

딥페이크 히트맵 분석 보고서

1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-06 16:42:49
- 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 모델이 특정 클래스에 대해 예측할 때 주의했던 이미지의 영역을 시각화하는 기법으로, 주로 CNN(Convolutional Neural Network) 모델의 해석에 사용됩니다. 질문에서 제공된 정보에 따라 Grad-CAM 히트맵을 해석하겠습니다.

Grad-CAM 해석:

1. **모델 예측 결과**:

- **Real (48.54%)**: 모델이 이미지를 실제라고 판단한 확률이 약 48.54%로, 실제 이미지일 가능성은 거의 50%에 가까운 상황입니다.

- **딥페이크 확률 (13.30%)**: 딥페이크일 가능성은 상대적으로 낮아 보이지만, 이는 해석의 중요한

부분입니다.

2. **붉은색 영역**:

- 붉은색 영역은 모델이 딥페이크 판단을 내리는 데 주목한 부분을 나타냅니다. 이 부분에서 주의 깊게 살펴봐야 할 몇 가지 시각적 요소는 다음과 같습니다.

기술적 분석:

1. **합성 흔적**:

- 딥페이크 영상은 종종 자연스러운 이미지와 구별되는 특징을 갖습니다. 붉은색 영역이 눈에 띄는 경우, 여기서 그레이디언트는 비정상적인 경계선이나 인공적인 패턴이 있는 부분을 강조할 수 있습니다. 피부의 색상 변화, 비정상적인 경계, 또는 인물의 얼굴에 적절히 일치하지 않는 모양 등에서 합성 흔적이 발견될 수 있습니다.

2. **피부 질감**:

- 딥페이크는 종종 피부 질감에서 부자연스러운 차이를 보입니다. 예를 들어, 지나치게 매끄럽거나 도포된 듯한 피부는 인공적인 특성을 드러냅니다. 붉은색 영역이 이러한 부자연스러운 피부 부분에 집중되어 있다면, 모델은 피부 질감의 불일치로 인해 딥페이크 가능성에 의심했을 수 있습니다.

3. **조명 왜곡**:

- 딥페이크 영상에서는 종종 조명의 일관성이 떨어지거나, 빛 반사가 비정상적인 경우가 많습니다. 붉은색 영역이 상대적으로 조명 차이가 큰 부분에 위치해 있다면, 이는 자연스러운 광원에 비해 부자연스러운 조명 변화를 의미할 수 있습니다.

신뢰도와 한계점:

1. **신뢰도**:

- Grad-CAM은 모델의 판단 근거를 시각적으로 제공하여, 모델이 어디에 주목했는지를 보여줍니다. 이 정보는 분석가 또는 전문가가 판단을 내리는 데 도움을 줄 수 있으며, 예측 결과의 신뢰도를 높이는 요소로 작용할 수 있습니다. 그러나 48.54%라는 결과는 확률적으로 명확한 결론을 내리기에는 부족한 수치일 수 있습니다.

2. **한계점**:

- Grad-CAM 역시 흐릿한 상황에 대한 해석은 주관적일 수 있습니다. 예를 들어, 모델이 제시한 붉은색 영역이 꼭 합성이라는 것이 아니며, 자연적인 이미지에서도 유사한 특징이 존재할 수 있습니다. 또한, Grad-CAM이 성공적으로 히트맵을 생성하더라도, 그 근거가 얼마나 실제적인지, 또는 더 정확한 다른 지표들과 통합해 해석해야 할 필요가 있습니다.

심층 결과:

- **추가적인 분석**: 모델이 특정 감정이나 표정에서 인식 오류를 범할 가능성도 있습니다. 예를 들어, 딥페이크가 특정 얼굴 표정을 과장했을 경우, 모델이 이를 비정상적이라고 판단할 수 있습니다. 이러한 미세한 특징을 인식할 수 있는지 여부도 평가해야 합니다.
- **컨텍스트**: 특정 문화적 배경이나 한국인의 특성을 반영하지 못한 경우, 모델의 판단이 왜곡될 수 있습니다. 한국인 이미지에 특화되어 훈련된 모델은 다양한 문화적 맥락에서 인식하는 데 한계가 있을 수 있습니다.

이러한 기술적 분석과 신뢰도, 한계점을 종합적으로 고려하여, 해당 Grad-CAM 히트맵을 해석하고 딥페이크 탐지의 정확성을 높일 수 있는 피드백으로 활용하십시오.

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.