Homework 2 – Computer Vision

Image Segmentation without User Input

Grad-CAM-based Point Selection for SAM

Mikołaj Woźniak[1]

[1]University of Warsaw, Faculty of Mathematics, Informatics and Mechanics

Abstract

This report addresses the problem of automatic image segmentation without user-provided prompts. The task is divided into two stages. First, we implement Grad-CAM, a gradient-based visualization method for convolutional neural networks, to identify image regions most relevant to a classifier's prediction. Second, we design a multi-stage pipeline that leverages Grad-CAM outputs to automatically generate foreground and background point prompts for the Segment Anything Model (SAM), enabling segmentation using only the image as input.

The proposed approach is evaluated on a custom dataset derived from CIFAR-10, containing geometric shapes with ground-truth segmentation masks. We report metrics for both the intermediate point-selection stage and the final segmentation stage using Intersection over Union (IoU). The results demonstrate that Grad-CAM-based point selection enables effective segmentation without manual intervention.

Keywords: Grad-CAM, image segmentation, SAM, explainable AI

1 Introduction

Most modern image segmentation models rely on user-provided prompts, such as bounding boxes or point coordinates, which limits their use in fully automated systems. The Segment Anything Model (SAM) achieves strong segmentation performance but still depends on external guidance.

Explainability techniques like Grad-CAM provide spatial insight into convolutional neural network predictions by highlighting regions most relevant to a given class. These activations preserve coarse spatial structure and can therefore be repurposed for automatic prompt generation.

In this work, we investigate the use of Grad-CAM to automatically generate point prompts for SAM, enabling image segmentation without user input. We implement Grad-CAM from scratch and propose two prompt-generation pipelines based on activation heatmaps.

2 Materials and Methods

2.1 Dataset

We use a custom dataset derived from CIFAR-10 consisting of images that contain exactly one geometric object: a circle, square, diamond, triangle, or star. Each image is paired with a ground-truth binary segmentation mask.

2.2 Grad-CAM

Grad-CAM is implemented following the original formulation [1]. For a chosen convolutional layer, gradients of the target class score with respect to feature maps are globally averaged to obtain channel-wise importance weights. The weighted sum of feature maps is passed through a ReLU function and normalized to the range $[0, 1]$.

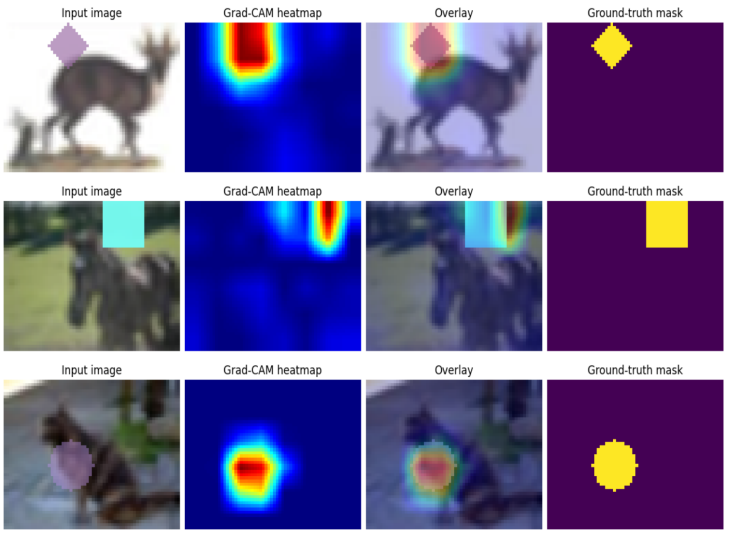

Figure 1: Example Grad-CAM activation maps for input images and predicted target class

2.3 Heatmap Transformation

To improve spatial coherence, Grad-CAM heatmaps are post-processed using an inverse-distance weighted convolution kernel. The transformed heatmap is obtained by convolving the original heatmap with this kernel and adding the result back to the original activation map. This operation smooths and expands salient regions, making point selection more robust.

Let $H \in \mathbb{R}^{m \times n}$ denote the original Grad-CAM heatmap. We apply an inverse-distance weighted convolution kernel $K \in \mathbb{R}^{(2r+1)\times(2r+1)}$, defined as

$$K(i,j) = \frac{1}{\sqrt{i^2 + j^2} + \varepsilon}, \quad i,j \in [-r,r] - \{0,0\}, \tag{1}$$

where $K(0,0) = 1$.

The transformed heatmap $\tilde{H}$ is defined as

$$\tilde{H} = H + H * K, \tag{2}$$

where $*$ denotes two-dimensional convolution.

This transformation improves the robustness of subsequent point selection by reinforcing spatially consistent activation regions.

2.4 SAM Pipelines

Empirically, we observe that SAM often requires only a single correctly placed foreground point to produce an accurate segmentation. Therefore, the objective of this stage is to identify pixels that are most likely to belong to the target object.

To achieve this, we use the transformed Grad-CAM heatmap to prioritize pixels located near clusters of high activation values. Based on this observation, we propose two automatic prompt-generation pipelines:

- **OnlyForegroundSamPipeline**: selects the top-K pixels with the highest transformed Grad-CAM activation and labels them as foreground points. We find that $K = 1$ yields the best performance.
- **BackgroundAndForegroundSamPipeline**: selects foreground points from high-activation regions, computes their center of mass, and samples background points at a sufficient distance from this center. We find that 1 foreground pixel and 1 background pixel with distance > 20 yielded the best performance.

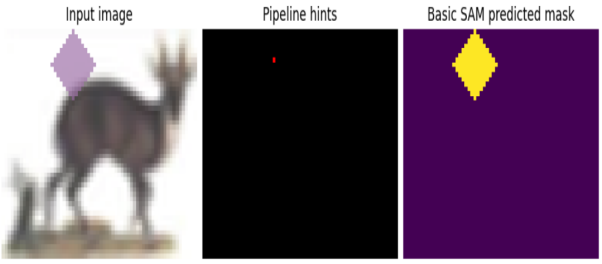

Figure 2: OnlyForegroundSamPipeline prompts and SAM prediction. Red points indicate the foreground class.

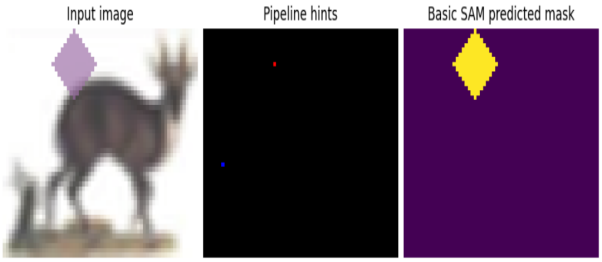

Figure 3: OnlyForegroundSamPipeline prompts and SAM prediction. Red points indicate the foreground class, while blue points indicate the background class.

3 Results

The foreground-only pipeline achieves a slightly higher hit rate and higher IoU, indicating that selecting points from the most salient Grad-CAM regions is sufficient for accurate segmentation on this dataset. Introducing background points slightly reduces performance, likely due to imperfect background sampling when object boundaries are ambiguous.

Nevertheless, both pipelines exceed the required IoU threshold, demonstrating that Grad-CAM provides reliable spatial cues for fully automatic image segmentation.

Table 1: Final dataset metrics for both SAM pipelines.

Pipeline	Hit Rate	IoU	Distance
OnlyForeground	0.773	0.769	9.290
Foreground + Background	0.770	0.765	9.468

Discussion

The results lead to two main conclusions. First, Grad-CAM performs particularly well on ResNet-based classifiers, highlighting image regions that are crucial not only for the model's decision but also for

human visual interpretation. Second, these highly activated pixels serve as effective prompts for SAM, enabling accurate segmentation without manual input. Overall, SAM did not require many points; however, the points that were correctly selected had a significant impact on the segmentation results.

Future improvements could focus on increasing the accuracy of foreground point selection, for example by sampling multiple points from different activation layers or regions. This could help reduce the influence of background artifacts. Additionally, introducing new hyperparameters or refining existing ones may further improve segmentation performance.

Acknowledgements

This work was completed as part of Homework 2 for the Computer Vision course at the University of Warsaw.

References

[1] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision (IJCV), 2017.