ADAPTIVE MASKED WEIGHT IMPRINTING FOR FEW-SHOT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning has mainly thrived by training on large-scale datasets. However, for a continual learning agent it is critical to incrementally update its model in a sample efficient manner. Learning semantic segmentation from few labelled samples can be a significant step toward such goal. We propose a novel method that constructs the new class weights from few labelled samples in the support set without back-propagation, while updating the previously learned classes. Inspiring from the work on adaptive correlation filters, an adaptive masked imprinted weights method is designed. It utilizes a masked average pooling layer on the output embeddings and acts as a positive proxy for that class. Our proposed method is evaluated on PASCAL-5i dataset and outperforms the state of the art in the 5-shot semantic segmentation. Unlike previous methods, our proposed approach does not require a second branch to estimate parameters or prototypes, and enables the adaptation of previously learned weights. Our adaptation scheme is evaluated on our proposed incremental version of PASCAL and has shown to outperform the baseline model.

1 INTRODUCTION

Few-shot learning literature has mainly focused on image classification [Finn et al. (2017), Koch et al. (2015), Lin et al. (2017), Vinyals et al. (2016), Snell et al. (2017), Qi et al. (2017), Qiao et al. (2017)]. However, unlike image classification, semantic segmentation requires to learn a pixel-wise classification and can provide multiple classes in the support set. Thus, segmentation is more challenging in the low-shot regime than classification. One of the methods that are based on learning a parameter predictor was proposed by Shaban et al. (2017). A conditional network method was proposed by Rakelly et al. (2018) based on sparse or dense labels to guide the segmentation network. Another method that inspired from prototypical networks was proposed by Dong & Xing (2018). The previous methods require the training of an additional branch to act as a prototype learner or a parameter prediction branch. They can only operate in a static setting where one support set is provided to the model. If a new support set is provided to the model that has annotations for both novel classes and previously learned classes from few data, there is no direct extension to adapt their model.

In this paper we propose an adaptive masked weight imprinting scheme for few-shot semantic segmentation. Our main inspiration is from classical approaches in learning adaptive correlation filters [Bolme et al. (2010), Henriques et al. (2015)]. Correlation filters date to 1980s by Hester & Casasent (1980) that proposed learning an averaged matched spatial filter constructed as a weighted linear combination of basis functions. Bolme et al. (2010) proposed a fast object tracking method based on adaptive correlation filters, where the filters are updated using a running average. Our method proposes a novel scheme to compute convolutional filters to match the objects through masked weight imprinting, while adapting the learned ones. Weight imprinting Qi et al. (2017) has been proposed for image classification and relates metric learning methods to softmax classification. It utilizes the normalized embeddings for the support set as proxies and concatenate it to the original weight matrix in the last classification layer. Since 1x1 convolution, that is typically used in segmentation networks, is equivalent to fully connected layers its filters can be imprinted as well. Nonetheless using the output embeddings directly can incorporate undesired features from other classes. Thus, a masked average pooling layer is utilized on the output embeddings with the segmentation label provided in the support set.
Figure 1: Adaptive Masked weight Imprinting using the Masked Average Pooling Layer. For simplicity it shows the imprinting on the final layer solely. Nonetheless, our scheme is applied on multiple resolution levels to improve the segmentation accuracy.

Unlike previous methods, our approach can easily operate with any pretrained network without the need to train a second branch. The contributions of this paper are: (1) we propose a masked weight imprinting scheme that is performed on multiple resolution levels. (2) we propose a novel adaptive weight imprinting scheme that inspires from adaptive correlation filters, in order to update the weights of previously learned classes. (3) Our method outperforms the state of the art on the 5-shot case on PASCAL-5i, and the adaptation method outperforms the baseline method. (4) We propose iPASCAL which is the incremental version of PASCAL-VOC to evaluate the continual learning mode for segmentation.

2 PROPOSED METHOD

We formulate a problem similar to Shaban et al. (2017) for the few-shot setting. However, for the continuous object segmentation setup we propose a novel setup using incremental PASCAL which we term as iPASCAL. The PASCAL VOC dataset Everingham et al. (2015) classes are split into $L_{\text{train}}$ and $L_{\text{incremental}}$ with 10 classes each, where $L_{\text{train}} \cap L_{\text{incremental}} = \emptyset$. The classes belonging to the $L_{\text{train}}$ are used to construct the training dataset $D_{\text{train}}$ and pre-train the segmentation network.

Unlike the static setting in the few shot case, the continuous segmentation mode provides the image-label pairs incrementally with different encountered tasks. Each task introduces two novel classes to learn. The tasks are in the form of triplets $(t_i, (X_i, Y_i))$, where $(X_i, Y_i)$ represent the overall batch of images and labels from task $t_i$. The batch labels are for the two novel classes belonging to task $t_i$, and the previously learned classes in the encountered tasks $t_0, ..., t_{i-1}$. In each task the model encounters each image-label pair from the current batch only once.

The backbone architecture used in our segmentation network is a VGG-16 Simonyan & Zisserman (2014) that is pre-trained on ImageNet Deng et al. (2009). Similar to FCN8s architecture Long et al. (2015) skip connections are used to benefit from higher resolution feature maps, and a 1x1 convolution layers are used to map from the feature space to the label space. Inspiring from the work in few-shot image classification with imprinted weights Qi et al. (2017) we propose to utilize a masked weight imprinting scheme. The weight imprinting method is based on the relation between metric learning methods and softmax classification.

Similar to Qi et al. (2017) we utilize the embeddings for the few labelled samples from the novel class as proxies. These proxies can be fused directly as weights in the final fully connected classification layer. There exist some major differences between the classification setting and the segmentation setting: (1) only convolutional layers are utilized in semantic segmentation. (2) the support set provides additional information not only to the novel class but it can include updated information about older classes as well. (3) The output embeddings are 3D tensors unlike in classification where the output embedding vector can be used directly. (4) Multi-resolution support is necessary to ensure the segmentation accuracy.
Table 1: Quantitative results for 1-way 1-shot segmentation on PASCAL-5i dataset. FT: denotes Fine-tuning. OSLSM method by Shaban et al. (2017). Baseline methods evaluation are reported from Shaban et al. (2017).

<table>
<thead>
<tr>
<th></th>
<th>1-NN</th>
<th>Siamese</th>
<th>FT</th>
<th>OSLSM</th>
<th>ours (FCN8s)</th>
<th>ours (DFCN8s)</th>
<th>ours (Red-DFCN8s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>25.3</td>
<td>28.1</td>
<td>24.9</td>
<td>33.6</td>
<td>33.4</td>
<td>36.0</td>
<td><strong>39.2</strong></td>
</tr>
<tr>
<td>Fold 1</td>
<td>44.9</td>
<td>39.9</td>
<td>38.8</td>
<td><strong>55.3</strong></td>
<td>46.8</td>
<td>48.0</td>
<td>48.0</td>
</tr>
<tr>
<td>Fold 2</td>
<td><strong>41.7</strong></td>
<td>31.8</td>
<td>36.5</td>
<td>40.9</td>
<td>38.7</td>
<td>39.3</td>
<td>39.3</td>
</tr>
<tr>
<td>Fold 3</td>
<td>18.4</td>
<td>25.8</td>
<td>30.1</td>
<td>33.5</td>
<td>33.2</td>
<td>33.8</td>
<td><strong>34.2</strong></td>
</tr>
<tr>
<td>Mean</td>
<td>32.6</td>
<td>31.4</td>
<td>32.6</td>
<td><strong>40.8</strong></td>
<td>38.0</td>
<td>39.3</td>
<td><strong>40.2</strong></td>
</tr>
</tbody>
</table>

It is well known that 1x1 convolution are equivalent to fully connected layers so the same concept can be utilized in fully convolutional networks. In this case the weight filter is used as a proxy that convolves the output feature map and computes the extent to which the different parts in the feature map matches these proxies. In order to incorporate the pixels that belong mainly to the novel class, masked feature maps with the binary labels provided in the support set are used. This is followed by average pooling the masked feature maps per channel as in equation equation 1, we denote this layer as masked average pooling [Zhang et al. (2018)].

\[ p_l = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{N} \sum_{x \in X} F^i(x)Y_l^i(x) \]

(1)

Where \( Y_l^i \) is a binary mask for \( i^{th} \) image with the novel class \( l \), \( F^i \) is the corresponding output feature maps for \( i^{th} \) image. \( X \) is the set of all possible spatial locations and \( N \) is the number of pixels that are labelled as foreground for class \( l \). The output from the masked average pooling layer \( p_l \) can be further used as proxies representing class \( l \). In the case of a novel class the imprinted weights can be utilized directly as the weight filter representing that new class. An average of all the masked pooling features for the k-shot samples provided in the support set is used.

In case of the older classes, the convolutional layer weights in our model can be updated with the newly imprinted weights for that class in an adaptive scheme. A running average is used to update the weights following equation equation 2 for older classes with the update rate \( \alpha \). Figure 1 shows the adaptive masked weight imprinting scheme. Masked weight imprinting is performed on the multiple resolution levels in order to improve the segmentation accuracy.

\[ W_{new} = (1 - \alpha)W_{old} + \alpha W_{imprinted} \]

(2)

3 EXPERIMENTAL RESULTS

The setup for pretraining the models to be tested on PASCAL-5i is detailed. The base network is trained using RMSProp [Hinton] with learning rate \( 10^{-6} \), and L2 regularization with a factor of \( 5\times10^{-4} \) on the 15 classes outside of the current fold. In the few-shot evaluation 1000 samples are used similar to OSLSM setup [Shaban et al. (2017)]. The alpha parameter used for adapting the previously learned weights is 0.5. Table 1 and Table 2 show the results for the 1-shot and 5-shot segmentation respectively on PASCAL-5i using mIoU of the foreground class. Our method is compared to OSLSM [Shaban et al. (2017)] and the baseline methods for few-shot segmentation. It shows that our method outperforms the baseline fine-tuning method by 7.6% in terms of mIoU, without the need for extra back-propagation iterations through directly using the imprinted weights. Our method performs on par with OSLSM [Shaban et al. (2017)] method in the 1-shot in terms of mean across folds, while it outperforms OSLSM in the 5-shot case. However, unlike OSLSM our method does not need to train an extra branch for predicting the parameters. Figure 3 shows the qualitative results on PASCAL-5i which shows both the support set image-label pair, and our predicted segmentation for the query image.

We conducted further experiments on iPASCAL, where triplets for the task, the corresponding images and semantic labels are provided. Semantic labels include the new classes in the current and previous
Figure 2: 1-way n-shot evaluation using the setup proposed in iPASCAL. Naive #1 denotes fine-tuning with 1 iteration per sample, while #10 uses 10 iterations. The Imprint method is utilized with different alpha parameters 0.05, 0.2, 0.5, 0.9.

Table 2: Quantitative results for 1-way 5-shot segmentation on PASCAL-5i dataset. OSLSM method by [Shaban et al., 2017]. LogReg baseline reported from [Shaban et al., 2017]

<table>
<thead>
<tr>
<th></th>
<th>LogReg</th>
<th>OSLSM</th>
<th>ours (FCN8s)</th>
<th>ours (DFCN8s)</th>
<th>ours (Red-DFCN8s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 0</td>
<td>35.9</td>
<td>35.9</td>
<td>37.4</td>
<td>40.5</td>
<td>45.3</td>
</tr>
<tr>
<td>Fold 1</td>
<td>51.6</td>
<td><strong>58.1</strong></td>
<td>50.9</td>
<td>52.5</td>
<td>51.4</td>
</tr>
<tr>
<td>Fold 2</td>
<td>44.5</td>
<td>42.7</td>
<td>44.0</td>
<td>44.8</td>
<td><strong>44.9</strong></td>
</tr>
<tr>
<td>Fold 3</td>
<td>25.6</td>
<td>39.1</td>
<td>39.1</td>
<td>39.9</td>
<td>39.5</td>
</tr>
<tr>
<td>Mean</td>
<td>39.3</td>
<td><strong>43.9</strong></td>
<td>42.9</td>
<td>44.4</td>
<td><strong>45.3</strong></td>
</tr>
</tbody>
</table>

Figure 3: Qualitative evaluation on PASCAL-5i. The support set and our proposed method prediction on the query image are shown in pairs for the 1-way 1-shot setting.

encountered tasks. Figure 2 shows the comparison between naive fine-tuning from random weights against our proposed adaptive masked weight imprinting without any fine-tuning operations in terms of mIoU. It shows that masked imprinting provides better mIoU in comparison to fine-tuning that will lead to over-fitting. Fine-tuning was conducted using RMSProp with learning rate $10^{-10}$. Fine-tuning is performed only to the last layers responsible for pixel-wise classification, while the feature extraction weights for VGG16 are fixed. It is worth noting that the current evaluation setting is a $n$-way 1-shot, where $n$ increases with 2 additional classes with each encountered task resulting in 10-way 1-shot evaluation in the last task.

4 CONCLUSION

In this paper we proposed a novel approach for few-shot semantic segmentation using an adaptive masked imprinting scheme. Our proposed method outperforms the state of the art few-shot segmentation methods in the 5-shot setting, while it alleviates the need for training a second branch as the previous literature. We proposed a novel setup iPASCAL to evaluate the effectiveness of our adaptation method.
REFERENCES


