# CAM과 Grad-CAM 성능 비교하기

**성연우**

January 10, 2025

## Abstract

본 연구는 Stanford-Dogs 데이터셋에서 ResNet50 모델을 활용하여 CAM과 Grad-CAM의 성능을 비교하였다. 에폭별로 모델을 분석한 결과, Grad-CAM이 CAM보다 좋은 성능을 보여 객체 전체를 더 효과적으로 검출하였다. IoU 기반 평가를 통해 에폭에 따른 Grad-CAM의 강점을 확인하였다.

## 1 Introduction

본 연구는 120개의 개 품종으로 구성된 Stanford-Dogs 데이터셋과 ResNet50 모델을 활용하여 CAM과 Grad-CAM을 비교하였다. 해당 데이터셋은 라벨링된 바운딩 박스를 포함하고 있어 IoU 기반 성능 평가에 적합하다. 에폭 3, 5, 7의 모델을 선정하여 Grad-CAM과 CAM의 과적합 경향 및 성능 차이를 분석하였다. 특히 Grad-CAM은 모든 에폭에서 CAM보다 객체 전체를 더 정확히 포착하며 높은 IoU 결과를 보였다. 이를 통해 Grad-CAM이 ResNet50 모델에서 국소화 성능을 크게 향상시키는 것을 확인하였다.

## 2 Background

CAM은 마지막 convolution layer의 feature map을 활용해 input 이미지의 내용을 함축적으로 표현하며, Global Average Pooling(GAP) layer와 Fully Connected(FC) layer의 weight를 학습해 heatmap을 생성한다. 그러나 GAP 및 FC layer가 반드시 필요하며, 중간 layer에서는 CAM 사용이 제한된다는 단점이 있다. Grad-CAM은 기존 CNN 구조를 수정하지 않고도 각 feature map의 gradient를 활용해 weight를 계산하므로, 재학습 없이도 heatmap을 생성할 수 있다. 또한, Grad-CAM은 CAM과 비교해 객체의 전체적인 영역을 더욱 정확히 표현하는 장점이 있다. 이러한 차이를 통해 두 기법의 성능과 국소화 능력을 비교하고자 한다.

## 3 Method

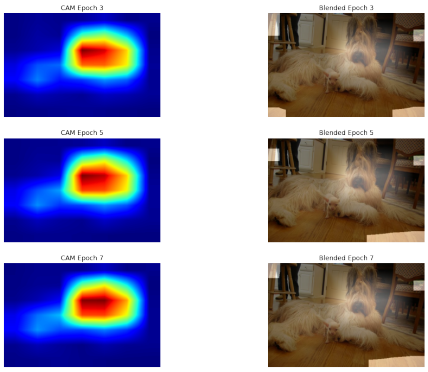
본 연구는 resnet50 모델을 활용해 CAM과 Grad-CAM의 성능을 비교하기 위해 진행하였다. Stanford-Dogs 데이터셋을 활용하여 실험을 진행 하였다. 해당 데이터셋은 120개의 개 품종으로 구성되어 있으며, 각 이미지에 라벨링된 바운딩 박스 정보를 포함하고 있다. 이러한 특성은 Grad-CAM과 CAM의 국소화 성능을 비교하기 위한 IoU 기반 평가에 적합하다. 데이터셋은 학습(train)과 테스트(test)로 분리된 상태에서 제공되며, 실험에서는 데이터셋의 테스트 세트를 사용하여 모델의 국소화 성능을 검증하였다. CAM은 마지막 convolution layer와 Global Average Pooling(GAP) layer를 통해 heatmap을 생성하였다. 반면, Grad-CAM은 기존 ResNet50 구조를 수정하지 않고도 gradient 정보를 사용해 각 feature map의 중요도를 계산한다. 이를 통해 Grad-CAM은 기존 모델의 추가 학습 없이 국소화 성능을 확인할 수 있는 장점을 가진다. 또한, Grad-CAM 논문 [Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization]에서 ResNet50을 활용한 점을 참고하였으며, Residual Learning 기반 설계로 매우 깊은 네트워크에서도 효과적으로 학습할 수 있다는 점을 고려해 비교 실험에 ResNet50을 선택하였다.

성능 비교를 위해 에폭 3, 에폭 5, 에폭 7의 총 3가지 모델을 선정하였다. 모델 학습을 3회 반복한 결과, 1회차와 2회차에서는 에폭 5에서 가장 높은 성능을 보였으나, 3회차에서는 에폭 3에서 가장 우수한 성능을 확인하였다. 모든 학습 회차에서 각 에폭 이후 오버피팅이 발생하는 경향을 관찰할 수 있었다. 특히 에폭 7에서는 오버피팅이 발생했음에도 불구하고,

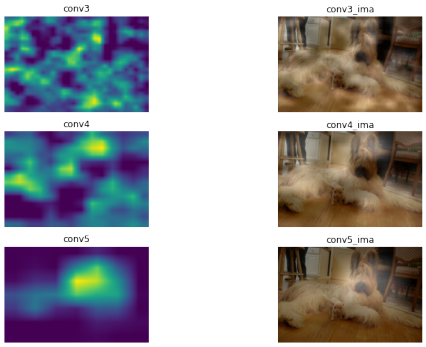
CAM과 Grad-CAM 이미지를 시각적으로 분석하고 IoU를 비교한 결과, 일부 사례에서 에폭 7이 더 나은 결과를 보이는 것을 확인하였다. 이에 따라 에폭별 성능 비교의 필요성을 인지하고, 추가적인 분석을 위해 accuracy가 가장 높고 loss가 가장 낮았던 2회차를 기준으로 비교를 진행하였다.

| | Epoch 3 | | Epoch 5 | | Epoch 7 | |
|---|---|---|---|---|---|---|
| | accuracy | loss | accuracy | loss | accuracy | loss |
| 1 회차 | 0.709655 | 1.025016 | 0.731460 | 0.964633 | 0.730294 | 0.993686 |
| 2 회차 | 0.727495 | 0.965132 | 0.739506 | 0.937959 | 0.734725 | 0.975292 |
| 3 회차 | 0.728195 | 0.957389 | 0.728195 | 0.980687 | 0.726912 | 1.005752 |

CAM을 사용하여 에폭 3, 에폭 5, 에폭 7의 결과를 이미지로 비교한 결과, 모든 에폭에서 대상을 정확히 검출하면서도 일부 배경이 함께 검출되는 현상이 관찰되었다. 에폭 3에서는 배경의 일부가 더 넓게 포함되어 검출되는 것을 시각적으로 확인할 수 있었다. 에폭 5에서는 배경의 검출 범위가 에폭 3에 비해 감소하여 더 정확한 결과를 보였다. 에폭 7의 경우, 에폭 5와 거의 동일한 영역을 검출하며 유사한 성능을 나타냈다.



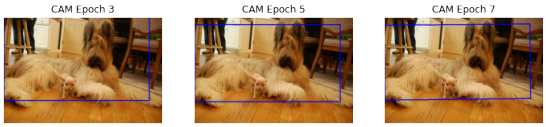

Grad-CAM은 각 레이어에서 모델이 검출하는 영역을 시각화할 수 있는 장점을 활용하여, conv3-block4-out, conv4-block6-out, conv5-block3-out의 3개 레이어에서 결과를 확인하였다. conv3-block4-out에서는 세부적인 특징을 포착하려는 경향으로 인해 이미지의 여러 부분에 산재된 형태를 보였다. conv4-block6-out에서는 이전 레이어보다 더 큰 특징을 포착하였으나, 여전히 강아지보다는 배경에서 주요 특징을 찾는 모습이 관찰되었다. 마지막 레이어인 conv5-block3-out에서는 강아지의 얼굴이 뚜렷하게 나타났으며, 몸통 부분도 연하게 시각화되는 것을 확인할 수 있었다.




CAM과 Grad-CAM의 결과를 비교한 결과, Grad-CAM이 CAM보다 강아지의 몸통까지 포함하여 더욱 전체적인 영역을 정확히 검출하는 것을 확인할 수 있었다.

# 4 Result

CAM의 각각의 epoch 별 바운딩 박스의 모양은 epoch 3이 가장 크고 epoch 5, epoch 7이 동일한 크기로 같다.




AM의 epoch별 바운딩 박스 크기는 epoch 3에서 가장 크고, epoch 5와 epoch 7은 비슷한 크기를 보였다. IoU 점수는 epoch 3에서 0.782, epoch 5에서 0.793, epoch 7에서 0.800으로 증가하는 경향을 보였다. 이는 학습이 진행되며 모델이 객체를 더 정확히 포착했음을 시사한다.

Grad-CAM의 각각의 epoch 별 바운딩 박스는 시각적으로 판단하기 어려울 정도로 큰 차이가 나타나지 않기 때문에 IoU를 활용하여 결과를 확인하였다.


Grad-CAM의 IoU 점수는 epoch 3에서 0.881, epoch 5에서 0.883, epoch 7에서 0.816으로 epoch 7에서 성능이 저하되는 양상을 보였다. 이러한 결과는 Grad-CAM이 모델의 학습 단계에 따라 객체를 더 효과적으로 포착하면서도, 과적합이 발생한 경우 성능이 저하될 수 있음을 나타낸다.

# 5 Discussion

CAM과 Grad-CAM의 결과를 비교했을 때, Grad-CAM이 CAM보다 IoU 점수에서 더 높은 성능을 보였다. 특히 Grad-CAM은 객체의 전체적인 형상을 더 정확히 반영하며, CAM이 배경을 포함하는 경향이 있는 것과 대조적이었다. 다만, Grad-CAM에서 epoch 7의 IoU가 낮아진 것은 과적합에 따른 성능 저하로 해석될 수 있다. 이는 학습 데이터에 과도하게 집중할 경우, 모델이 일반화 능력을 잃게 되는 현상을 반영한다. 이러한 분석은 Grad-CAM이 CAM 대비 우수한 국소화 성능을 제공하면서도, 과적합 방지를 위한 추가적인 기법이 필요함을 시사한다.

# 6 References

[1] Ramprasaath R, [Grad-CAM: Visual Explanations from Deep Networksvia Gradient-based Localization], cs.CV, 3 Dec 2019. [2] Bolei Zhou, [Learning Deep Features for Discriminative Localization], cs.CV, 14 Dec 2015.