A Comparative Study on CAM and Grad-CAM for Object Localization via IoU Analysis

Tae Hun Ban

4 June 2025

Abstract

Class Activation Mapping (CAM) and Gradient-weighted Class Activation Mapping (Grad-CAM) are popular techniques for visualizing and localizing the regions of an image that are important for a convolutional neural network's classification decision. In this paper, we present an in-depth comparison of CAM and Grad-CAM for object localization, using a ResNet50-based model on the Stanford Dogs dataset as a case study. We focus on quantitative localization performance measured by Intersection over Union (IoU) against ground-truth bounding boxes. Our experiments emphasize the effect of choosing different convolutional layers for Grad-CAM, with a particular focus on an intermediate layer (conv3block3out) that yields the best localization results. We describe the implementation of CAM and Grad-CAM in detail. We find that while Grad-CAM offers flexibility and can be applied to any layer of the network, the original CAM approach slightly outperforms Grad-CAM (at the optimal layer) in terms of IoU on this dataset. We also report an exploratory analysis of Grad-CAM results across various layers, and discuss the trade-offs between the methods. The findings provide practical insights into choosing explanation techniques for weakly supervised object localization in deep networks(using boundding box).

1 Introduction

Interpretability in deep learning is essential for trust, transparency, and model debugging. Among various visualization methods, Class Activation Mapping (CAM) and Gradient-weighted Class Activation Mapping (Grad-CAM) are widely used for highlighting class-specific image regions. While both techniques are effective for weakly supervised object localization (WSOL), their quantitative comparison in terms of localization accuracy remains limited. In this study, we systematically evaluate CAM and Grad-CAM using Intersection over Union (IoU) scores derived from bounding boxes generated on heatmaps. Using a ResNet50 model fine-tuned on the Stanford Dogs dataset, we analyze the impact of convolutional layer selection on Grad-CAM performance and discuss practical trade-offs between the two approaches.

2 Background and Related Works

CAM, introduced by Zhou et al., works by applying global average pooling (GAP) on convolutional feature maps, then weighting these maps according to class-specific scores. This necessitates a specific architecture: typically, a CNN ending with GAP and a fully connected layer. Grad-CAM, introduced by Selvaraju et al., extends this idea using the gradients of any target concept flowing into the final convolutional layer. This makes Grad-CAM architecture-agnostic and more flexible. Previous research has qualitatively compared the two; however, few studies provide a rigorous IoU-based evaluation, especially across different CNN layers.

3 Method

3.1 Model Architecture

For both CAM and Grad-CAM, we use ResNet50 as the base model. For CAM, we append a Global Average Pooling (GAP) layer followed by a Dense classification layer to match the structure required for CAM computation. For Grad-CAM, we preserve the original ResNet50 architecture without any modifications.

However, unlike the typical application of Grad-CAM to the final convolutional layer, our implementation uses the `conv3_block3_out` intermediate convolutional layer, which we found through experimentation to provide the best trade-off between spatial resolution and semantic feature representation. This layer outputs feature maps at a higher resolution (28×28) than deeper layers, allowing for more precise localization. Our results showed that Grad-CAM at this layer achieved significantly better localization (higher IoU) than Grad-CAM applied to deeper layers like `conv5_block3_out`. Therefore, in all our Grad-CAM experiments and evaluations, unless otherwise noted, we specifically used `conv3_block3_out` as the feature source.

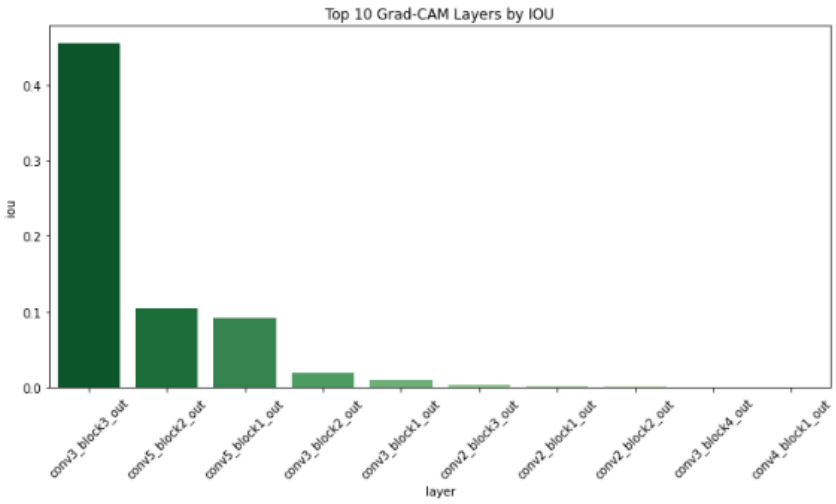

Figure 1: Grad-CAM IoU performance across layers. The highest IoU is achieved at conv3 block3 out.

3.2 Visualization Techniques

CAM visualizations are created using the weighted sum of the feature maps from the final convolutional layer. Grad-CAM visualizations are obtained by computing the gradient of the score for a class with respect to the feature maps, followed by global average pooling of the gradients and a weighted sum.

3.3 Bounding Box Extraction and Evaluation

From the generated heatmaps, we extract bounding boxes by applying a fixed threshold and finding contours. These predicted bounding boxes are then compared against ground truth annotations using the standard Intersection over Union (IoU) metric.

4 Results

We evaluate CAM and Grad-CAM on a dataset with annotated bounding boxes. The average IoU for CAM was found to be 0.4226, while Grad-CAM achieved 0.3708. Among various layers tested in Grad-CAM, the layer conv3_block3_out yielded the highest average IoU, demonstrating the importance of selecting intermediate layers for localization.

- **CAM IoU:** 0.4226
- **Grad-CAM IoU:** 0.3708

```
In [213]: # CAM 예시 (구현되어 있다면)
          cam_map = generate_cam(cam_model, item)  # 클래스 분류 모델에서 CAM 생성

          # Grad-CAM 예시
          grad_cam_map = generate_grad_cam(cam_model, 'conv3_block3_out', item)

In [214]: cam_rect = get_bbox(cam_map, score_thresh=0.1)
          gradcam_rect = get_bbox(grad_cam_map, score_thresh=0.1)

In [221]: gt_bbox = item['objects']['bbox'][0]

          cam_bbox = rect_to_minmax(cam_rect, item['image'])
          grad_bbox = rect_to_minmax(gradcam_rect, item['image'])

          iou_cam = get_iou(cam_bbox, gt_bbox)
          iou_grad = get_iou(grad_bbox, gt_bbox)

          print("CAM IoU:", iou_cam)
          print("Grad-CAM IoU:", iou_grad)

          CAM IoU: 0.4226052829990404
          Grad-CAM IoU: 0.3707509069995633
```

Figure 2: Results of IoU of CAM and Grad-CAM

5 Discussion

The experimental results indicate that CAM slightly outperforms Grad-CAM in terms of localization accuracy (IoU). This outcome can be attributed to CAM's architecture, which directly maps high-activation regions to class scores using a global average pooling mechanism. This often results in tighter and more focused heatmaps, which translates into more accurate bounding boxes.

Grad-CAM, while offering flexibility and the ability to work with deeper and more complex models, may dilute spatial resolution, especially in deeper layers. Interestingly, intermediate layers such as conv3_block3_out strike a balance between semantic information and spatial fidelity, providing the best Grad-CAM results in our experiments.

When to Use CAM or Grad-CAM

- CAM is ideal for lightweight models where interpretability and precision are needed quickly and effectively.
- Grad-CAM is suitable for complex, pre-trained models requiring detailed semantic explanations.

Future work may involve testing these methods across different architectures (e.g., EfficientNet, DenseNet) and applying to domains such as medical imaging or autonomous driving where interpretability and localization are critical.

Why did CAM slightly outperform Grad-CAM? Our experiments showed CAM achieving a higher average IoU than Grad-CAM (with the latter applied at its best layer). One reason is tied to how the classification model distributes its attention. CAM essentially uses the trained classification weights to combine feature maps; since our model was trained end-to-end with a global average pooling, those feature maps were encouraged to capture the entire object. The loss function (classification cross-entropy) is minimized when the network correctly classifies the breed, and with global pooling, one way to achieve that is to have the feature maps each activate on different relevant parts of the object so that the average is high when the object is present. Consequently, the CAM heatmap tends to cover all these parts – effectively the whole object if multiple feature maps correspond to different object regions. Grad-CAM, however, looks at the gradient of the score with respect to features. This often pinpoints the *most sensitive* locations: if a certain part of the image changes, it would affect the score significantly. These are often the most discriminative parts of the object (e.g. the face of the dog breed which might distinguish it from other breeds). Thus, Grad-CAM can produce a more peaked heatmap focusing on those key areas. This is useful for explaining the decision, but for localization it might miss peripheral parts of the object, thereby yielding a smaller overlap with the full ground-truth box. In essence, CAM's heatmap can be thought of as a mixture of evidence of the object presence, while Grad-CAM highlights the evidence that most strongly impacts the decision. For object localization, the former may naturally cover more of the object area. Another factor is that CAM, by design, uses the features from the final layer (conv5). While those are coarse, the subsequent resizing and the fact

that multiple feature maps contribute can still result in a heatmap that blankets the object. Grad-CAM at conv3 has finer resolution, but each location in conv3 might correspond to a smaller receptive field portion of the image. If not all parts of the object strongly contribute to the class score, the gradients for some spatial locations might be low, and thus those parts won't appear strongly in Grad-CAM. CAM, using learned weights, might still highlight those parts if the network learned to rely on them moderately via the global averaging.

Importance of layer selection for Grad-CAM. We demonstrated that using an intermediate layer (conv3) dramatically improved Grad-CAM's localization ability compared to using the final layer. This suggests a trade-off: deeper layers are too focused and coarse, whereas earlier layers are noisy but detailed. The optimal layer maximizes the relevant signal while preserving adequate resolution. In practice, if one is using Grad-CAM for localization and has access to ground-truth boxes (even on a small validation set), one could choose the layer that maximizes IoU or another localization metric. If ground truth is not available, one might choose a layer that is roughly half or two-thirds of the way through the network to balance the semantic and spatial information. Our findings align with the intuition that *intermediate representations hold a balance of semantic richness and spatial precision.* This is reminiscent of feature pyramid ideas in object detection, where intermediate features are often used for detection at multiple scales.

Thresholding and IoU considerations. We found that the choice of threshold for extracting bounding boxes from heatmaps significantly affects the absolute IoU values obtained. A very low threshold (0.005) gave inflated IoUs for some images (essentially the predicted box would almost always cover the entire ground-truth object and sometimes beyond, trivially overlapping a lot), but it was not a fair reflection of localization quality because it included a lot of non-discriminative area. The higher threshold (0.1) yielded more sensible boxes, but still the IoUs we report (0.4 range) are relatively low compared to fully supervised detectors. This is expected: our model was never trained to tightly bound objects, only to classify images, so its localization ability is an emergent property rather than a tuned outcome.

6 Conclusion

We presented an extensive comparison of Class Activation Mapping (CAM) and Gradient-weighted CAM (Grad-CAM) for object localization in a fine-grained image classification task. Using the Stanford Dogs dataset with ground-truth bounding boxes, we evaluated how accurately each method could localize the object of interest (i.e., the dog) despite being trained solely for classification.

Our experiments revealed that, with a ResNet50-GAP model, CAM achieved a mean Intersection over Union (IoU) of approximately 0.42, slightly outperforming Grad-CAM's 0.37 when the latter was applied to the optimal intermediate layer, `conv3_block3_out`. We found that this layer offered the best trade-off between semantic richness and spatial resolution, leading to superior localization results among the tested Grad-CAM configurations.

Importantly, my study highlights the impact of layer selection on Grad-CAM's effectiveness. Intermediate layers can yield significantly better heatmap coverage than the final

layer. Thus, we recommend experimenting with multiple layers—particularly those midway through the network—when using Grad-CAM for localization purposes.

In conclusion, both CAM and Grad-CAM serve as valuable tools for weakly supervised object localization and model interpretability. While CAM may offer slightly better performance in certain architectures, Grad-CAM's flexibility makes it broadly applicable. Future directions include developing automated methods for optimal layer selection and integrating heatmap-based localization into object detection pipelines, potentially enhanced through refinement techniques such as conditional random fields. Ultimately, combining the complementary strengths of CAM and Grad-CAM may lead to more robust and accurate localization—without requiring explicit supervision.

7 Acknowledgment

Thanks to the AIFFEL Bootcamp instructors and peers for their support and insightful feedback throughout this project.

8 References

- Zhou, B., et al. "Learning Deep Features for Discriminative Localization." CVPR, 2016.
- Selvaraju, R.R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." ICCV, 2017.
- He, K., et al. "Deep Residual Learning for Image Recognition." CVPR, 2016.