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. While effective, this requirement limits their applicability in fully automated systems. The Segment Anything Model (SAM) provides strong segmentation performance but still depends on external guidance.

At the same time, explainability techniques such as Grad-CAM offer spatial insight into convolutional neural network predictions by highlighting regions most responsible for a given class decision. Since these activations preserve coarse spatial structure, they can be repurposed as a signal for automatic prompt generation.

In this work, we investigate whether Grad-CAM can be used to generate point prompts automatically for SAM, enabling image segmentation without any user input. We implement Grad-CAM from scratch and design two prompt-generation pipelines based on the resulting heatmaps.

2 Materials and Methods

2.1 Dataset

We use a custom dataset derived from CIFAR-10 consisting of images containing exactly one geometric object: 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 and normalized to the range $[0, 1]$.

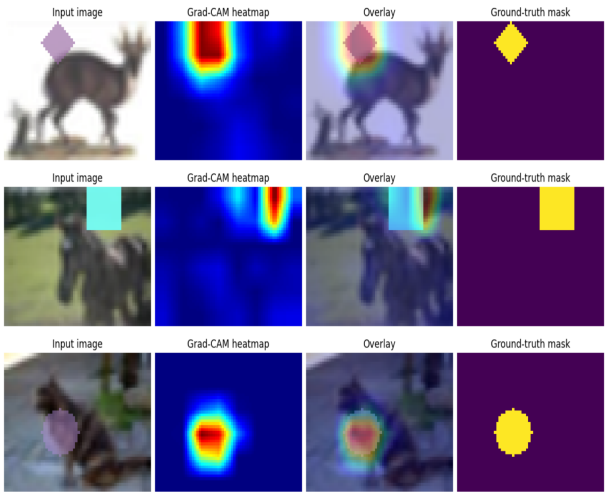

Figure 1: Grad-CAM results

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. To improve spatial coherence, we apply an inverse-distance weighted convolution kernel $K \in \mathbb{R}^{(2r+1)\times(2r+1)}$, defined as

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

where $\varepsilon > 0$ is a small constant preventing division by zero.

The transformed heatmap $\tilde{H}$ is obtained by convolving H with K and adding the result back to the original heatmap:

$$\tilde{H} = H + H * K, \qquad (2)$$

where $*$ denotes the two-dimensional convolution operator.

This transformation smooths and spatially expands salient regions of the heatmap, thereby improving the robustness of subsequent point selection.

2.4 SAM Pipelines

I observed that it is usually enough to hint him with one pixel, so my objective was, to find a way, to classify pixels correctly. For this purpose i used previuosly implemented Grad-Cam, with transformation, to prioritize the points that near a big number of highly heated points. Two automatic prompt-generation pipelines are proposed:

- **OnlyForegroundSamPipeline** selects the top-K pixels with the highest transformed Grad-CAM activation and labels them as foreground points. $K = 1$ was the best setting.
- **BackgroundAndForegroundSamPipeline** selects foreground points from high-activation regions, computes their center of mass, and samples background points sufficiently far from this center. Both foreground and background points are provided to SAM.

3 Results

The foreground-only pipeline achieves higher hit rate and IoU, indicating that selecting points from the most salient Grad-CAM regions is sufficient for accurate segmentation in 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 segmentation.

Table 1: Final dataset metrics for both SAM pipelines.

Pipeline	Hit Rate	IoU	Distance
OnlyForeground	0.773	0.769	9.290
Foreground + Background	0.733	0.725	9.690

Discussion

From above we can deduce two things. Grad-Cam works very well on ResNET, and shows us that the key pixels for model, are also key pixels for human in order to classify the image.

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. In: International Journal of Computer Vision (IJCV), 2017.