Few-shot Learning with Weakly-supervised Object Localization

Jinfu Lin\textsuperscript{1}, Xiaojian He\textsuperscript{1}
\textsuperscript{1}South China University of Technology
\{Jinfu Lin\}kingfoulin@foxmail.com

Abstract

Few-shot learning (FSL) aims to learn novel visual categories from very few samples, which is a challenging problem in real-world applications. Many data generation methods have improved the performance of FSL models, but require lots of annotated images to train a specialized network (e.g., GAN) dedicated to hallucinate new samples. We argue that localization is a more efficient approach because it provides the most discriminative regions without using extra samples. In this paper, we propose a novel method to address the FSL task by achieving weakly-supervised object localization within performing few-shot classification. To this end, we design (i) a triplet-input module to obtain the initial object seeds and (ii) an Image-To-Class-Distance (ITCD) based localizer to activate the deep descriptors of the key objects, thus obtaining the more discriminative representations used to perform few-shot classification. Extensive experiments show our method outperforms the state-of-the-art methods on benchmark datasets under various settings. Besides, our method achieves superior performance over previous methods when training the model on miniImageNet and evaluating it on the different datasets (e.g., Stanford Dogs), demonstrating its superior generalization capacity. Extra visualization shows the proposed method can localize the key objects accurately.

1 Introduction

Deep Convolutional Neural Networks (ConvNets) has achieved excellent performance in numerous computer vision tasks in recent years. Trained with a large amount of annotated data, the ConvNets can extract robust and effective representations for classification. However, ConvNets suffers from its weak generalization ability and poor performance when the annotated samples for training are very limited. In contrast, we humans are able to identify novel classes with only a single or few samples. Thus, recognizing novel categories from very few samples is an important and significant problem, which is often termed Few-shot learning (FSL). Recently, the FSL problem has attracted increasing attention, and a number of methods have been proposed to tackle this task. Fine-tuning the pre-trained model on the novel datasets is a common and simple method. Besides, to generalize the model to novel datasets, an emerging direction is to apply the meta-learning paradigm on few-shot learning. Meta-learning trains an across-task meta-learner which can accumulate transferable knowledge in one task and generalize to other novel tasks quickly. Another common approach is based on metric-learning, which learns an informative similarity metric between the query and the support samples, thus performing few-shot classification.

However, due to the very limited samples, most of the approaches encounter over-fitting and poor generalization. Thus, many data augmentation (DA) methods have been proposed. [Wang \textit{et al.}, 2018b] and [Hariharan and Girshick, 2017] are both data generation based method, which can generate additional examples for data-starved classes. However, they need a large of extra annotated data to train such a specialized data generator or hallucinator. Recently, [Alfassy \textit{et al.}, 2019] performs different set operations (e.g., union and intersection) on the label set of multi-label samples in the embedded space to fulfill DA. [Schwartz \textit{et al.}, 2019] leverages...
extra multiple semantic and feature fusion to train a more robust embedded module. Nevertheless, these methods contain additional combination networks and complicated feature fusion networks, leading to a more difficult training phase.

Different from DA, localization can distinguish the most discriminative regions from distractors without using extra annotated samples. Existing studies [Zhou et al., 2016; Oquab et al., 2015] show there is a built-in attention mechanism in classification networks (e.g., VGG16) to localize the objects of interest. Such a learned localization has positive feedback effects on classification but requiring abundant expensive annotated data to begin bootstrapping. Inspired by this, we argue that guiding to localize the key object during performing few-shot classification should make a significant improvement of the FSL model. However, due to very limited samples and with only image-level labels, achieving such a localization mechanism would be a very challenging problem for few-shot classification.

To bridge this gap, we propose a novel end-to-end network to achieve weakly-supervised object localization to address the FSL task. The proposed model can focus on the key objects to obtain the most discriminate features to perform few-shot classification. To this end, we utilize the Grad-CAM [Selvaraju et al., 2017] to generate the initial object seeds, which is a common initial step in weakly-supervised object localization [Zhang et al., 2018] and weakly-supervised semantic segmentation [Wang et al., 2018a]. Initial object seeds are the coarse regions that may include the key objects. For obtaining the more accurate localization of the key object, we use a set of deep descriptors to represent the embedded feature. Besides, we adopt the Naive Bayes (NB) assumption that each deep descriptor is independent, thus guaranteeing the translation invariance of the key object. Compared with using a vector obtained by flattening embedded features to represent input, a set of deep descriptors can provide spatial information, which is a key point of our method. Secondly, inspired by NBNN [Boiman et al., 2008], this paper designs the ITCD based localizer to activate each deep descriptor of the key object by its nearest neighbor. More concretely, the localizer achieves localization by correctly performing classification for each deep descriptor of the key object. Without additional supervisory loss, the localizer is trained only useful for classification. Thus, the localizer serves as the classifier at the same time, and achieve the localization within performing few-shot classification. The proposed model is a positive feedback network that the localization and the few-shot classification will bootstrap from each other.

Below, we list our main contributions: (1) We propose a novel end-to-end model that can localize the most discriminative region for the FSL problem. To the best of our knowledge, we are the first to achieve the weakly-supervised localization within few-shot classification. (2) We design the ITCD based localizer/classifier to activate each deep descriptor of the key region without extra supervised loss, helping the localization and the classification bootstrap from each other. (3) Extensive experiments on several benchmark datasets and the generalization evaluation experiment all show the superiorities of our method. Meanwhile, the visualization shows the proposed method can localize the key objects accurately.

2 Related Work

2.1 Meta-learning and Metric-learning

Meta-learning based method trains an across-task meta-learner with the meta-learning paradigm. MAML [Finn et al., 2017] trained a model agnostic meta-learner and found the initial parameters adapting to a variety of tasks with similar distribution, such that the model can quickly generalize to the new tasks. Meta-Learning LSTM [Ravi and Larochelle, 2017] proposed a model based on LSTM to learn an optimization method as well as the general initialization of the classifier. Metric-learning tackles the FSL by learning an embedding space where the input of the samples of the same categories is closer than those of different categories. Combining with attention mechanism, Matching Network [Vinyals et al., 2016] used the cosine distance to train a k-nearest neighbor classifier on the learned embedding space. [Snell et al., 2017] proposed a prototypical network to learn a prototype representation of each category and performed classification by the Euclidean distance between the query and prototype on embedding space. Relation network [Zhang et al., 2018] learned a nonlinear comparator to compare the distance metric between the query and the support images. Different from these metric-learning approaches that directly using the vector obtained by flattening the embedded feature, we use a set of deep descriptors to represent the embedded feature.

2.2 Object Localization

Many works [Wei et al., 2017; Fu et al., 2017] have shown that utilizing localization can help to learn the more discriminative embedded feature for classification. In this context, Wei et al., 2017 presented the Selective Convolutional Descriptor Aggregation method to achieve unsupervised localization in fine-grained datasets. For utilizing localization to address FSL, SalNet [Zhang et al., 2019] employed a readily available saliency detector to obtain the foreground and background, thus fulfilling DA by combining them differently. The ablation study of SalNet demonstrated that performance improvements stem from the localization ability of the saliency detector. Similarly, Sun et al., 2019 used the class attention map (CAM) generated by classification networks to locate the key region and fused with other different scale features for fine-grained few-shot classification. Different from applying readily available localization, Wertheimer and Har-tharan, 2019 used the assistant bounding box annotations to achieve the localization within the few-shot classification, thus to address the FSL task over heavy-tailed datasets. These methods can improve few-shot classification performance by leveraging localization. However, in real-world applications, bounding box annotations and readily available localization (e.g., saliency map) may be hard to meet or impracticable. Therefore, we propose a novel method to achieve weakly-supervised object localization, thus addressing the FSL problem with realistic settings.

3 The Proposed Model

3.1 Problem Definition

There is the train dataset $D$, support dataset $S$, and the query dataset $Q$ in the FSL task. $D = \{(x_i, y_i)\}_{i=1}^{N}$ contains $N$
samples, where \( y_i \) is the label of image \( x_i \). The support set \( S = \{(x_j, y_j)\}_{j=1}^M \) (\( M = C \times K \)) includes \( M \) examples in test phase and there is \( K \) labeled samples for each of \( C \) novel categories (C-way K-shot problem). The query dataset \( Q = \{(x_j, y_j)\}_{j=1}^N \) shares the same label space with \( S \). Their relationship is denoted as \((S \cup Q) \cap D = \emptyset\). FSL aims to train a model from \( D \), then classify the novel samples from \( Q \) based on the \( S \) during the testing phase. To mimic the few-shot learning task, the episodic training mechanism is adopted to train the model. Episode is a mini-batch includes \( D_{\text{support}} \) and \( D_{\text{query}} \), where we randomly sample \( C \) categories from \( D_{\text{train}} \) and for each category of \( C \) categories, its labeled samples are randomly split into subset \( D_{\text{support}} \) with \( K \) samples and subset \( D_{\text{query}} \) with the rest samples. Through this training mechanism, the model can learn transferable knowledge.

3.2 Model

Deep descriptors For an image \( X \), the activation of a convolution layer can be formatted as an order-3 tensor denoted as \( E(X) \in \mathbb{R}^{d \times w \times h} \). On the one hand, \( E(X) \) includes \( d \) feature maps with the size of \( w \times h \) and is denoted as \( M = \{M_n\}_{n=1, 2, 3, \ldots, d} \), \( M_n \) also known as the feature map in \( n \)th channel. On the other hand, \( E(X) \) can be considered as including \( m = (w \times h) \) deep descriptors and each deep descriptor is a \( d \)-dimension vector. We denote it as:

\[
D = \{d(i,1), d(i,2), \ldots, d(i,j), \ldots, d(i,w,h)\} = \{d_1, d_2, \ldots, d_m\}
\]

where \((i, j)\) is the position of the descriptor and \( d(i,j) \in \mathbb{R}^d \). Thus, a set of deep descriptors is the representation containing spatial information.

Triplet-input Module This paper adopts the Grad-CAM method to obtain the initial object seeds for weakly-supervised localization. Simultaneously, to maximally mine the information of the original images, we further design the Triplet-input module. Figure 3 shows the special implementation. Firstly, we obtain the focus-area mask by binarizing with the threshold of the mean of the Grad-CAM maps. Then use the focus-area mask to generate the foreground (initial object seeds) and background for forming the triplet input.

\[
MASK_i = \text{Grad} - \text{CAM} (I_i)
\]
\[
F_i = I_i \ast MASK_i
\]
\[
B_i = I_i \ast (1 - MASK_i)
\]
\[
T_i = \{I_i, F_i, B_i\}
\]

Embedding Module We adopt the shallow convolutional neural network containing only four convolutional blocks as the embedding module to extract embedded features. More concretely, there are a convolutional layer with 64 filters of size \( 3 \times 3 \), a batch normalization layer, and Leaky ReLU nonlinear layer in each convolutional block. Besides, behind each of the first two convolutional blocks, there is an extra \( 2 \times 2 \) max-pooling layer to reduced spatial resolution. We also adopt the ResNet backbone as the embedding module to obtain more detailed and richer features from images.

Naive Aggregation Module We aggregate the triplet output \( E_i = \{f(I_i), f(F_i), f(B_i)\} \) to make fully use of the information of foreground, background and original image. To pay more attention to the key regions of the images while suppressing background noise, the naive aggregation method is to enhance the foreground embedded feature while restraining the background embedded feature.

\[
A_i = f(I_i) + f(F_i) - f(B_i)
\] (6)

ITCD Localizer/Classifier Module In our model, we chose the cosine distance to measure the similarity between two descriptors \( d_i \) and \( d_j \):

\[
cos(d_i, d_j) = \frac{d_i^T d_j}{\|d_i\| \|d_j\|} \in [-1, 1], \quad (d_i, d_j \in \mathbb{R}^d)
\] (7)

For the query image of class \( k \), its embedded features is denoted as: \( q_k = \{d_1, d_2, \ldots, d_m\} \in \mathbb{R}^{d \times m} \). Similarly, the deep descriptors of all support embedded features of class \( k \) are denoted as: \( s_k = \{d_{1\prime}, d_{2\prime}, \ldots, d_{K\times m}\} \in \mathbb{R}^{d \times \lambda} \). In order to correctly classify the query image, the model need to guarantee the distance between \( q_k \) and its support embedded feature set \( s_k \) to be highest (nearest), which means the localizer will guarantee each deep descriptor of the query image can be accurately activate by its nearest neighbor deep descriptor (\( NN(d_{i\prime}) \)) from support set. Thus, the final distance between the embedded feature of query image and the embedded feature of the support category \( k \) is:

\[
D_{\text{ITC}} (q_k, s_k) = \sum_{i=1}^{n} \|d_i - NN(d_{i\prime})\| = \sum_{i=1}^{n} NN_\text{cos} (d_i, d_{i\prime})
\] (8)

where \( NN_\text{cos} (d_i, d_{i\prime}) \) represents \( \hat{d}_i \) is the nearest neighbor descriptor of \( d_i \) among \( \{d_{1\prime}, d_{2\prime}, \ldots, d_{K\times m}\} \) over cosine distance. Since the deep descriptors from the key regions of the query will be mostly activated by the deep descriptors from the regions belonging to the same class, thus the localizer can fulfill the weakly-supervised object localization with the only image-level labels. Therefore, we directly perform classification by the class of its nearest neighbor descriptor.

\[
p_k = p(y = k|x) = \frac{\exp (D_{\text{ITC}} (q_k, s_k))}{\sum_{k' \in C} \exp (D_{\text{ITC}} (q_k, s_{k'}))}
\] (9)

For \( N \) query images \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_N, y_N)\} \) in each episode, we minimize the loss \( L_e \):

\[
L_e = \sum_{i=1}^{N} - \log p(y = y_i|x_i)
\] (10)

4 Experiments

4.1 dataset miniImageNet As the mini-version of ImageNet, it contains 100 classes with 600 color images per class. The splits proposed in Matching Net is becoming the standard splitting rule of miniImageNet, where 64 categories are for training, 16 categories for validation, and 20 categories for testing. In this work, we adopt this split to compare our approach with state-of-the-art methods.
Fine-Grained Datasets In this paper, we pick three fine-grained datasets, e.g., Stanford Dogs [Khosla et al., 2011], Stanford Cars [Makadia and Yumer, 2014] and CUB-200 [Wah et al., 2011] to conduct the fine-grained few-shot learning task. Stanford Dogs contains 120 categories with 20,580 color images, where 70, 20, and 30 categories are used for training, validation, and testing, respectively. Stanford Cars containing 16,185 color images of 196 classes of cars, which is divided into 130, 17, and 49 categories for training, validation, and testing, respectively. CUB-200 is an image dataset with 6033 color images of 200 bird species for the FGVC task. Similarly, we split it into 130, 20 and 50 categories for training, validation, and testing, respectively.

4.2 Settings and Experiments

Settings We adopt the episodic training mechanism to make the training phase more faithful to the testing phase. For each training episode, besides the $K$ support images in each class, 5-way 1-shot contains 15 query images while 5-way 5-shot contains 10 query images for each of the $C$ randomly sampled categories. To be specific, for the 5-way 1-shot task, there 5 support images and 15 query images per class, thus that each episode contains $5 \times 1 = 5$ support images and $15 \times 5 = 75$ query images totally. Similarly, for the 5-way 5-shot task, there are $5 \times 5 = 25$ support images and $10 \times 5 = 50$ query images totally. In addition, we resize all the input images to $84 \times 84$. During the training phase, we randomly sample 300,000 episodes and select Adam as the optimizer with an initial learning ratio $5 \times 10^{-2}$ which will be reduced by half for every 100,000 episodes to train our model. During the testing phase, we also randomly sample 600 episodes from the test set to evaluate our model. We adopt the mean accuracy with 95% corresponding confidence interval as the performance indicator. It is worth mentioning that all our model is trained from scratch in an end-to-end manner, without any finetuning in the test phase.

Few-shot Classification on minImageNet We report the experiment results in table 1. When adopting the Resnet as the embedding module, our model can achieve state-of-the-art results both in the 5-way 1-shot and 5-shot task, especially in the 5-shot task (3.95 % higher than the 74.44% reported by DN4 [Li et al., 2019a]). Besides, when using the Conv as embedding module, our model also achieves the highest accuracy on the 5-way 5-shot task, gaining the 5.57%, 2.2%, and 0.41% over CovaMNet [Li et al., 2019b], DN4, and Dynamic-Net [Gidaris and Komodakis, 2018]. We also obtain very competitive accuracy on 5-way 1-shot task with Conv embedding module, gaining 3.8%, 2.11%, 1.76% improvement over R2D2 [Bertinetto et al., 2019a], CovaMNet, and DN4. As for Dynamic-Net and SalNet on the 5-way 1-shot task, they perform very complicated training steps to obtain state-of-the-art results. The former utilizes a two-stage model and needs to pre-train the model while our approach does not. The latter utilizes the state-of-the-art saliency detection model to generate the saliency map, thus to directly locate the key object. On the contrary, our approach achieves weakly-supervised localization with only image-level. Our approach is more simple but efficient and outperforms over state-of-the-art methods both on 5-way 1-shot and 5-shot.

Few-shot Classification on fine-grained datasets Compared with the generic few-shot classification task, its more challenging to perform fine-grained few-shot (FGFS) classification due to the smaller inter-class and larger intra-class variations of the fine-grained datasets. However, since FGFS receives very little attention, most of the exiting few-shot learning methods do not report their performance on such fine-grained datasets. Therefore, we implement and evaluate our approach on fine-grained datasets. Meanwhile, we also report the DN4, CovaMNet, GNN [Satorras and Estrach, 2018], Proto Net, and the baseline K-NN to make a compar-
eralization performance of the few-shot learning models, we conduct the evaluation on five different datasets, the seed from Stanford Cars is more coarse, leading to mistake guidance for the shallow network: Conv. CAM can not generate very well initial object seeds over the Conv module which is a shallow network. (ii) the Grad-CAM can not generate very well initial object seeds over Stanford Cars. As Figure 4 shown, compare with the other two datasets, the seed from Stanford Cars is more coarse, leading to mistake guidance for the shallow network: Conv.

Table 1: The mean accuracies of the 5-way 1-shot and 5-shot tasks on the miniImageNet dataset, with 95% confidence intervals.

| Model          | Embedding    | 5-Way Accuracy (%) | 1-shot | 5-shot |
|----------------|--------------|---------------------|--------|--------|
| Proto Net      | Resnet       | 51.15±0.85          | 69.02±0.75 |
|                | Conv         | 49.42±0.78          | 68.20±0.66 |
| Relation Net   | Resnet       | 52.13±0.82          | 64.72±0.72 |
|                | Conv         | 50.44±0.82          | 65.32±0.70 |
| R2D2           | Resnet       | 51.80±0.20          | 68.70±0.20 |
|                | Conv         | 49.50±0.20          | 65.40±0.20 |
| DN4            | Resnet       | 54.37±0.36          | 74.44±0.29 |
|                | Conv         | 51.24±0.74          | 71.02±0.64 |
| Dynamic-Net    | Resnet       | 55.45±0.89          | 70.13±0.68 |
|                | Conv         | 56.20±0.86          | 72.81±0.62 |
| Ours           | Resnet       | 58.12±0.92          | 78.39±0.58 |
|                | Conv         | 53.30±0.32          | 73.22±0.45 |

Methods with Conv Embedding

| Matching Nets  | Conv         | 43.56±0.84          | 55.31±0.73 |
| Meta-Learn LSTM| Conv         | 43.44±0.77          | 60.60±0.71 |
| MAML           | Conv         | 48.70±1.84          | 63.11±0.92 |
| CovAMNet       | Conv         | 51.19±0.76          | 67.65±0.63 |
| SalNet         | Conv         | 57.45±0.88          | 72.01±0.67 |

Generalizing to other datasets To better reflect the generalization performance of the few-shot learning models, we evaluate the few-shot learning model on the completely different datasets. A new dataset that totally different from the training dataset may present data distribution shift [Recht et al., 2019], which will cause significant performance degradation of the model. According to section 3.1, the training classes and the testing classes do not share the same label space, but they still possess the same data distribution because of coming from the same dataset. In this section, we train the model on miniImageNet and conduct the testing on the novel datasets to evaluate the generalization capability. The experiment results show that our model outperforms previous work (Proto Net, Relation Net, and K-tuplet loss [Li et al., 2019c]) on the three novel datasets, which demonstrates the superior generalization capacity of our approach.

5 Discussion

5.1 Ablation study

To demonstrate the influence of various modules, we perform the ablation study and the results are shown in Table 4. Firstly, we replace the ITCD with Image-To-Image distance (ITID) both in the triplet-input and no-triplet-input combination. The implementation of the ITID classifier is similar to the prototype network, where we use the vector flattened from the embedded feature to represent the embedded feature. The ITCD outperforms ITID in different settings, especially with the Resnet and triplet-input module, gaining approximately 26% improvement in 5-shot and 15% in 1-shot. Secondly, when disabling the triplet-input module, the accuracy drops down in the triplet-input+NA+ITCD combination. Conversely, disabling the triplet-input in the triplet-input+NA+ITCD obtains great performance improvement, gaining 8% improvement in 1-shot and 17% improvement in 5-shot. This may be caused by the inability of ITID to utilize the initial object seeds provided by the triplet-input module, leading to negative cooperation results. The ablation study shows our method is an organic whole and each module is elaborate for achieving localization.

Table 2: The mean accuracies of the 5-way 1-shot and 5-shot accuracies (%) on three fixed-trained datasets using the model trained on miniImageNet, with 95% confidence intervals. All the experiments are conducted with the same network for fair comparison.

| Dataset       | Proto Net | Relation Net | K-tuplet loss | ours       |
|---------------|-----------|--------------|---------------|------------|
| Stanford Dog  | 31.54±0.41| 31.24±0.61   | 37.33±0.65    | 43.69±0.72 |
|               | 31.27±0.41| 42.47±0.68   | 49.97±0.66    | 61.50±0.73 |
| Stanford Car  | 29.19±0.40| 28.83±0.55   | 31.20±0.58    | 32.87±0.57 |
|               | 38.00±0.42| 35.43±0.58   | 47.10±0.62    | 50.10±0.69 |
| CUB200        | 37.55±0.51| 38.30±0.71   | 40.16±0.68    | 43.30±0.75 |
|               | 55.03±0.49| 50.89±0.69   | 56.96±0.65    | 62.21±0.73 |

5.2 Visualization

We visualize the activated deep descriptors to show the localization. In the background image, the pixel value of the
Table 3: The mean accuracies of the 5-way 1-shot and 5-shot tasks on three fine-grained datasets, with 95% confidence intervals.

| Model       | Embedding | Stanford Dogs 1-shot | Stanford Cars 1-shot | Stanford DOgs 5-shot | Stanford Cars 5-shot | CUB-200 1-shot | CUB-200 5-shot |
|-------------|-----------|----------------------|----------------------|----------------------|----------------------|----------------|----------------|
| Baseline K-NN | Conv      | 26.14 ± 0.91         | 43.14 ± 1.02         | 23.50 ± 0.88         | 34.45 ± 0.98         | 25.81 ± 1.00  | 45.34 ± 1.03  |
| Proto Net   | Conv      | 37.59 ± 1.00         | 48.19 ± 1.03         | 40.90 ± 1.01         | 52.93 ± 1.03         | 37.36 ± 1.00  | 45.28 ± 1.03  |
| GNN         | Conv      | 46.98 ± 0.98         | 62.27 ± 0.95         | 55.85 ± 0.97         | 71.25 ± 0.89         | 51.83 ± 0.98  | 63.69 ± 0.94  |
| GNN         | Conv      | 49.10 ± 0.76         | 63.04 ± 0.65         | 56.65 ± 0.86         | 71.33 ± 0.62         | 52.42 ± 0.76  | 63.76 ± 0.64  |
| DN4         | Conv      | 45.41 ± 0.76         | 63.51 ± 0.62         | 59.84 ± 0.80         | 88.65 ± 0.44         | 46.84 ± 0.81  | 74.92 ± 0.64  |
| Ours        | Conv      | 52.50 ± 0.90         | 67.37 ± 0.76         | 46.46 ± 0.89         | 82.28 ± 1.02         | 55.31 ± 0.78  | 75.03 ± 0.64  |
| Ours        | Resnet    | 54.45 ± 0.95         | 73.63 ± 0.92         | 72.71 ± 0.43         | 93.94 ± 0.33         | 60.21 ± 0.87  | 78.70 ± 0.62  |

Figure 4: Illustration of the proposed few-shot weakly-supervised object localization. From up to down, every two rows are sampled from the miniImageNet, Stanford Dogs, Stanford Cars, and CUB-200, respectively. For each dataset, the first column is the initial object seeds and focus-area mask generated by Grad-CAM; the second column is the localization and activated-area mask generated by the proposed ITCD localizer. We resize activated-area mask using bicubic interpolation to show the localization in input images.

6 conclusions

This paper proposes a method to achieve weakly-supervised localization with very limited samples and only image-level labels, thus obtaining the most discriminative representations to perform few-shot classification. Thereinto, the specialized ITCD localizer can help the classification and the localization bootstrap from each other. Extensive experiments show the proposed method obtains superior performance over state-of-the-art methods on many benchmark datasets. The visualization shows our method can localize the key objects, which explains the high accuracies of our method in the FSL tasks.
References

[Alfassy et al., 2019] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shток, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. In CVPR, 2019.

[Bertinetto et al., 2019] Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. arXiv preprint arXiv:1805.08136, 2019.

[Boiman et al., 2008] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In CVPR, 2008.

[Finn et al., 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In ICML-Volume 70, 2017.

[Fu et al., 2017] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In CVPR, 2017.

[Gidaris and Komodakis, 2018] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In CVPR, 2018.

[Hariharan and Girshick, 2017] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In ICCV, 2017.

[Khosla et al., 2011] Aditya Khosla, Nityananda Jayadeva-prakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), 2011.

[Li et al., 2019a] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In CVPR, 2019.

[Li et al., 2019b] Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo. Distribution consistency based covariance metric networks for few-shot learning. In AAAI, 2019.

[Li et al., 2019c] Xiaomeng Li, Lequan Yu, Chi-Wing Fu, Meng Fang, and Pheng-Ann Heng. Revisiting metric learning for few-shot image classification. arXiv preprint arXiv:1907.03123, 2019.

[Makadia and Yumer, 2014] Ameesh Makadia and Mehmet Ersin Yumer. Learning 3d part detection from sparsely labeled data. In 2014 2nd International Conference on 3D Vision, 2014.

[Oquab et al., 2015] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?—weakly-supervised learning with convolutional neural networks. In CVPR, 2015.

[Ravi and Larochelle, 2017] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. ICLR, 2017.

[Recht et al., 2019] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? arXiv preprint arXiv:1902.10811, 2019.

[Satorras and Estrach, 2018] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043, 2018.

[Schwartz et al., 2019] Eli Schwartz, Leonid Karlinsky, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Baby steps towards few-shot learning with multiple semantics. arXiv preprint arXiv:1906.01905, 2019.

[Selvaraju et al., 2017] Ramprasaath R Selvaraju, MichaelCogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In ICCV, 2017.

[Snell et al., 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In NeurIPS, 2017.

[Sun et al., 2019] Xin Sun, Hongwei Xu, Junyu Dong, Qiong Li, and Changrui Chen. Few-shot learning for domain-specific fine-grained image classification. arXiv preprint arXiv:1907.09647, 2019.

[Vinyals et al., 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In NeurIPS, 2016.

[Wah et al., 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Wang et al., 2018a] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In CVPR, 2018.

[Wang et al., 2018b] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In CVPR, 2018.

[Wei et al., 2017] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing, 2017.

[Wertheimer and Hariharan, 2019] Davis Wertheimer and Bharath Hariharan. Few-shot learning with localization in realistic settings. In CVPR, 2019.

[Zhang et al., 2018] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In ECCV, 2018.

[Zhang et al., 2019] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In CVPR, 2019.

[Zhou et al., 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In CVPR, 2016.