Leveraging Weak Labels for Fine-Grained Pet Image Segmentation with MaskFormer

Ronen Raj Roy (21137474)

No Institute Given

Abstract. In this study, we investigate the effectiveness of various pseudo-mask generation techniques for weakly supervised semantic segmentation using MaskFormer. We explore different sources of pseudo-labels, including Mask R-CNN, trained ResNet models, and GradCAM, to generate masks for the Oxford-IIIT Pet dataset. These pseudo-masks serve as weak supervision for training the MaskFormer model. We evaluate the performance of the model trained with these pseudo-masks against ground truth masks using metrics such as mean Intersection over Union (mIoU) and class accuracy. Our results demonstrate the impact of pseudo-mask quality on the MaskFormer's performance and highlight the potential of weakly supervised learning for image segmentation tasks when limited annotations are available. We analyze the strengths and weaknesses of different pseudo-label generation strategies, providing insights into the trade-offs between efficiency and accuracy in weakly supervised segmentation.

Keywords: Semantic Segmentation · Weak Supervision · Machine Learning.

1 Introduction

1.1 Background

Semantic image segmentation, the task of assigning a class label to each pixel in an image, is a fundamental problem in computer vision with diverse applications in medical imaging, autonomous driving, and scene understanding. Deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved remarkable success in this domain. However, these models heavily rely on large amounts of pixel-level annotations, which are often expensive and time-consuming to obtain. This limitation has spurred significant interest in weakly supervised learning approaches that can leverage weaker forms of supervision, such as image-level labels or pseudo-labels.

1.2 Literature

Recent advances in weakly supervised semantic segmentation have explored various techniques to mitigate the reliance on pixel-level annotations. These include methods based on Class Activation Maps (CAMs) [5], saliency maps [4],

and pseudo-label generation using pre-trained models [1]. MaskFormer [2] has emerged as a powerful framework for semantic segmentation, achieving state-of-the-art results with full supervision. However, its application to weakly supervised scenarios remains largely unexplored.

1.3 Motivation

This study is motivated by the need for efficient and accurate semantic segmentation methods that can be trained with limited supervision. We aim to investigate the potential of MaskFormer in weakly supervised settings by leveraging different pseudo-mask generation strategies. Our focus is on the Oxford-IIIT Pet dataset, which presents a challenging fine-grained segmentation task with a diverse set of categories, useful for both binary classification or per species classification. The dataset also provides ground truth masks which are excellent for evaluating model performance. We address the following research questions through two main studies:

1. **Implement a simplified MaskFormer architecture on a weakly-supervised setting, and perform an investigation to find out how this weakly supervised model compares to a fully supervised baseline**.(Minimum Required Project)
2. **To build on our first study, find out the simplest version of the architecture which can still produce satisfying results.** (Open Ended Question)

2 Methodology

This section describes the methodologies employed in our study, focusing on the weakly supervised algorithms, network architectures, and the novel approaches explored in the OEQ.

2.1 Generating Pseudo Masks

Our approach involves generating pseudo-masks to provide weak supervision for training the MaskFormer model. We look 3 different ways of doing it.

1. Mask-R-CNN : We utilize a pre-trained Mask R-CNN model [3] to generate pseudo-masks by applying it to the Oxford-IIIT Pet dataset images. We extract the predicted masks for the foreground (pet) class, threshold them at a suitable value, and use these as pseudo-labels for training. This is primarily used to answer our first study, as the pseudo masks generated were of the best quality.
2. Pretrained ResNet50 + GradCAM : To find a simpler way, we explore using image-labels to generate class activation maps using GradCAM and generate pseudo binary masks by thresholding the activations. The image labels are generated by using a pretrained ResNet50 model.

3. Pretrained Resnet50 + GradCAM finetuned on the Oxford IIIT-Pet Dataset - This is similar to the previously mentioned experiment, however for this experiment ResNet50 was finetuned on by training only the last layer on Oxford IIIT-Pet Dataset.

2.2 Network Architecture

MaskFormer: We adopt the MaskFormer architecture [2] as our primary segmentation model.There are several key simplifications made as described below:

- **Backbone:** Uses frozen pretrained ResNet50 instead of a full Feature Pyramid Network (FPN).
- **Pixel Decoder:** Employs sequential upsampling with four identical blocks (Conv2d → GroupNorm → ReLU → bilinear upsampling) rather than multi-scale feature fusion.
- **Transformer:** Utilizes standard PyTorch TransformerDecoder (3 layers, 8 heads) without custom attention mechanisms.
- **Prediction Heads:**
  - Class head: Simplified MLP with single hidden layer (768 neurons)
  - Mask head: Single 1×1 convolution instead of complex mask prediction network
- **Query Embeddings:** Basic embedding layer without specialized positional encoding.

3 Experiments

We shall now perform some experiments to test our Weakly Supervised MaskFormer model with pseudo masks generated from our previously described models.

Table 1. Model Performance Metrics

Metric	Train	Test
Class IoU	0.9698	0.9054
Mean Mask IoU	0.8525	0.7524
Mean Mask IoU w.r.t. Original Dataset	0.7415	0.7212
Mean Mask IoU of Pseudo-masks w.r.t. Original Dataset	0.8066	0.8074

3.1 Mask-R-CNN

We use a pretrained Mask-R-CNN model to generate masks on the Oxford IIIT-Pet Dataset and use these masks to train our MaskFormer implmentation. Note that the Mask-R-CNN model is not trained on the specific dataset, which qualifies the generated masks as weak supervision. We used IoU i.e Intersection Over

Table 2. Fully Supervised MaskFormer Performance Metrics

Metric	Train	Test
Class IoU	0.9535	0.8910
Mean Mask IoU	0.8372	0.7775
Mean Mask IoU w.r.t. Original Dataset	0.8372	0.7775
Mean Mask IoU of Pseudo-masks w.r.t. Original Dataset	1.0000	1.0000

Union as a metric to evaluate model performance. IoU signifies the percentage of pixels common between the masks generated by MaskFormer and the pixel level annotations provided by the dataset.

We achieve a mean Training IoU of 74.1% and a mean Test IoU of 72.1% w.r.t to the pixel level annotations , which falls in line with our expectations as the masks generated by Mask-R-CNN were of quite high quality and this ensures that our implementation of MaskFormer was able to learn effectively from them. Comparing this to our fully supervised models with a test IoU of 75.5% we can ascertain that our weakly supervised model was able to generate high quality masks, without the need of pixel level annotations.

3.2 Pretrained ResNet 50 + GradCAM

Table 3. Model Performance Metrics with ResNet-50 and GradCAM Pseudo-masks

Metric	Train	Test
Class IoU	0.9448	0.8915
Mean Mask IoU	0.4958	0.0257
Mean Mask IoU w.r.t. Original Dataset	0.0655	0.0375
Mean Mask IoU of Pseudo-masks w.r.t. Original Dataset	0.0900	0.0900

Even though we obtained a weakly supervised model with similar performance to the baseline, we experimented more to find out how simple we can make the pseudo mask generation process while retaining satisfactory performance. The following 2 experiments are aimed to answer our second study, **How simple can we make the pseudo mask generation process while ensuring satisfactory performance?**

We use a pre-trained ResNet-50 model (often pre-trained on ImageNet) to extract features from the input image. We were particularly in the activations of the final convolutional layer, as these activations capture high-level semantic information about the image.

GradCAM Grad-CAM is a technique used to visualize the regions of an image that are most important for a particular classification decision made by a CNN. It computes the gradients of the target class score (the score for the class we

are interested in) with respect to the activations of the final convolutional layer. These gradients are then globally averaged to produce a weight for each feature map in the final convolutional layer. These weights are used to linearly combine the feature maps, resulting in a coarse localization map. This map highlights the regions of the image that are most relevant to the target class.The localization map produced by Grad-CAM is then thresholded to create a binary pseudo-mask. Pixels above the threshold are considered to belong to the target class, while pixels below the threshold are assigned to the background.

These generated pseudo masks were used to train our simplified MaskFormer model and the results are defined in Table 3. Grad-CAM produces coarse localization maps that highlight the most discriminative regions for classification. However, these maps often lack the precise boundaries and detailed object shapes required for accurate segmentation. This coarseness leads to noisy and inaccurate pseudo-masks, hindering the MaskFormer's ability to learn precise segmentations.

3.3 Pretrained ResNet50 + GradCAM Finetuned On Oxford IIIT-Pet

Table 4. Model Performance Metrics with Trained ResNet-50 and GradCAM Pseudo-masks

Metric	Train	Test
Class IoU	0.9514	0.8850
Mean Mask IoU	0.9219	0.8156
Mean Mask IoU w.r.t. Original Dataset	0.5441	0.5544
Mean Mask IoU of Pseudo-masks w.r.t. Original Dataset	0.5441	0.5431

This experiment is similar to the previous experiment, however instead of using the pretrained ResNet50 as is, we finetuned it on our target dataset hoping to generate better masks. The results are present in Table 4. These results indicate a crucial observation : finetuning the model on the target dataset greatly improves the performance of our Weakly Supervised Model. These results can be further improved by applying postprocessing such as Conditional Random Fields (CRF) or rigorous hyperparameter tuning.

3.4 Using Binary Class Labels

We also performed these experiments using binary class labels instead of the 37 labels used for the earlier MaskFormer models. Instead of generating species specific segmentation, we considered an even simpler case, performing segmentation only with regards to if the pet was a cat or a dog.The IoU's for all 4 models are as follows.

- **MaskRCNN:**
  - Train Mean Mask IoU w.r.t. Original Dataset: 0.7359
  - Test Mean Mask IoU w.r.t. Original Dataset: 0.7177
- **Untrained ResNet:**
  - Train Mean Mask IoU w.r.t. Original Dataset: 0.0546
  - Test Mean Mask IoU w.r.t. Original Dataset: 0.0297
- **Finetuned ResNet:**
  - Train Mean Mask IoU w.r.t. Original Dataset: 0.4389
  - Test Mean Mask IoU w.r.t. Original Dataset: 0.4577
- **Fully supervised:**
  - Train Mean Mask IoU w.r.t. Original Dataset: 0.8404
  - Test Mean Mask IoU w.r.t. Original Dataset: 0.7842

As is clearly visible from the results, binary class labels caused a drop in the performance of our ResNet models because with binary labels, the ResNet models loses the opportunity to learn object-specific features that are crucial for accurate segmentation. For example, the model might learn to identify general pet-like features, but it might not be able to distinguish between a dog and a cat based on subtle differences in their shapes, textures, or poses. This lack of object-specific information can result in pseudo-masks that are less accurate and complete, particularly for objects that share similar visual characteristics.

4 Conclusion

In this study, we investigated the potential of weakly supervised semantic segmentation using MaskFormer on the Oxford-IIIT Pet dataset. We explored various pseudo-mask generation techniques to provide weak supervision and compared their impact on segmentation performance.We can now answer our study questions.

1. We have implemented a weakly supervised and Simplified MaskFormer model to work with Oxford IIIT-Pet dataset. After performing various experiments, we have determined that the masks generated by Mask-R-CNN, can be used as weakly supervised data to train our model to generate promising results.
2. From our other experiments we have also learnt that pseudo masks generated by applying GradCAM on a FineTuned ResNet50 are of a lower, yet still usable quality. Training our weakly supervised MaskFormer on this model yields satisfactory results which have a lot of room for improvement. This is especially important because fine-tuning ResNet50 and generating GradCAM masks takes only a fraction of time required for generating Mask-R-CNN which can lead to much greater efficiency.

4.1 Future Work

Future research could explore further refinement techniques for pseudo-masks, such as iterative refinement and incorporating contextual information. Additionally, GradCAM masks may be improved by applying post-processing and rigorous hyperparameter tuning.

References

1. Anh, P., Kwak, S.: Weakly supervised learning with context-aware pixel prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2436–2445 (2021)
2. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: Advances in Neural Information Processing Systems. vol. 34, pp. 17864–17875 (2021)
3. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
4. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
5. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)