된 특성을 발견했음을 나타냅니다.

2. **붉은색 영역의 해석**:
   - **합성 흔적**: 붉은색 영역이 주로 관찰되는 부분은 일반적으로 얼굴 주변, 특히 눈 주위, 입술, 그리고 피부의 결이 일관되지 않은 지역일 가능성이 높습니다. 이곳에서의 불일치는 합성 과정에서 나타나는 흔한 현상입니다.
   - **피부 질감**: 피부의 질감이 비정상적으로 매끄럽거나 비대칭인 경우, 이는 합성 기술의 자주 나타나는 특징입니다. 모델이 붉은색 히트맵으로 이러한 지역을 강조했다면, 실제 이미지에서 기대되는 자연스러운 질감이 부족하다는 것을 의미할 수 있습니다.
   - **조명 왜곡**: 얼굴의 조명이나 그림자가 부자연스럽게 나타난 부분도 붉은 영역으로 표현될 수 있습니다. 예를 들어, 조명 방향이 비일관된 경우 또는 그림자가 명확하지 않은 경우, 이는 모델이 인식한 딥페이크의 증거일 수 있습니다.

인간 전문가의 관점에서의 신뢰도와 한계:

- **신뢰도**:
  - 98.11%의 신뢰도로 이미지를 실제로 분류했다는 것은 모델이 많은 정보를 바탕으로 하여 신뢰성이 높다는 것을 시사합니다. 그러나 23.03%의 딥페이크 확률은 이 이미지에 대한 불확실성이 존재함을 의미합니다. 이런 낮은 확률로 인해 전문가나 사용자가 추가적인 검증 과정을 요구할 수 있으며, 모델에 대한 신뢰도를 높이기 위한 근거로 기능할 수 있습니다.

- **한계점**:
  - Grad-CAM은 주로 시각적 특징을 해석하는 도구지만, 실제 판별의 정확도나 신뢰성을 완전히 보장하지는 않습니다. 특정 코로나와 같은 조명이 강하게 작용하는 경우나 복잡한 배경이 포함되면, 잘못 해석될 수 있습니다.
  - 또한, 모델이 학습한 데이터셋의 다양성에 따라 결과가 달라질 수 있으며, 만약 학습 데이터에 딥페이크가 포함되어 있지 않다면, 모델이 이들의 패턴을 잘 포착하지 못할 수도 있습니다.

심층 결과:

이 분석을 통해 진단된 피쳐나 특정 부분 외에도, 사용자가 추가로 고려해야 할 요소들이 존재합니다:

- **얼굴의 비율**: 사람의 얼굴 비율이 비정상적으로 왜곡된 경우, 이는 모델이 인식하는 불일치로 나타날 수 있습니다.
- **동작 및 표정의 자연스러움**: 감정 표현이나 표정의 자연스러움이 떨어지는 경우에도 모델은 피쳐를 감지할 수 있습니다.

결론적으로, Grad-CAM을 통한 분석은 딥페이크 탐지에서 중요한 도구가 될 수 있지만, 모델의 한계와 함께

여러 가지 외부 요인과 개별적인 검토가 필요하다는 사실은 항상 유념해야 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.