

# 딥페이크 히트맵 분석 보고서

## 1. 분석 개요

- 모델명: MobileNetV3-Small
- 모델 유형: 외국인 전용 모델
- 분석 일시: 2025-11-06 15:21:03
- 예측 결과: Real (7.69%)
- 딥페이크 확률: 14.70%

## 2. Grad-CAM 시각화



## 3. LangChain 기반 AI 해석

Grad-CAM(Gradient-weighted Class Activation Mapping)은 딥러닝 모델의 예측에서 중요하게 여긴 이미지의 특정 영역을 시각화하는 도구입니다. 여기서 붉은색 영역은 모델이 '딥페이크' 판단을 내리는 데 기여한 부분을 나타냅니다. 주어진 정보를 기반으로 Grad-CAM 히트맵을 해석하고, 모델의 판단 근거를 기술적으로 분석해보겠습니다.

### 모델의 판단 근거 해석

### 1. \*\*합성 흔적\*\*:

- 붉은 영역에서 합성 흔적이 관찰되는 경우, 특히 피부의 경계나 표정의 인위적인 변화가 있을 수 있습니다. 이러한 부분은 진짜 이미지에서 흔히 나타나는 자연스러운 피부 결이나 주름과 달리, 인위적이고

부자연스러운 형태로 나타날 수 있습니다. 따라서 이 부분은 모델이 합성 여부를 판단하는 중요한 요소가 될 수 있습니다.

2. **\*\*피부 질감\*\***:

- 모델이 붉은색으로 강조한 부분에 피부 질감의 고르지 못한 부분이 포함되었다면, 이는 자연적이지 않은 표면 처리나 색상의 불균형을 나타낼 수 있습니다. 딥페이크 기술은 종종 피부 질감을 부드럽게 처리하거나 인위적으로 조작하여 자연스러운 피부 톤을 재현하는 데 어려움을 겪기 때문에 이러한 요소는 딥페이크 탐지의 중요한 신호가 됩니다.

3. **\*\*조명 왜곡\*\***:

- 조명은 이미지의 분위기와 현실감을 결정짓는 중요한 요소입니다. 만약 붉은색 영역에서 조명에 대한 불일치가 보이면, 이는 딥페이크에서 조명 조건의 불일치를 나타낼 수 있습니다. 즉, 얼굴과 배경의 조명이 서로 다르거나 전반적인 조명 패턴이 일관성이 없을 경우 이는 합성된 이미지를 암시할 수 있습니다.

### 신뢰도와 한계점

- **\*\*신뢰도\*\***:

- Grad-CAM 히트맵은 모델이 주목하는 이미지의 특정 영역을 시각적으로 표현하기 때문에, 모델이 어떤 근거로 결정을 내렸는지를 파악하는 데 유용합니다. 하지만 이 결과만으로 절대적인 결정을 내리기엔 위험할 수 있습니다. 모델이 특정 특징에 과도하게 의존하여 시각적 판단을 내릴 수도 있기 때문에, 이러한 시각화는 함께 사용되는 여러 검증 방법의 하나로 이해해야 합니다.

- **\*\*한계점\*\***:

- Grad-CAM은 주로 CNN(Convolutional Neural Network) 기반 모델에 효과적이며, 해당 모델이 항상 100% 정확한 판단을 한다고 보장할 수는 없습니다. 예를 들어, 특정 조명을 가정하면 그것이 진짜로 보일 수 있는 경우, 이럴 때 모델이 오판할 가능성도 존재합니다.

- 또한, 히트맵 자체는 모델의 예측의 '믿음도'를 표현하기는 하지만, 이는 단순히 모델의 계산적인 해석에 의존하므로, 복잡한 심리적, 문화적 요소, 사람의 감정 판단을 충분히 반영하지 못할 수 있습니다.

### 추가적 심층 결과

- 만약 모델이 특정 기본적인 인물 특징(예: 눈, 코, 입 등)을 감지하고 붉은색 영역으로 강조했다면, 이는 에칭작업의 흔적이 있을 수도 있습니다. 인물의 표정, 움직임이 자연스럽지 않거나 부자연스럽게 조정된 경우, 이러한 요소도 함께 분석해야 신뢰성을 높일 수 있습니다.

- 마지막으로, 각기 다른 딥페이크 기술의 특성을 이해하고 관련 데이터셋에서 훈련된 모델의 한계를 알고 사용해야 합니다. 예를 들어, 다양한 문화적 배경을 가진 인물의 얼굴을 다루는 경우, 그 특성을 모델이 얼마나 잘 학습했는지가 중요한 평가 기준이 됩니다.

이러한 다양한 관점에서 분석하고 결과를 평가함으로써, 보다 신뢰성 있는 딥페이크 탐지 작업을 진행할 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.