Learning Task-aware Local Representations for Few-shot Learning

Chuanqi Dong , **Wenbin Li** , **Jing Huo*** , **Zheng Gu** and **Yang Gao**
State Key Laboratory for Novel Software Technology, Nanjing University, China
{dongchuanqi, guzheng}@smail.nju.edu.cn, {liwenbin, huojing, gaoy}@nju.edu.cn

Abstract

Few-shot learning for visual recognition aims to adapt to novel unseen classes with only a few images. Recent work, especially the work based on low-level information, has achieved great progress. In these work, local representations (LRs) are typically employed, because LRs are more consistent among the seen and unseen classes. However, most of them are limited to an individual image-to-image or image-to-class measure manner, which cannot fully exploit the capabilities of LRs, especially in the context of a certain task. This paper proposes an *Adaptive Task-aware Local Representations Network (ATL-Net)* to address this limitation by introducing episodic attention, which can adaptively select the important local patches among the entire task, as the process of human recognition. We achieve much superior results on multiple benchmarks. On the *mini*Imagenet, ATL-Net gains 0.93% and 0.88% improvements over the compared methods under the 5-way 1-shot and 5-shot settings. Moreover, ATL-Net can naturally tackle the problem that how to adaptively identify and weight the importance of different key local parts, which is the major concern of fine-grained recognition. Specifically, on the fine-grained dataset Stanford Dogs, ATL-Net outperforms the second best method with 5.39% and 9.69% gains under the 5-way 1-shot and 5-shot settings.

1 Introduction

Deep learning based methods [Krizhevsky *et al.*, 2012; He *et al.*, 2016] have achieved state-of-the-art performance on a variety of visual recognition tasks. These supervised methods need a lot of labeled data with diverse visual variations to effectively train a network. However, collecting a large amount of labeled data is time-consuming and laborious. In contrast, humans can recognize classes with extremely few labeled examples. Therefore, for machine learning algorithms, how to recognize classes with extremely few labeled examples, *i.e.,* few-shot learning, has attracted a lot of interests. Few-shot learning attempts to transfer the knowledge like humans, for generalizing to novel classes with very few supervisions.

To address few-shot learning tasks, a lot of methods have been proposed. However, most of these methods [Vinyals *et al.*, 2016; Snell *et al.*, 2017] adopt an image-level feature for classification and make an assumption that the image-level deep embedding space for the seen classes is extensively effective for the unseen classes, which is somewhat idealistic in practice. Fortunately, although the image-level embedding space is not equally effective for the seen and unseen classes, the low-level information, *i.e.,* the local representations (LRs) of semantic patches, among the seen and unseen classes, generally remain similar. Some recent methods [Li *et al.*, 2019c; Li *et al.*, 2019b; Sung *et al.*, 2018] have taken feature representations of semantic patches (*i.e.,* LRs) into consideration, but they do not fully exploit the capabilities of LRs in the context of the entire task. Recall the way that humans recognize an instance (object) into one of several unseen classes. It is quite natural that he/she will look for the distinct semantic patches which are only shared between the certain class and the query image. *In other words, the semantic patches commonly shared by all classes are not truly important for recognizing a novel instance.* For example, the way we recognize a "bird" among the "dog" and "cat" is quite different from the one among the "airplane" and "dragonfly". For the former one, the wings are important but are not the key concern for the latter one. Similarly, the fur and feather are more important for the latter one than the former. In other words, the importance of the semantic patches changes with the tasks.

As described above, the existing LRs based few-shot methods have not yet made full use of the information provided by the LRs mainly in two aspects: (1) the LRs are only considered inside one image or one class individually (*i.e.,* image-to-image or image-to-class manner), rather than the entire task; (2) the semantic local patches are weighted equally, rather than the more discriminative patches enjoy the higher weights. To overcome these two limitations, we design a *episodic attention mechanism*, which can select and weight the key patches without paying too much attention to the common parts among the entire task. Note that, in the work of [Li *et al.*, 2019b], a rank-based selection, *i.e.,* k-nearest neighbor (k-NN) selection, is utilized to select k (*e.g.,* $k = 3$) most related patches in each class for a query local instance. However, the number of related semantic patches for a query local

*Contact Author

instance shall be dynamically changed according to the context of the current task. It means that we may need more discriminative patches to recognize an object in one task, but just need much fewer patches to recognize the same object in another task. By contrast, the value range of the relationship between the related semantic patches mainly depends on semantic patches' nature, and remains relatively stable in changing tasks, so that we propose a *value-based selection* with a threshold to replace the rank-based selection. And then, we name the network with the above episodic attention mechanism as *Task-aware Local Representations Network (TL-Net)*.

Although the above mentioned value-based selection is appealing, it roughly sets a global manual threshold for all the semantic patches of tasks, which is hard to be effective for different semantic patches at the same time. To this end, we develop a trainable module to adaptively learn this threshold for each semantic patch, *i.e., adaptive value-based selection*. In this way, for each certain semantic patch, we can obtain its own relation threshold according to its nature. Typically, we call the extended TL-Net with learnable thresholds as *ATL-Net, Adaptive Task-aware Local Representations Network,* to show the additional adaptive ability relative to TL-Net.

Our contributions can be summarized as follows:

- We propose a novel *episodic attention mechanism* by exploring and weighting discriminative semantic patches inside the entire task, aiming to learn task-aware local representations for few-shot learning. Moreover, instead of the rank-based selection, a feasible *value-based selection* strategy is proposed.
- We further develop a trainable module to design an *adaptive value-based selection* strategy, making it possible to dynamically and adaptively select discriminative semantic patches for different tasks.
- We conduct comprehensive experiments on the challenging *mini*Imagenet and three fine-grained datasets to verify that the proposed ATL-Net achieves superior performance over the state-of-the-art methods.

2 Related Work

The recent literature of few-shot learning mainly comes from the following two categories: **meta-learning** based methods and **metric-learning** based methods.

2.1 Meta-learning based Methods

Meta-learning based methods learn the learning algorithm itself. [Santoro *et al.*, 2016] proposes an LSTM-based meta-learner to interact with an external memory module. The proposed framework in [Santoro *et al.*, 2016] adopts an LSTM-based meta-learner to learn a distinct optimization algorithm to train a classifier as well as learning a task-aware initialization for this classifier. MAML and its variants [Finn *et al.*, 2017] train a meta-leaner to provide suitable parameter initialization, so that they can be quickly adapted to a novel task. Similarly, [Li *et al.*, 2017] adjusts the update direction and learning rate for quickly adapting to a novel task. [Cai *et al.*, 2018] introduces the memory slots to construct a contextual learner for predicting the parameters of an embedding module for unlabeled images.

Nevertheless, these methods often need costly higher-order gradients or need another complicated memory structure, making these methods difficult to train and may lead to failure when scaling to deeper network architectures [Mishra *et al.*, 2018]. Compared with methods in this branch, the proposed ATL-Net can achieve competitive results with a much simpler network architecture, which is trained from scratch without fine-tuning.

2.2 Metric-learning based Methods

Metric-learning based methods address the few-shot classification problem by "learning to compare". [Koch *et al.*, 2015] proposes a Siamese neural network to learn generic image representations, which is conducted as a binary classification network and trained by a regularized cross-entropy loss. [Vinyals *et al.*, 2016] introduces an episodic training mechanism into few-shot learning and proposes the Matching Net by using attention and memory together. [Snell *et al.*, 2017] proposes a Prototypical Net by measuring the Euclidean distance between the class-mean feature and the query feature.

However, the above methods usually adopt an image-level global feature to represent each image based on a somewhat ideal assumption that the seen and unseen classes sharing a relatively consistent embedding space. In contrast, the low-level information, *i.e.,* the local representations (LRs) of semantic patches, is more consistent and transferable than the high-level global features among the seen and unseen classes, which has been verified in some recent work. For example, [Sung *et al.*, 2018] measures the distances between the query images and the support images by applying convolution layers on the concatenated feature maps, which implicitly uses the LRs. [Li *et al.*, 2019b] proposes DN4 to explicitly utilize the LRs through a k-nearest neighbor selection and enlarges the image-to-image search space to a more effective image-to-class one. However, these methods only consider the relationship between query images and classes at an image-level or a class-level without adequately mining the important information hidden behind the LRs at the task-level.

Different from the methods above, our ATL-Net can explore richer information of the LRs at the task-level and can adaptively select the key semantic patches for a specific task, as the progress of the human beings. Experiments on the challenging general and fine-grained datasets show the superiority of our method compared with other state-of-the-art methods.

3 The Proposed Method

3.1 Problem Definition

In this paper, we follow the common settings of few-shot learning methods. Given a small support set $\mathcal{S}$ which consists of N unseen classes with K samples per class, our goal is to classify a query sample $q \in \mathcal{Q}$ into one of the N support classes, which is called an N-way K-shot task. To achieve this goal, an auxiliary set $\mathcal{A}$ is employed to learn transferable knowledge using the episodic training mechanism [Vinyals *et al.*, 2016]. We divide $\mathcal{A}$ into many N-way K-shot tasks $\{\mathcal{T}\}$, where each $\mathcal{T}$ contains an *auxiliary support set* $\mathcal{A}_\mathcal{S}$ and an *auxiliary query set* $\mathcal{A}_\mathcal{Q}$. In the training stage, hundreds of tasks are fed into the model, encouraging the model to learn

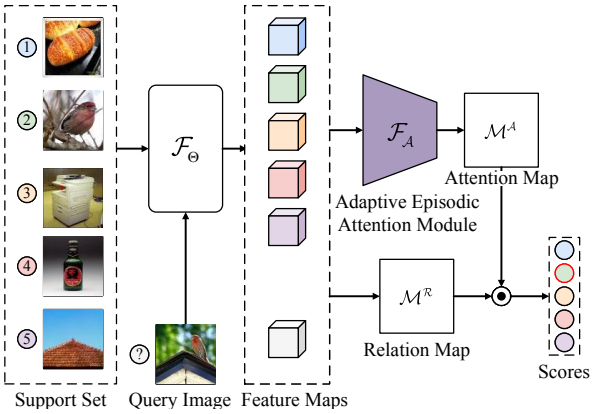

Figure 1: The overview of the proposed method under 5-way 1-shot setting. The model mainly consists of two parts: the embedding module $\mathcal{F}_\Theta$ to learn local representations and the adaptive episodic attention module $\mathcal{F}_\mathcal{A}$ to generate adaptive episodic attention for selecting discriminative patches for a special task. The score with the red circle indicates the predicted label. (Best viewed in color.)

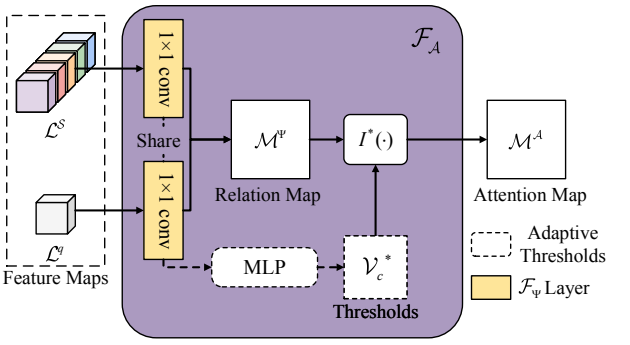

Figure 2: The framework of adaptive episodic attention module $\mathcal{F}_\mathcal{A}$. Through this module, we obtain episodic attention by $\mathcal{L}^q$ and $\mathcal{L}^\mathcal{S}$. The components with dashed lines generate adaptive thresholds $\mathcal{V}_c^*$, which is a fixed manual defined hyperparameter in TL-Net.

transferable knowledge that can be used in new N-way K-shot tasks (*i.e.,* $\mathcal{S}$ and $\mathcal{Q}$) with unseen classes. Note that $\mathcal{S}$ and $\mathcal{A}$ own different label spaces with no intersection.

Our overall framework is illustrated in Figure 1. All the images are first embedded into feature representations by an embedding module $\mathcal{F}_\Theta$. A local relation map $\mathcal{M}^\mathcal{R}$ is then calculated to capture the local relationship between the query image and the support set. Meanwhile, the adaptive episodic attention module $\mathcal{F}_\mathcal{A}$ learns an episodic attention map $\mathcal{M}^\mathcal{A}$, which can adaptively select the discriminative local patches among the support set for a certain query patch, as the process of human recognition. Note that the episodic attention focuses on the relations between the local patches, not isolated individuals. After that, we apply the attention map $\mathcal{M}^\mathcal{A}$ onto the relation matrix $\mathcal{M}^\mathcal{R}$ through an element-wise multiplication to eliminate noise, *i.e.,* the relation constructed by the commonly shared patches among the task, and then enhance the discriminative information. Finally, we can directly get the final score for classification from the processed relation matrix through naive methods, like addition.

3.2 Task-aware Local Representations

Let $x \in \mathcal{S} \cup \mathcal{Q}$ denote an input image, we first feed it into the embedding module $\mathcal{F}_\Theta$ to obtain a feature representation $\mathcal{F}_\Theta(x) \in \mathbb{R}^{C\times H\times W}$. Typically, we can get HW C-dimensional LRs for each input image, making up a total number of $NKHW$ support LRs, *i.e.,* $\mathcal{L}^\mathcal{S} = \mathcal{F}_\Theta(\mathcal{S}) \in \mathbb{R}^{C\times NKHW}$ and HW query LRs, *i.e.,* $\mathcal{L}^q = \mathcal{F}_\Theta(q) \in \mathbb{R}^{C\times HW}$. Then we calculate the relation matrix of these LRs as below:

$$\mathcal{M}^\mathcal{R}_{i,j} = g(\mathcal{L}^q_i, \mathcal{L}^\mathcal{S}_j)\,, \tag{1}$$

where $i \in \{1,\dots,HW\}$, $j \in \{1,\dots,NKHW\}$ and $g(\cdot,\cdot)$ is a similarity metric, which is implemented as cosine similarity in this paper. In contrast to previous methods that build image-level [Sung *et al.*, 2018] or class-level [Li *et al.*, 2019b] relationship, we aim to build a task-level relationship while maintaining discriminative relations at the same time.

Further more, we apply a transformation layer $\mathcal{F}_\Psi$ (*i.e.,* the 1×1 conv layer in Figure 2) on the original LRs, and then learn another relation matrix $\mathcal{M}^\Psi$ for subsequent operations:

$$\mathcal{M}^\Psi_{i,j} = g(\mathcal{F}_\Psi(\mathcal{L}^q_i), \mathcal{F}_\Psi(\mathcal{L}^\mathcal{S}_j))\,, \tag{2}$$

where $i \in \{1,\dots,HW\}, j \in \{1,\dots,NKHW\}$. Each row in this matrix represents the adaptive subspace relationship of each position in the query image to all positions of all images in the support set. Moreover, we eliminate the noises (*i.e.,* the trivial relations) in the relation matrix $\mathcal{M}^\Psi$ by a threshold $\mathcal{V}_c$, and then produce an episodic attention map $\mathcal{M}^\mathcal{A}$ as below:

$$\mathcal{M}^\mathcal{A}_{i,j} = \frac{I(\mathcal{M}^\Psi_{i,j})}{\sum_j I(\mathcal{M}^\Psi_{i,j})} \tag{3}$$

$$I(x) = \begin{cases} x, & \text{if } x > \mathcal{V}_c \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

As Eq. (3) shown, the common patches shared by multiple classes among the entire task will "dilute" the attention, and thus they will enjoy relatively small attention values. Meanwhile, we find that although the influence of each noise (*i.e.,* each trivial relation) is slight, it still greatly affects the distribution of the episodic attention due to the large number. For this reason, we apply Eq. (4) to construct a sparse episodic attention. In fact, this sparse episodic attention is more similar to a selection process or a hard attention rather than a soft attention. Next, we perform an element-wise multiplication between $\mathcal{M}^\mathcal{A}$ and $\mathcal{M}^\mathcal{R}$ to obtain a weighted relation matrix $\mathcal{M}^\mathcal{A} \odot \mathcal{M}^\mathcal{R}$, and then collect the weighted relation between query q and the n-th class to obtain the score for n-th class:

$$\text{Score}_n = \frac{\mathcal{V}_s}{HW} \sum_{i=1}^{HW} \sum_{j=\mathcal{Z}^n_1}^{\mathcal{Z}^n_{KHW}} (\mathcal{M}^\mathcal{A} \odot \mathcal{M}^\mathcal{R})_{i,j}\,, \tag{5}$$

where $\mathcal{V}_s$ is a temperature for the following cross-entropy loss, and $\mathcal{Z}^n_k$ indicates the k-th relation of KHW relations belong to the n-th class in the $NKHW$ relations of the entire support set $\mathcal{S}$. Finally, we can obtain the classification probability $\mathcal{P}^q$ of query q by a softmax function. Note that based on the above process, we can develop the *Task-aware Local Representations Network (TL-Net)*, which can be easily implemented by two matrix multiplications, an element-wise multiplication as well as some convolution operations.

Dataset	Stanford Dogs	Stanford Cars	CUB-200
N_{all}	120	196	200
N_{train}	70	130	130
N_{val}	20	17	20
N_{test}	30	49	50

Table 1: The splits of three fine-grained datasets. N_{all} is the total number of classes. N_{train}, N_{val} and N_{test} indicate the number of classes in training (auxiliary) set, validation set and test set.

3.3 Adaptive Threshold for Episodic Attention

In the sections above, we introduce the TL-Net, where a fixed threshold $\mathcal{V}_c$ (*i.e.,* a global scalar) is used to select the most informative relationships. However, such kind of selection is sensitive to the value of $\mathcal{V}_c$ and not flexible for different query patches, which is shown in Figure 3. To handle this problem, we propose a novel *adaptive episodic attention module* $\mathcal{F}_{\mathcal{A}}$, which can learn different thresholds for different patches.

Figure 2 shows the framework of our adaptive episodic attention module. Different from the method mentioned in Eq. (4), we use a Multi-Layer Perceptron (MLP) $\mathcal{F}_\Gamma$ to adaptively predict the threshold for each LR of the query image. Specifically, $\mathcal{F}_\Gamma$ takes the query LRs as input and outputs a threshold $\mathcal{V}_c^*$:

$$\mathcal{V}_c^* = \sigma(\mathcal{F}_\Gamma(\mathcal{L}_i^q)), \tag{6}$$

where σ is a sigmoid function. Beyond that, to narrow the search space for $\mathcal{V}_c^*$, we change the output range of sigmoid function σ. However, the step function used in Eq. (4) is indifferentiable. So we approximate it using a variant $I^*(\cdot)$ of sigmoid function with a hyperparameter k:

$$I^*(x) = x/(1+\exp^{-k(x-\mathcal{V}_c^*)}), \tag{7}$$

where $\mathcal{V}_c^*$ is the corresponding threshold value for x, and x denotes one of the values in $\mathcal{M}^{\mathcal{A}}$. Theoretically, when k is large enough, the $I^*(\cdot)$ can be considered as $I(\cdot)$. Moreover, we call the extended TL-Net with learnable thresholds as *ATL-Net, Adaptive Task-aware Local Representations Network,* to show its additional adaptive ability. The training process of the proposed ATL-Net is shown in Algorithm 1.

4 Experiments

4.1 Datasets

***mini*ImageNet** [Vinyals *et al.*, 2016] is a subset of ImageNet [Deng *et al.*, 2009], which consists of 100 classes and 600 images per class. Following the commonly used strategy, we divide the dataset into training (auxiliary)/validation/test set with a percentage of $64/16/20$ respectively.

We also evaluate our method on three fine-grained image classification datasets. **Stanford Dogs** [Khosla *et al.*, 2011] contains 120 categories with a total number of $20,580$ images. **Stanford Cars** [Krause *et al.*, 2013] contains 196 classes of cars and $16,185$ images. **CUB-200** [Welinder *et al.*, 2010] contains 200 bird species with a total number of $6,033$ images. For fair comparisons, we use the data splits of [Li *et al.*, 2019b; Li *et al.*, 2019c; Huang *et al.*, 2019], as Table 1 shows.

Algorithm 1 Training of ATL-Net

Input: Episodic task $\mathcal{T} = \{\mathcal{A}_\mathcal{S}, \mathcal{A}_\mathcal{Q}\}$

while no converge **do**
  $\mathcal{L}^\mathcal{S} \leftarrow \mathcal{F}_\Theta(\mathcal{A}_\mathcal{S})$
  $\mathcal{L}^\mathcal{Q} \leftarrow \mathcal{F}_\Theta(\mathcal{A}_\mathcal{Q})$
  for $\mathcal{L}^q$ in $\mathcal{L}^\mathcal{Q}$ **do**
    Get relation matrix $\mathcal{M}^\mathcal{R}$ by Eq. (1)
    Calculate adaptive threshold $\mathcal{V}_c^*$ for $\mathcal{L}^q$ by Eq. (6)
    Construct adaptive episodic attention $\mathcal{M}^\mathcal{A}$ by Eq. (2), Eq. (3) and Eq. (7)
    Calculate probability $\mathcal{P}^q$ for $\mathcal{L}^q$ by Eq. (5)
  end for
  $L \leftarrow -\sum \mathcal{Y}\log(\mathcal{P})$
  Mini-batch Adam to minimize L, update Θ, Ψ and Γ
end while

4.2 Implementation Details

Network architecture. We follow the basic feature extraction network which is used in previous works [Li *et al.*, 2019b; Li *et al.*, 2019c]. The feature extraction network $\mathcal{F}_\Theta$ consists of 4 convolution blocks, each of which contains a convolutional layer, batch normalization and LeakyReLU activation. The transformation layer $\mathcal{F}_\Psi$ consists a 1×1 convolutional layer followed by batch normalization and LeakyReLU activation. The MLP module $\mathcal{F}_\Gamma$ is implemented by two fully connected layers. In fact that only a few parameters are introduced by $\mathcal{F}_\Psi$ and $\mathcal{F}_\Gamma$, which will be discussed in Section 5.

Training and testing detail. We implement our experiments using PyTorch [Paszke *et al.*, 2019]. All the images are resized to 84×84. During the training stage, we randomly construct $250,000$ episodes from the training (auxiliary) set for the *mini*Imagenet dataset and the Stanford Car dataset, and $150,000$ for the other two datasets to avoid overfitting. In each episode, we collect 15 query images per class. For example, under 5-way 1-shot setting, we have 5 support images and 75 query images in each task. We use Adam [Kingma and Ba, 2015] optimizer with a cross-entropy loss to train the network. The initial learning rate is set to 0.001. During the test, we evaluate the proposed ATL-Net on 600 randomly sampled tasks. The mean accuracy, as well as the 95% confidence interval will be reported after being repeated five times. Note that the whole model is trained from scratch in an end-to-end manner without any data augmentation and weight decay, neither do fine-tune in the test stage [1].

4.3 Baselines

To evaluate the proposed ATL-Net on the *mini*Imagenet, we make comparisons with eleven state-of-the-art models, including Matching Net [Vinyals *et al.*, 2016], MAML [Finn *et al.*, 2017], Prototypical Net [Snell *et al.*, 2017], GNN [Satorras and Estrach, 2018], Relation Net [Sung *et al.*, 2018], MetaGAN [Zhang *et al.*, 2018], MM-Net [Cai *et al.*, 2018], MEPS [Chu *et al.*, 2019], CovaMNet [Li *et al.*, 2019c], DN4 [Li *et al.*, 2019b] and GCR [Li *et al.*, 2019a].

[1] The source code can be available from https://github.com/LegenDong/ATL-Net

Model	Backbone	Additional Stage	5-way 1-shot	5-way 5-shot
Matching Net [Vinyals *et al.*, 2016]	Conv-64F	N	43.56 ± 0.84	55.31 ± 0.73
MAML [Finn *et al.*, 2017]	Conv-32F	Y	48.70 ± 1.84	63.11 ± 0.92
Prototypical Net [Snell *et al.*, 2017]	Conv-64F	N	49.42 ± 0.78	68.20 ± 0.66
GNN [Satorras and Estrach, 2018]	Conv-256F	N	50.33 ± 0.36	66.41 ± 0.63
Relation Net [Sung *et al.*, 2018]	Conv-64F	N	50.44 ± 0.82	65.32 ± 0.70
MetaGAN [Zhang *et al.*, 2018]	Conv-64F	N	52.71 ± 0.64	68.63 ± 0.67
MM-Net [Cai *et al.*, 2018]	Conv-64F	N	53.37 ± 0.48	66.97 ± 0.35
MEPS [Chu *et al.*, 2019]	Conv-64F	N	51.03 ± 0.78	67.96 ± 0.71
CovaMNet [Li *et al.*, 2019c]	Conv-64F	N	51.19 ± 0.76	67.65 ± 0.63
DN4 [Li *et al.*, 2019b]	Conv-64F	N	51.24 ± 0.74	71.02 ± 0.64
GCR [Li *et al.*, 2019a]	Conv-64F	Y	53.21 ± 0.40	72.34 ± 0.32
ATL-Net (Ours)	Conv-64F	N	$\mathbf{54.30 \pm 0.76}$	$\mathbf{73.22 \pm 0.63}$

Table 2: Comparisons with other methods on *mini*Imagenet. The second column shows which kind of embedding module is employed. The third column denotes whether the model contains additional training stage, *e.g.* pretrain stage or fine-tune stage. We use the officially provided results for all the other methods. For each setting, the best and the second best results are highlighted.

Model	Stanford Dogs		Stanford Cars		CUB-200	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Matching Net	35.80 ± 0.99	47.50 ± 1.03	34.80 ± 0.98	44.70 ± 1.03	45.30 ± 1.03	59.50 ± 1.01
Prototypical Net	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	46.98 ± 0.98	62.27 ± 0.95	55.85 ± 0.97	71.25 ± 0.89	51.83 ± 0.98	63.69 ± 0.94
DN4	45.41 ± 0.76	63.51 ± 0.62	59.84 ± 0.80	88.65 ± 0.44	46.84 ± 0.81	74.92 ± 0.64
CovaMNet	49.10 ± 0.76	63.04 ± 0.65	56.65 ± 0.86	71.33 ± 0.62	52.42 ± 0.76	63.76 ± 0.64
PABN+$_{\text{cpt}}$	45.65 ± 0.71	61.24 ± 0.62	54.44 ± 0.71	67.36 ± 0.61	-	-
LRPABN$_{\text{cpt}}$	45.72 ± 0.75	60.94 ± 0.66	60.28 ± 0.76	73.29 ± 0.58	-	-
ATL-Net (Ours)	$\mathbf{54.49 \pm 0.92}$	$\mathbf{73.20 \pm 0.69}$	$\mathbf{67.95 \pm 0.84}$	$\mathbf{89.16 \pm 0.48}$	$\mathbf{60.91 \pm 0.91}$	$\mathbf{77.05 \pm 0.67}$

Table 3: Comparisons with other methods on three fine-grained datasets. We adopt the results from [Li *et al.*, 2019c] for the first three methods and the officially provided results for the other methods. For each setting, the best and the second best results are highlighted.

For fine-grained image classification datasets, we compare our method with six few-shot methods, Matching Net [Vinyals *et al.*, 2016], Prototypical Net [Snell *et al.*, 2017], GNN [Satorras and Estrach, 2018], DN4 [Li *et al.*, 2019b] and CovaMNet [Li *et al.*, 2019c], and the fine-grained methods PABN+$_{\text{cpt}}$/LRPABN$_{\text{cpt}}$ [Huang *et al.*, 2019].

4.4 Comparisons with the SOTA Methods

We make comparisons with several state-of-the-art methods under 5-way 1-shot and 5-way 5-shot settings.

Results on *mini*Imagenet. The results on *mini*Imagenet are summarized in Table 2. It can be seen that our method significantly outperforms other methods under both settings. We achieve 54.30% under the 5-way 1-shot setting, with an improvement of 0.93% from the second best [Cai *et al.*, 2018]. Moreover, compared with [Cai *et al.*, 2018], the proposed ATL-Net introduces simpler additional structures (*i.e.,* $\mathcal{F}_\Psi$ and $\mathcal{F}_\Gamma$) than the complex memory-addressing architectures. Similarly, our ALT-Net also gets higher performance, 0.88% improvement than previous methods [Li *et al.*, 2019a] that uses data augmentation, data hallucination [Wang *et al.*, 2018] and pretrains the feature extractor on the whole training set. Note that the proposed ATL-Net achieves an improvement of $3.06\%/2.20\%$ under 5-way 1-shot/5-shot settings than the most relevant work [Li *et al.*, 2019b], which exploits the relation at the class-level by k-NN selection. Such a great improvement further proves the superiority of our method that select the distinct patches, which are only shared between a certain class and query images.

Results on fine-grained datasets. The results on the three fine-grained datasets are summarized in Table 3. Due to the results for [Huang *et al.*, 2019] on the CUB-200 [Welinder *et al.*, 2010] dataset is not provided, we leave them blank. It can be observed that our method achieves the best performance compared with both general and fine-grained-specific few-shot learning methods. Compared with the general few-shot learning methods, our method is 5.39%, 8.11% and 8.49% better than the second best under the 5-way 1-shot setting. The results compared with the fine-grained few-shot learning methods are similar, we obtain 7.67% improvements at least. The reason for these great improvements is that ATL-Net can naturally tackle the major challenge of identifying and weighting the importance of the key parts [Sun *et al.*, 2018]. The proposed method will not be fooled by the similar global geometry and appearances, and thus pay more attention to their subtle differences behind the key parts.

4.5 Ablation Study

To further verify the effectiveness of the proposed ATL-Net, we conduct ablation studies on *mini*Imagenet, the results are reported in Table 4. We remove $\mathcal{F}_\Psi$ and $\mathcal{F}_\Gamma$ from the network respectively to confirm that each part of the model is indis-

Factor	5-way 1-shot	5-way 5-shot
(i) baseline	50.94 ± 0.79	65.16 ± 0.72
(ii) w/o $\mathcal{F}_\Gamma$ (TL-Net)	53.24 ± 0.80	71.87 ± 0.65
(iii) w/o $\mathcal{F}_\Psi$	53.80 ± 0.81	72.95 ± 0.64
ATL-Net (Ours)	$\mathbf{54.30 \pm 0.76}$	$\mathbf{73.22 \pm 0.63}$

Table 4: Ablation study on *mini*Imagenet for the proposed ATL-Net.

Model	Params	5-way 5-shot
Prototypical Net	**0.113M**	68.20
Relation Net	0.229M	65.32
GNN	1.619M	66.41
DN4	**0.113M**	71.02
GCR w/o Hallucinator	1.755M	72.34
ATL-Net (Ours)	0.117M	**73.22**

Table 5: The number of trainable parameters along with 5-way 5-shot performance of different models.

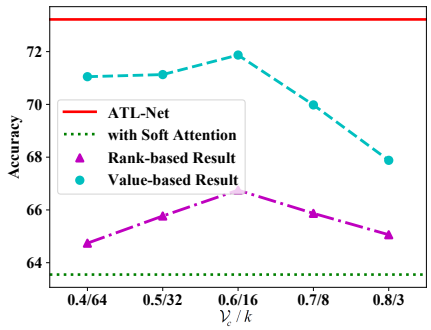

Figure 3: The results of the rank-based and value-based selections under the 5-way 5-shot setting. The abscissa is the value of hyper-parameters near the peak for value-based/rank-based methods, from introducing noise to losing information. The red line and green dash line are the results of ATL-Net and TL-Net with soft attention, respectively. (Best viewed in color.)

pensable, and them both as the baseline for comparison. It can be observed that the main improvement comes from the adaptive threshold module $\mathcal{F}_\Gamma$, without it about 1.06% and 1.35% performance loss occurring. Intuitively, $\mathcal{F}_\Psi$ can somewhat alleviate the effects of an unsuitable threshold for a certain relation, by measuring in an adaptive subspace without harming the original features. In contrast, $\mathcal{F}_\Gamma$ tries to directly assign an adaptive threshold, making it more effective than $\mathcal{F}_\Psi$. However, $\mathcal{F}_\Gamma$ still cannot handle all changeable cases, so both $\mathcal{F}_\Gamma$ and $\mathcal{F}_\Psi$ are indispensable. Moreover, even for the TL-Net (*i.e.,* ATL-Net without $\mathcal{F}_\Gamma$), we also obtain competitive results with only 0.13% and 0.47% lower than [Cai *et al.*, 2018] and [Li *et al.*, 2019a] under 5-way 1-shot and 5-shot settings, but much better than the other methods.

5 Discussion

5.1 Value-based v.s. Rank-based Selection

To verify the superiority of value-based selection than rank-based one [Li *et al.*, 2019b], we replace the adaptive attention module by a k-NN selection, the peak and surrounding results under 5-way 5-shot setting are reported in the Figure 3. The results of the value-based selection are generally better than the results of the rank-based selection. Meanwhile, we observe that it's difficult for the rank-based method to select enough corresponding LRs without too much noise at the task-level. For example, the rank-based methods cannot select enough corresponding LRs for $k = 3$ or 8 , while the corresponding LRs contain numerous noise when $k = 32$ or 64.

5.2 Influence of the Threshold $\mathcal{V}_c$

The results of the threshold $\mathcal{V}_c$ are reported in Figure 3, which shows that the choice of the threshold $\mathcal{V}_c$ has a mild impact on performance, and we observe that the performance degradation due to information loss is more severe than the noise introduced. To verify the importance of episodic attention's sparsity, we replace the hard attention by a soft attention. The result shows that the trivial relations, even with relatively small attention values, still greatly affect the distribution of episodic attention. Beyond that, the proposed adaptive episodic attention module $\mathcal{F}_\mathcal{A}$ outperforms 1.35% than the best result with a manual selection for $\mathcal{V}_c$. Such a great improvement shows that the global manually set $\mathcal{V}_c$ can only achieve a suboptimal solution, but the adaptive threshold $\mathcal{V}_c^*$ is more robust in complex situations.

5.3 Number of Trainable Parameters

We also compare the number of trainable parameters to verify the efficiency of the proposed ATL-Net, as the Table 5 shows. Since no other trainable parameters are introduced except for the embedding module $\mathcal{F}_\Theta$, [Snell *et al.*, 2017; Li *et al.*, 2019b] become the most light-weight models. [Satorras and Estrach, 2018] adopts a larger embedding module (*i.e.,* the filter number is 256), which draws a great contribution from the number of parameters. [Sung *et al.*, 2018; Li *et al.*, 2019a] adopt additional architectures to boost the result, which also introduce a huge number of trainable parameters. However, the proposed ATL-Net only introduces a small number of the trainable parameters, while achieves a better result than the methods above.

6 Conclusions

In this paper, we propose an *Adaptive Task-aware Local Representation Network (ATL-Net)* for few-shot learning, aiming to learn more discriminative local representations by taking a view of the entire task. Specifically, an adaptive episodic attention mechanism is designed to adaptively select the key semantic patches for a special task, without distracting attention by the common parts shared by most classes. Extensive experimental results on the benchmarks verify the effectiveness and superiority of the proposed ATL-Net.

Acknowledgements

This work is supported by National Key R&D Program of China (2018YFB1402600), NSFC (61806092) and Jiangsu Natural Science Foundation (No. BK20180326).

References

[Cai *et al.*, 2018] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *CVPR*, pages 4080–4088, 2018.

[Chu *et al.*, 2019] Wen-Hsuan Chu, Yu-Jhe Li, Jing-Cheng Chang, and Yu-Chiang Frank Wang. Spot and learn: A maximum-entropy patch sampler for few-shot image classification. In *CVPR*, pages 6251–6260, 2019.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135. JMLR. org, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Huang *et al.*, 2019] Huaxi Huang, Junjie Zhang, Jian Zhang, Jingsong Xu, and Qiang Wu. Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. *arXiv preprint arXiv:1908.01313*, 2019.

[Khosla *et al.*, 2011] Aditya Khosla, Nityananda Jayadevaprakash, 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.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.

[Koch *et al.*, 2015] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, pages 554–561, 2013.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[Li *et al.*, 2017] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[Li *et al.*, 2019a] Aoxue Li, Tiange Luo, Tao Xiang, Weiran Huang, and Liwei Wang. Few-shot learning with global class representations. In *ICCV*, pages 9715–9724, 2019.

[Li *et al.*, 2019b] 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*, pages 7260–7268, 2019.

[Li *et al.*, 2019c] 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*, volume 33, pages 8642–8649, 2019.

[Mishra *et al.*, 2018] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

[Santoro *et al.*, 2016] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, pages 1842–1850, 2016.

[Satorras and Estrach, 2018] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *ICLR*, 2018.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087, 2017.

[Sun *et al.*, 2018] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, pages 805–821, 2018.

[Sung *et al.*, 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638, 2016.

[Wang *et al.*, 2018] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, pages 7278–7286, 2018.

[Welinder *et al.*, 2010] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[Zhang *et al.*, 2018] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *NeurIPS*, pages 2365–2374, 2018.