

딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-06 16:19:14
- 예측 결과: Real (7.69%)
- 딥페이크 확률: 14.40%

2. Grad-CAM 시각화



3. LangChain 기반 AI 해석

Grad-CAM(Gradient-weighted Class Activation Mapping)은 딥러닝 모델, 특히 CNN(합성곱 신경망)의 시각적 인사이트를 제공하는 기법으로, 모델이 특정 클래스를 예측할 때 주목하는 이미지의 영역을 강조합니다. 주어진 정보에 따라 Grad-CAM 히트맵을 해석해보겠습니다.

Grad-CAM 해석

1. **붉은색 영역(딥페이크 판단 근거)**:

- 모델이 딥페이크 판단의 근거로 삼은 붉은색 영역은 통상적으로 다음과 같은 요소들을 포함할 수 있습니다:

- **합성 흔적**: 영상의 합성된 부분에서 경계가 부자연스럽거나 뚜렷한 경계선이 나타날 수

있습니다. 예를 들어, 얼굴과 배경의 자연스러운 혼합이 이루어지지 않은 경우에 이러한 흔적이 발견됩니다.

- **피부 질감**: 합성된 이미지에서는 피부 질감이 비현실적일 수 있습니다. 예를 들어, 매끈하거나 지나치게 부드러운 피부 표면이 보일 경우, 이는 합성의 특성을 나타낼 수 있습니다.
- **조명 왜곡**: 자연스럽지 않은 조명이나 그림자의 배치도 신호가 될 수 있습니다. 실제 인물의 그림자와 조명에서의 반사와 함께 잘 어우러지지 않는 모습을 보일 수 있습니다.

2. **합성 흔적**:

- 컴퓨터 비전 시스템은 픽셀 간의 불일치나 유사성을 분석하여 합성이 이루어진 영역을 탐지하는 데 강점을 가집니다. 만약 합성된 부분이 주변 영역과 일관성이 떨어지면, 모델은 이러한 것을 키포인트로 작용할 수 있습니다.

3. **피부 질감**:

- 딥페이크 이미지에서는 피부의 텍스처가 지나치게 매끄럽거나 불균형적으로 나타날 수 있습니다. 이는 원본 이미지와 비교했을 때 비현실적인 느낌을 줄 수 있으며, 모델이 이러한 점을 감지했다면, 해당 부분이 히트맵에서 강조될 가능성성이 높습니다.

4. **조명 왜곡**:

- 인간의 피부와 조명의 상호작용은 매우 복잡합니다. 특정 인물의 조명이 자연스럽지 않게 이루어진 경우, 모델이 해당 부분에 주목할 수 있습니다. 예를 들어, 한쪽 얼굴에만 강한 조명이 있을 경우 비정상적인 조명 패턴으로 인식될 수 있습니다.

전문가의 관점에서 신뢰도 및 한계점

- **신뢰도**:

- Grad-CAM은 딥러닝 모델이 어떻게 결정을 내리는지 시각적으로 보여주는 강력한 도구입니다. 주목한 영역이 실제로 딥페이크 판단의 근거가 되는 경우가 많기 때문에, 신뢰도가 상당히 높다고 평가할 수 있습니다.

- **한계점**:

- 그러나 모델의 신뢰도는 14.40%와 같이 낮은 딥페이크 확률로 나타나 최근의 기술에서의 잘못된 판단 가능성을 반영합니다. 모델이 적합한 데이터셋으로 훈련되지 않았거나, 특정 환경에서의 이미지 요소가 필수적으로 잘못 인식될 수 있습니다.

- Grad-CAM의 결과는 단지 CNN의 마지막 레이어가 보여주는 특성에 기반하기 때문에, 더 깊은 맥락이나 세부 사항을 놓칠 수 있습니다. 이는 하드웨어 및 소프트웨어의 강력한 반사 처리에 기반한 신뢰성 있는 분석을 제공하는 데 어려움이 있을 수 있습니다.

심층 결과

추가적으로, Grad-CAM은 모델의 최종 판단에 영향을 미치지는 않지만 주목한 특정 부분이 더욱 구체적으로

설명될 수 있습니다. 예를 들어, 해당 모델이 사용한 학습 데이터셋의 다양성이나, 데이터 전처리 과정에서의 노이즈나 변형, 그리고 전체적인 모델의 성능 또한 고려되어야 합니다. 이러한 요소들은 최종적인 판단 결과에 중요한 영향을 미칠 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.