딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-06 17:20:06
- 예측 결과: Real (48.54%)
- 딥페이크 확률: 13.30%

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 중요한 영역을 시각화해 주는 기법입니다. 주어진 정보를 바탕으로 Grad-CAM 히트맵을 해석해 보겠습니다.

모델의 판단 근거 분석

1. **붉은색 영역 해석**:
    - 붉은색으로 표시된 영역은 모델이 딥페이크 판단을 하는 데 있어 "중요한" 요소들입니다. 이 영역들은 합성된 이미지가 실제 이미지와 비교했을 때 자주 나타나는 특이점을 반영합니다.

2. **합성 흔적**:
   - 일반적으로 딥페이크 이미지에서는 경계가 부자연스럽거나, 머리와 배경의 경계가 흐릿하게 연결되는 경우가 많습니다. 따라서 붉은색 영역이 이 경계에 집중되어 있다면, 모델은 이러한 합성 흔적을 감지했을 가능성이 있습니다.

3. **피부 질감**:
   - 딥페이크는 종종 피부 질감이 불균형하게 나타납니다. 예를 들어, 합성이 자연 이미지를 재현할 수 없기 때문에 피부에 비정상적인 반짝임이나 텍스처의 불일치가 있을 수 있습니다. 붉은색 히트맵이 피부 영역에 나타난다면, 모델은 피부 질감의 왜곡을 감지한 것으로 해석됩니다.

4. **조명 왜곡**:
   - 딥페이크 이미지에서 조명이 일관되지 않은 경우가 많으며, 조명 방향이나 세기가 물체 간에 일관성을 가지지 않을 수 있습니다. 붉은 영역이 조명 차이가 있는 부분에 집중되어 있다면, 모델은 이러한 불일치를 감지한 것일 수 있습니다.

5. **표정 및 얼굴의 비대칭**:
   - 인간의 얼굴 특징은 대칭성이 높은데, 딥페이크에서는 이러한 대칭성이 잘 유지되지 않을 수 있습니다. 붉은 영역이 얼굴의 비대칭 부분에 나타나면, 모델이 이러한 특성을 인식하고 있다면, 이는 중요한 판단 근거가 됩니다.

신뢰도와 한계점

- **신뢰도**:
   - Grad-CAM은 모델이 주목하는 요소를 시각적으로 나타내므로, 해당 히트맵을 통해 피사체의 어떤 부분이 중요하게 작용했는지를 이해할 수 있습니다. 이를 통해 모델의 의사결정을 어느 정도 설명할 수 있어 신뢰성을 높입니다.

- **한계점**:
   - 그러나 Grad-CAM에서 나타나는 붉은 영역이 항상 실제로 딥페이크의 불일치를 의미하는 것은 아닙니다. 일반 이미지에서도 비슷한 시각적 특징이 나타날 수 있으므로, 도메인 일반성을 고려해야 합니다. 또한, Grad-CAM이 모델 백본의 특성이나 데이터셋의 불균형에 민감할 수 있어 오판을 유도할 가능성도 있습니다.

추가 심층 결과

- **대규모 데이터셋**:
   - 모델이 어떤 데이터셋에 기반하여 훈련되었는지에 따라 Grad-CAM의 해석이 달라질 수 있습니다. 만약 훈련 데이터에서 특정한 패턴이 있다면, 모델이 이를 과도하게 의존해 불필요한 특징을 강하게 인식하게 될

수 있습니다.

- **다양한 조건의 탐지**:
  - 모델이 다양한 조명, 표정, 각도 등을 갖춘 데이터를 학습했다면, 더 나은 일반화를 이룰 수 있지만, 이는 반대로 특정 유사한 이미지에 잘못된 예측을 할 수도 있음을 의미합니다.

이러한 요소들은 Grad-CAM 해석에 깊이를 더해주며, 딥페이크 탐지에 대한 신뢰성을 높이기 위해 지속적인 연구와 개선이 필요함을 강조합니다.

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.