

# 딥페이크 히트맵 분석 보고서

## 1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 한국인 전용 모델
- 분석 일시: 2025-11-06 14:22:14
- 예측 결과: Fake (56.71%)
- 딥페이크 확률: 12.30%

## 2. Grad-CAM 시각화



## 3. LangChain 기반 AI 해석

Grad-CAM(Gradient-weighted Class Activation Mapping)은 신경망 모델의 예측 결과를 시각적으로 설명하는 방법 중 하나입니다. 주어진 정보를 바탕으로 Grad-CAM 히트맵을 해석해보겠습니다.

### ### 모델 해석

#### 1. \*\*붉은색 영역\*\*:

모델이 붉은색으로 표시한 영역은 딥페이크 판단의 근거가 되는 중요한 부분입니다. 이 지역들은 모델이 관찰한 특징들이며, 이는 주로 다음 세 가지 측면에서 분석할 수 있습니다:

- \*\*합성 흔적\*\*: 만약 붉은색 영역이 얼굴의 턱선이나 귀 주위, 머리카락 경계 등 불규칙한 경계면에 위치하고 있다면 이는 합성의 흔적으로 해석될 수 있습니다. 딥페이크 기술이 종종 얼굴의 윤곽을

부자연스럽게 만들기 때문에, 이러한 특징들은 모델이 '가짜'로 판단하는 중요한 근거가 될 수 있습니다.

- **피부 질감**: 특정 붉은 영역이 피부의 질감에 관한 질문을 제기한다면, 이는 합성으로 인한 피부의 비정상적인 매끄러움 또는 결점의 부재와 관련이 있을 수 있습니다. 예를 들어, 자연적인 피부의 주름이나 결점에서 실패한 합성의 징후가 발견된다면 이는 모델이 진짜 이미지를 Fake로 판단하게 만드는 요소로 작용할 수 있습니다.

- **조명 왜곡**: 조명이 붉은색 영역과 관련될 경우, 이는 주로 인공적으로 생성된 조명 효과가 있을 때 발생합니다. 즉, 얼굴의 특정 부분에 비정상적인 그림자나 밝기 차이가 있다면 이는 모델이 딥페이크를 감지하게 되는 이유가 됩니다. 조명이 일관되지 않거나 부자연스러운 경향을 보인다면 이는 신뢰성을 떨어뜨리는 요인이 됩니다.

2. **딥페이크 확률**: 예측된 딥페이크 확률 12.30%는 모델이 확신하는 수준에 비해 상대적으로 낮습니다. 일반적으로 모델이 적은 확률로 딥페이크라고 판단할 경우, 해당 프레임이 내용 왜곡이 심하지 않거나 자연스러움을 유지하는 경우일 수 있습니다.

### ### 신뢰도와 한계점

- **신뢰도**: Grad-CAM은 딥러닝 모델이 어떤 특징으로 선택을 했는지를 시각적으로 나타내 주므로, 그 해석은 딥페이크 탐지에서 매우 유용할 수 있습니다. 전문가의 시각에서 이 히트맵은 부자연스러운 요소를 빠르게 식별하는 데 도움을 줄 수 있으며, 합성 이미지의 일관성 확인이나 추가적인 검증을 위한 기초 자료로 사용될 수 있습니다.

- **한계점**: 그러나 Grad-CAM은 언제나 완벽한 해석을 제공하지는 않습니다. 모델이 중요하게 여기는 특징이 반드시 사람에게도 명백하지 않을 수 있으며, 노이즈나 풀에서 발생하는 작은 그래디언트 변화로 인해 잘못된 판단을 유도할 수 있습니다. 또한, 딥페이크 기술이 발전하면서 생성된 이미지가 더욱 자연스럽고 일관되게 되기 때문에 Grad-CAM이 제공하는 해석이 시간이 지남에 따라 신뢰성을 잃을 수 있습니다.

### ### 추가적인 심층 결과

- **다양한 각도 분석**: 얼굴의 다양한 각도나 표현에서 모델의 성능이 어떻게 달라지는지도 중요한 요소입니다. 다양한 표정이나 조명 아래에서 테스트하면 모델의 성능 또는 의심할 만한 패턴을 찾아낼 수 있습니다.

- **배경과의 상호작용**: 배경의 특징, 즉 조명이나 색상의 변화도 중요한 분석 요소가 될 수 있습니다. 배경이 실제 촬영한 장면과 잘 어우러지지 않는다면 그 또한 딥페이크의 신호로 작용할 수 있습니다.

결론적으로, Grad-CAM 히트맵은 딥페이크 탐지의 신뢰성을 높이는 유용한 도구이지만, 신중한 해석이 필요하며, 전문가의 정교한 분석과 추가 데이터가 항상 요구됩니다.

## 40. 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.