Weakly Supervised One-Shot Segmentation

Hasnain Raza  
Karlsruhe Institute of Technology  
Karlsruhe, Germany  
hasnainraza@outlook.com

Mahdyar Ravanbakhsh  
University of Genova  
Genova, Italy  
mahdyar.ravan@ginevra.dibe.unige.it

Tassilo Klein  
SAP Machine Learning Research  
Berlin, Germany  
tassilo.klein@sap.com

Moin Nabi  
SAP Machine Learning Research  
Berlin, Germany  
m.nabi@sap.com

Abstract

One-shot learning is a challenging discipline of machine learning since it gnaws at the concept of learning from large amounts of data. This is akin to making machine learning algorithms generalize from a few examples, much like how humans learn. We explore another novel dimension to this problem, of using weak supervision (labels only) in the one-shot domain, and specifically analyse it in the context of semantic segmentation. This is a challenging problem since we operate in the scarcity of data and supervision. We present a simple yet effective approach, whereby exploiting information from the base training classes in the current one-shot segmentation set-up allows for weak supervision to be easily used. We show that this strategy can be leveraged to achieve nearly the same results as full supervision, but with no pixel annotations, allowing fully automated segmentation. Comparisons to several fully supervised methods show convincing results. As well as better results than a weakly supervised baseline. Also presented is a baseline for generalized segmentation under one-shot and weak supervision assumptions.

1. Introduction

Deep learning traditionally requires large amounts of data to learn features that can be generalized across inter, as well as intraclass variations in data. Segmentation itself can be seen as a harder problem compared to classification. It requires modelling spatial correlations and an explicit background class. Add onto this the constraint of weak supervision (class label only), limited data (only one example per class) and the difficulty of the problem increases dramatically. This is the exact set-up (weak supervision & one-shot) the paper aims to explore. The set-up makes sense in the real world, since segmenting objects from a single example reference image is a natural use case.

Substantial work has come out in the domain of few shot classification such as [10, 23, 22, 19, 7, 16] but most of these approaches remain limited to classification as their extension to image segmentation (general and semantic) [5, 11, 18] is non trivial. For weakly supervised segmentation, recent work [24] suggests using a pre-trained network to propose probable object regions in images, which are used to iteratively refine and produce segmentations. Since for the few shot setup the test classes are non-overlapping with the train classes, such an approach is inapplicable directly. However, in [8], it is proposed to learn the segmentation task by transferring knowledge from a source domain of fully annotated images into a target domain of class label only images. However, they do not make the one-shot assumption on the target domain, [14] learn segmentation on weak supervision of bounding boxes and class labels, whereas the assumptions of one-shot and class label only are not fulfilled. In [9], the goal is to learn to segment from large collections of data of box annotations and mask annotations by a weight transfer strategy. We have a much stricter constraint on the amount of data and supervision. In [15], the output of object and boundary detectors are used to learn to segment in open-set conditions. Whereas we have specific classes that need to be segmented.

Background. For one-shot segmentation, the meta-learning strategy has been used by [20] to regress parameters for a model to do segmentation. Other approaches build on the idea to extend to multiple classes by using prototypes [3], or guiding segmentation via similarity [25]. [13] use a siamese backbone to encode the scene and a reference object to do instance segmentation. In [12] they address the problem of segmentation under clutter by adopting the strategy of segment-first, classify later. [1] learn to segment objects
in videos, using a single fully annotated frame. [17] learn to extract a task representation with which segmentation is done, their model is what we extend for our weakly supervised benchmark since it allows for the fast merging of support features because of their late fusion strategy. While their approach allows for lower supervision in the form of sparse pixel annotations, there is still human annotation involvement, and the pipeline is not completely automated. We examine the task of one-shot segmentation, assuming pixel-wise labels are not available during testing. The idea is to extract features from known classes in the training data-set and use these features to construct a rough segmentation for a support image. This segmentation can then be used as a “guide” to select features corresponding to that class which in turn, can help segment a testing image.

Contributions. Our major contributions in this paper are: First, we introduce a novel low-cost extension for a fully supervised one-shot segmentation method to support weak (class label) supervision. Second, we propose an approach to exploit feature similarity between training and testing classes to guide segmentation in the weak supervision one-shot domain, showing competitive results to fully supervised one-shot methods. W e explore two problems, the first is one-shot segmentation (M), where a given class from a specific image (query image) is assumed to be large. Furthermore, if C \in M is a semantic class in M, then the set of semantic classes in D_{train} and D_{support} are disjoint, i.e., \{C_{train}\} \cap \{C_{support}\} = \emptyset.

2. Problem definition

In this section our proposed one-shot segmentation approaches are formalized, and the notations for each are introduced. We explore two problems, the first is one-shot segmentation, but under weak supervision at test time. This has been explored mostly in the existing setup proposed by [20], but we assume weak supervision during testing. The second approach is the generalized segmentation, concerning all the seen and unseen classes together. The objective of this task is to perform evaluation for seen and unseen classes at once, while making no assumptions on what classes the query image might contain. This provides insight into how much the model remembers to segment the base classes, and how well that information can be used to segment the new concepts.

2.1. One-shot, weakly supervised

This set-up builds on the problem defined in [20], and introduces weak supervision at test time. The task is to segment a given class from a specific image (query image), provided one example image and class label to learn the novel class (support set). This set-up is investigated under the following definitions:

- i) D_{train} = \{\{I_1, M_1\}, ..., \{I_{N_{train}}, M_{N_{train}}\}\} is the set of training images (I) and their corresponding ground truth segmentation (M), where N_{train} is assumed to be large.
- ii) D_{support} = \{\{I_1, C_1\}, ..., \{I_{N_{support}}, C_{N_{support}}\}\} images (I) and image level labels (C) constitute the support set.  It should be noted that N_{support} is relatively limited (only a single image-label pair is present per class).
- iii) D_{query} = \{\{I_1\}, ..., \{I_{N_{query}}\}\} contains images. The task is to obtain the set of segmentation masks \{M_1, ..., M_{N_{query}}\} for each image in D_{query} where the class to segment in each image is defined by the support.

Furthermore, if C \in M is a semantic class in M, then the set of semantic classes in D_{train} and D_{support} are disjoint, i.e., \{C_{train}\} \cap \{C_{support}\} = \emptyset.

2.2. Generalized segmentation

The one-shot, weakly supervised set-up investigates results on novel/support classes. The evaluation for generalized segmentation is done by picking query and support images containing that class. During the test phase, we evaluate how well the segmentation performance across the classes in D_{train} (base classes) is, and how well it is for D_{support} (novel classes). In other words, the evaluation is performed over novel and base classes without making assumptions on what classes the query image contains. For this case, D_{train}, D_{support} and D_{query} remain as dis-
cussed, but samples in $D_{query}$ need to segmented for all classes in the set $C_{query} = C_{train} \cup C_{support}$. Hence, $C_{support}$ always contains one image of each unseen class.

The explicit assumption that a support image $I_{support}$ can not contain more than one novel class label is also made. Also, $I_{support}$ can not contain more than one base and one novel class label at the same time. These are reasonable restrictions, since obtaining an image of a novel class only is not difficult. Also note that we use the words “base” to refer to the set of training/base classes, and “novel” to refer to the set of support/novel classes interchangeably.

3. Method

In this section, our proposed methodology for one-shot segmentation is described for the two setups defined in section 2. First, the standard fully supervised Co-FCN notation and methodology is reviewed, then our weakly supervised extension is introduced. Finally, the generalized segmentation methodology is explained.

3.1. Fully supervised Conditional FCN (Co-FCN)

We revise here the idea for the Co-FCN [17]. The Co-FCN method uses an FCN to encode a support image $I_s$ to produce a feature map $F_s \in \mathbb{R}^{h_s \times w_s \times d}$. The binarized segmentation masks for foreground/background (FG-BG) $\{M_{s_fg}, M_{s_bg}\}$ for $I_s$ is used to encode the class information in the support feature map, producing $F_{s_fg}$ and $F_{s_bg}$. These feature maps are pooled spatially, resulting in $F_{pooled} \in \mathbb{R}^{1 \times 1 \times d}$ for each of the FG-BG segmentation masks. These pooled maps are tiled and then concatenated depth-wise across the feature map $F_q \in \mathbb{R}^{h_q \times w_q \times d}$ of a query image to produce a resulting feature map $F_{\text{guided}} \in \mathbb{R}^{h_s \times w_s \times 3d}$. The network is trained to produce dense segmentation of the support class in the query image from $F_{\text{guided}}$. In this way, the encoded features from support can be seen as a guide $g$, while the segmentation network can be seen as a function $f_\theta$ parametrized by $\theta$.

**Learning base classes and testing.** The Co-FCN is trained on the standard training class splits of PASCAL 5' as defined by [20] and their extended annotations from SBD [6]. In the training phase for base classes, we select all the images that contain a training class for training, while mapping any occurrence of test class annotations to the background. At test time, a support set is sampled. From this support, features are extracted. These features are encoded using the segmentation mask. Then encoded features are used as a guide to segment a given test image (query), as proposed in [17].

3.2. Weakly Supervised Co-FCN (WS Co-FCN)

It has been shown that the features from deeper layers of a CNN generalize enough to perform other classification tasks [21, 2]. Each high dimensional feature vector can be viewed as a bag of features. Hence, for similar visual concepts, these latent representations are clustered together. Using this insight, one can extract feature vectors for training classes (as well as background) using $D_{train}$. Then use nearest neighbor classification to model resemblance of a novel concept, to a base concept (like a cat to a tiger). Using this approach, one can produce a segmentation proposal for a support image. This segmentation proposal can be used to produce the guide $g$ to segment a query image using $f_\theta$ in the co-FCN framework. A schematic representation of our proposed weakly supervised Co-FCN approach is illustrated in Fig. 1.

**Generalized one-shot, weakly supervised.** For this an FCN32s is trained on the split-wise $D_{train}$ from PASCAL 5' and their extended annotations from SBD [6], while removing all images containing the test classes. Then the cluster centres/prototypes are formed using the fc7 features, from the images in $D_{train}$.

**Training the novel classes.** During testing, fc7 features are extracted from the support images, and the feature space is updated by adding a new cluster centre/prototype for the novel classes. Then a nearest neighbor classifier is fitted on this new space. This new feature space is then used to segment each query image using the nearest neighbor classification. The idea is illustrated in Figure 2. The preservation of base class feature clusters, along with updates from novel class feature vectors allows the segmentation to be done in a generalized fashion.

4. Experiments

In order to evaluate our proposed weakly supervised one-shot segmentation approach, the challenging PASCAL VOC dataset [4] is chosen. The evaluation has been performed for both foreground/background (FG-BG) binary segmentation, and the generalized segmentation tasks. In this section first the evaluation dataset for our experiments is reviewed, then we describe the evaluation procedures and report the results for FG-BG segmentation (in Sec. 4.1), and generalized segmentation (in Sec. 4.2).

**Dataset.** The dataset used for the experiments is PASCAL VOC [4] with extended annotation from SBD [6]. We formed it into Pascal 5' following the proposed sequential splits by [20]. Pascal 5' is composed of 4 splits of the data, where for each split, 5 classes are reserved for testing, while 15 are used for training.

4.1. One-shot weakly supervised evaluation:

The weakly supervised nearest neighbor baseline is formed by training an FCN32s [11] on $D_{train}$ while mapping all occurrences of test annotations to the background. The features are extracted from the fc7 layer of the network, to do nearest neighbors classification for each support feature, by euclidean distance to the base class feature proto-
All that map to the background, are background, and everything else is assumed to be the support class.

**Evaluation setup.** Testing is performed using the standard testing process defined by [20], where we randomly sample a $I_{\text{query}}$ from PASCAL VOC validation set, and a support image $I_{\text{support}}$, containing a test class $C_{\text{test}}$. We then use the methodology in section 3 to obtain the dense segmentation $M_{\text{query}}$, for which we evaluate the IoU over the FG-BG binary segmentation task as defined in [17].

**Experimental results.** Table 1 lists the results of our proposed method. It can be seen that the proposed weakly supervised co-FCN achieves nearly the same results as the fully supervised case, and shows improvement over the strong FG-BG segmentation baseline. This shows that the conditioning features from the support set is still effective. This is because the co-FCN pools the coarse pseudo segmentations obtained by the nearest neighbor spatially. As long as most features belong to the support class, the guide $g$ remains informative about the class features. Another reason is based on the observation that nearest neighbor search forms a good baseline for one-shot segmentation, as shown by [20]. This is also supported by literature [2] [21]. We know deep features are discriminative for class, and the vectors can be assumed to be a bag of features. Hence in their latent space, the vectors belonging to the same class are clustered together, providing a powerful mechanism for classification via nearest neighbor search. One thing of interest here is that, if the image contains more than one class, we assume that any non-background object is the class object. This introduces false-positive features into the guide. Possibly the reason for the reduced score compared to the fully supervised case.

### Table 1. Mean IoU on the FG-BG segmentation task on held out classes for all splits suggested by [17]. All the previous results are reported from [17] and [3].

<table>
<thead>
<tr>
<th>Supervision</th>
<th>Method</th>
<th>IoU%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Supervised</td>
<td>FG-BG [17]</td>
<td>55.0</td>
</tr>
<tr>
<td></td>
<td>Fine-Tuning [17]</td>
<td>55.1</td>
</tr>
<tr>
<td></td>
<td>OSLSM [20]</td>
<td>61.3</td>
</tr>
<tr>
<td></td>
<td>Co-FCN [17]</td>
<td>60.1</td>
</tr>
<tr>
<td></td>
<td>PL+SEG [3]</td>
<td>61.2</td>
</tr>
<tr>
<td>Weakly Supervised</td>
<td>Nearest neighbor baseline</td>
<td>51.5</td>
</tr>
<tr>
<td></td>
<td>Ours (WS Co-FCN)</td>
<td><strong>58.7</strong></td>
</tr>
</tbody>
</table>

For generalized segmentation, $I_{\text{query}}$ is randomly sampled from the validation set of PASCAL VOC. A $D_{\text{support}}$ with 5 images (each containing one unique novel class) is then sampled. Then following the strategy explained in section 3, the segmentation mask of $M_{\text{query}}$ is obtained. The result is evaluated as the mean IoU over the 20 classes in PASCAL. This is repeated for 500 iterations. We report the IoU score for each split in PASCAL 5’s, further divided over the base (training) and the novel (test/support) classes.

**Generalized vs. non-generalized segmentation.** Tab. 3, shows results for one-shot segmentation results over the novel only test queries (OSL) but 5-way, since support contains all five novel classes, and our proposed generalized setup (G-OSL). According to Tab. 3, the final mean IoU between generalized and non-generalized setups are comparable. However, it can be observed that if $D_{\text{support}}$ is restricted to novel classes only (non-generalized), the scores on novel classes are higher compared to cases when base classes are allowed to be present in the query set (generalized). This is to be expected, since including the base classes in the query set increases the possibility of false positives and false negatives, both lowering the IoU. Also because the cluster centres are formed by the base features themselves, so naturally nearest neighbor classification becomes more biased towards base classes. Some qualitative results are shown in Figure 3. An FCN32s trained in a fully supervised manner on base classes is also shown for reference.

**Generalized segmentation testing benchmarks.** The proposed generalized segmentation is also evaluated over different input data for the novel training phase. Tab. 2, shows the results of generalized segmentation (G-OSL) under two benchmarks: i) Single label, when $D_{\text{support}}$ restricted to novel samples only, and ii) Multi label, when base classes
are allowed to be present in $D_{support}$. It can be seen for multi-label, the scores are slightly lower, indicating that mostly novel classes are affected. Since an image including both base and novel classes has a lower number of features to represent both. The lower number of features in the support sample may not affect the base classes too much, but for novel do not provide a well representative cluster centre.

5. Conclusion and Future Work

In this paper, we present an approach to exploit feature similarity to generate segmentation proposals in a weakly supervised fashion. Furthermore, we show this approach can simply be plugged into existing methods as a low-cost extension to enable weak supervision. Then, the proposed approach for feature similarity search for segmenting seen (base) and unseen (novel) classes in a generalized segmentation setup is analyzed. Our reported results show that in general, the nearest neighbor segmentation performs well on novel classes, as long as images do not contain any base classes. However, involving the base classes biases the predictions, weakening performance on the novel classes. As future work, we will focus on improving the overall performance and to narrow down the performance gap between the seen classes and unseen classes.

References


