Towards Minimalist MaskFormer: Simplifying Transformer-Based Segmentation with GradCAM Supervision

Student ID: 2307—

University College London
ucakrps@ucl.ac.uk

Abstract. This paper investigates the performance of simplified versions of MaskFormer architectures under weak supervision, particularly leveraging pseudo labels derived from pretrained models like GradCAM. Using the Oxford-IIIT Pet Dataset, we explore the potential of lightweight transformer models in segmenting images with limited supervision, aiming to understand the trade-offs between model simplicity and segmentation accuracy.

1 Introduction

Semantic segmentation has long been one of the core challenges in computer vision, traditionally addressed using fully-supervised models trained on densely annotated data. State-of-the-art architectures such as SAM [4] and MaskFormer [3] have demonstrated exceptional accuracy, yet remain prohibitively resource-intensive due to their dependence on exhaustive pixel-level labels. This dependency creates a bottleneck for deploying these models in settings where annotation is costly or unavailable.

To address this limitation, weakly supervised segmentation has emerged as a practical compromise, where models learn from sparse supervision such as image-level labels, bounding boxes, or pseudo-masks [1,?]. In this work, I explore the extent to which the MaskFormer architecture can be simplified, both in terms of its transformer decoder depth and training labels, while maintaining satisfactory performance. The Oxford-IIIT Pet Dataset [5], with its relatively clean and focused class structure, provides an ideal testbed.

Open-ended Question (OEQ): Can a significantly simplified MaskFormer architecture, trained solely on GradCAM-derived pseudo-labels, still deliver meaningful segmentation results in a weakly supervised context?

This project explores a minimal configuration of MaskFormer trained without ground-truth segmentation masks. GradCAM [6] is used to generate weak spatial supervision, serving as a substitute for hand-labeled masks. The study contributes to the ongoing discussion of how explainability tools can double as supervision signals and how lightweight transformers might remain competitive under real-world constraints.

2 Methods

2.1 Dataset

The Oxford-IIIT Pet Dataset [5] contains over 3,600 annotated images of cats and dogs. For weak supervision, only image-level labels are used during training.

2.2 Model Architecture

We simplify the MaskFormer [3] by reducing the number of decoder layers to one and using ResNet-50 as the backbone. This architecture balances performance with reduced complexity, inspired by works like DeepLabv3+ [2] and FastFCN [7].

2.3 GradCAM Pseudo-supervision

GradCAM heatmaps [6] from the final ResNet layer are thresholded and converted to binary pseudo-masks [1]. These fixed masks serve as supervision throughout training.

2.4 Training

Implemented in PyTorch. Adam optimizer is used with a learning rate of 1 $times 10^{-4}$, batch size 16, and 50 epochs. The loss combines cross-entropy and Dice loss.

3 Experiments

3.1 Evaluation

We compare:

- Fully-supervised baseline (with GT masks)
- Weak supervision with GradCAM
- Simplified MaskFormer with 1-layer decoder (OEQ)

3.2 Metrics

Metrics: mIoU, pixel accuracy, Dice coefficient.

4 Results

5 Discussion

GradCAM pseudo-labels offer a reasonable approximation for weakly-supervised segmentation. Although the one-layer transformer loses some accuracy, it still performs competitively, showing potential for deployment in resource-constrained environments.

6 Conclusion

Simplifying MaskFormer while leveraging GradCAM can deliver meaningful segmentation without pixel-level masks. This work supports further exploration of hybrid supervision and transformer efficiency.

References

1. Ahn, J., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. CVPR (2019)
2. Chen, L.C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)

3. Cheng, B., et al.: Per-pixel classification is not all you need for semantic segmentation. arXiv preprint arXiv:2107.06278 (2021)
4. Kirillov, A., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
5. Parkhi, O.M., et al.: Cats and dogs. In: CVPR (2012)
6. Selvaraju, R.R., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
7. Wu, H., et al.: Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. arXiv preprint arXiv:1903.11816 (2019)