Looking Wider for Better Adaptive Representation in Few-Shot Learning

Jiabao Zhao[1,2], **Yifan Yang**[3], **Xin Lin**[1,2,*], **Jing Yang**[2], **Liang He**[1,2,†]

[1] Shanghai Key Laboratory of Multidimensional Information Processing, ECNU, Shanghai, China
[2] School of Computer Science and Technology, East China Normal University, Shanghai, China
[3] Transwarp Technology (Shanghai) Co., Ltd, China
jiabao.zhao@ica.stc.sh.cn, yifan.yang@transwarp.io, {xlin, jyang, lhe}@cs.ecnu.edu.cn

Abstract

Building a good feature space is essential for the metric-based few-shot algorithms to recognize a novel class with only a few samples. The feature space is often built by Convolutional Neural Networks (CNNs). However, CNNs primarily focus on local information with the limited receptive field, and the global information generated by distant pixels is not well used. Meanwhile, having a global understanding of the current task and focusing on distinct regions of the same sample for different queries are important for the few-shot classification. To tackle these problems, we propose the Cross Non-Local Neural Network (CNL) for capturing the long-range dependency of the samples and the current task. CNL extracts the task-specific and context-aware features dynamically by strengthening the features of the sample at a position via aggregating information from all positions of itself and the current task. To reduce losing important information, we maximize the mutual information between the original and refined features as a constraint. Moreover, we add a task-specific scaling to deal with multi-scale and task-specific features extracted by CNL. We conduct extensive experiments for validating our proposed algorithm, which achieves new state-of-the-art performances on two public benchmarks.

Introduction

Traditional deep neural networks for recognization tasks require large-scale and category-balanced data for training, and the categories are fixed. However, the model needs to train from scratch when new categories emerge. Few-shot learning addresses this problem by recognizing the novel category (unseen during training) based on a few labeled samples (Vinyals et al. 2016; Chelsea et al. 2017). Recently, several few-shot algorithms have been proposed for various tasks in CV (ComputerVision) and NLP (Natural Language Processing) domains. Metric-based methods achieve excellent performances on multiple tasks with the simplicity (Vinyals et al. 2016; Snell et al. 2017; Li et al. 2020). This method learns to build an appropriate feature space, where similar samples are close and different samples are distant. Then it measures the similarities of the samples for prediction. The key to metric-based approaches is relying on extracting high-quality representation and choose an appropriate metric function.

Figure 1: There are two 2-way 1-shot tasks. For different queries, the model should focus on different regions of the same images in support set. In addition to local features, the global understanding of the current task is also essential.

Extracting context-aware features dynamically by considering the current task's global information is essential for few-shot tasks, especially for the samples containing multiple objects. As shown in Figure 1, there is a need to focus on distinct regions of the same samples in the support set for different queries because the correlativity between the distant pixels in an image and the relationship between the labeled sample and the query are valuable. Hence, capturing long-range dependency (Zhao et al. 2017) plays a crucial role in few-shot image classification. However, most metric-based approaches take the CNNs as the backbone. It cannot cover distant areas with correlativity and transmit a message between distant positions efficiently due to the limited receptive field. Non-local neural network (Wang et al. 2018) was proposed for capturing the long-range dependency, which computes the response at a position as a weighted sum of the features at all positions of the image. However, it only captures the global context and cannot extract adaptive representation based on various tasks. Besides, some essential local features are likely to be lost during the refinement.

To address the above problems, we propose a Cross Non-Local Neural Network (CNL) for capturing the long-range dependency of the samples and its related task, as shown in Figure 4. CNL aims to learn different attention weights

*Corresponding author: Xin Lin
†Corresponding author: Liang He

on the same samples for different queries with a global context of the current task. To this end, CNL strengthens the features at a position of the sample via aggregating information from all positions of itself and the current task, which can extract task-specific and multi-scale features and build a better feature space for prediction. To maintain the primary information during the reorganization process, we concatenate on the local features and global features in CNL. Besides, we maximize the mutual information (MI) between the local features extracted by the backbone and the task-specific representation extracted by CNL, which is treated as a constraint for training CNL. However, MI is challenging to compute in either high-dimensional or continuous settings. Here, we use Noise-Contrastive Estimation (NCE) (Gutmann and Hyvärinen 2010) as the lower bound on mutual information, and we adopt infoNCE (Oord, Li, and Vinyals 2018) to estimate the mutual information between the two features.

The proposed CNL with MI-based constraint extracts the features at local and global levels, and the features are dynamic according to different tasks. Cosine similarity cannot deal with multi-scale and task-specific features well because it is not sensitive to absolute value and cannot measure the angle exactly between the vectors without eliminating the offsets in patterns. To this end, we propose a task-specific scaling for normalization based on different tasks. The cosine similarity with the task-specific scaling outperforms other metric functions and can achieve new state-of-the-art performance without higher computational complexity and additional operation. Additionally, the task-specific scaling can be adapted to other metric functions.

We have carried out extensive experiments on multiple datasets. Our proposed method outperforms the baseline methods and achieves new state-of-the-art performances. The ablation experiments show that each key component in our proposed method plays a critical role, respectively. This method aims to build a better feature space and pay different attention to the same images based on different queries. And this idea also can be used in other tasks for extracting query-specific features, such as image retrieval and visual question answering.

In summary, our main contributions are three-fold:

- We propose Cross Non-Local Neural Network (CNL) to capture the long-range dependency for few-shot learning. CNL extracts the task-specific and context-aware features by considering the global information of the sample itself and the relationship between the specific query and the samples in the support set.
- We adopt infoNCE to estimate the mutual information between the task-specific representation and the local features. We regard it as a constraint for training the CNL, which aims to avoid losing important information during the refining process.
- We propose a task-specific scaling for dealing with the multi-scale and task-specific features extracted by CNL. It can be added to the many metric functions, which is efficient and can improve the performance significantly.

Related Work

Few-Shot Learning

Recently, few-shot learning attracts more attention in both CV (ComputerVision) and NLP (Natural Language Processing) fields, which aims to recognize a new category only based on a few labeled samples. There are many advanced Few-shot learning algorithms, such as metric-based method (Vinyals et al. 2016; Li et al. 2020; Zhang et al. 2020), data augmentation method (Zhang et al. 2018; Tsutsui et al. 2019; Alfassy et al. 2019), memory network (Geng et al. 2020), graph neural network (Kim et al. 2019), and meta-learning methods (Ravi and Larochelle 2017; Jamal and Qi 2019).

The metric-based method consists of two main modules, feature extraction and a metric function. It aims to represent the samples in an appropriate feature space where similar samples are closer and dissimilar samples are farther. It can be broadly classified as task-invariant (Vinyals et al. 2016; Allen et al. 2019), task-specific (Triantafillou et al. 2017; Hou et al. 2019), and hybrid models (Bertinetto et al. 2016; Oreshkin et al. 2018). Most of them adopt CNNs to extract features. However, CNNs are better at capturing local information, and the stacked CNNs cannot capture the long-range dependency adequately. In the few-shot learning, the global information is important for building the feature space because the adaptability and transferability can be improved if the model can focus on different parts of the same samples based on different queries. In this paper, we propose to capture the long-range dependency for the samples in the current task and extract task-specific features for building better feature space. (Hou et al. 2019) propose Cross Attention Network extract discriminative features, which calculates the correlation map between the samples in the support set and the queries by cosine similarity. In comparison with (Hou et al. 2019), our proposed Cross Non-Local Neural Network extracts the global information and local information of all samples in the current task at first and then captures the long-range dependency between the labeled samples and queries by using the refined features. Besides, we propose an MI-based constraint for maintaining the primary information during the refinement.

Long-Range Dependency Modeling

Convolutional Neural Networks have an excellent performance in many visual tasks (Krizhevsky et al. 2012; He et al. 2016; Huang et al. 2017). Many studies (Liu, Rabinovich, and Berg 2015; Zhao et al. 2017) have validated that capturing long-range dependencies can improve the performance in many domains. CNNs capture the long-range dependencies by deeply stacking convolution layers. However, it is difficult to deliver messages between distant positions because of the limited receptive field of a single convolutional layer. Moreover, it is ineffective to directly repeat the convolutional layers and use a big kernel to enlarge the receptive field for covering other areas.

There are many methods to capture long-range dependence, such as conditional random fields for semantic segmentation (Krähenbühl and Koltun 2011), feedforward networks for modeling sequences in language (Gehring et al. 2017), self-

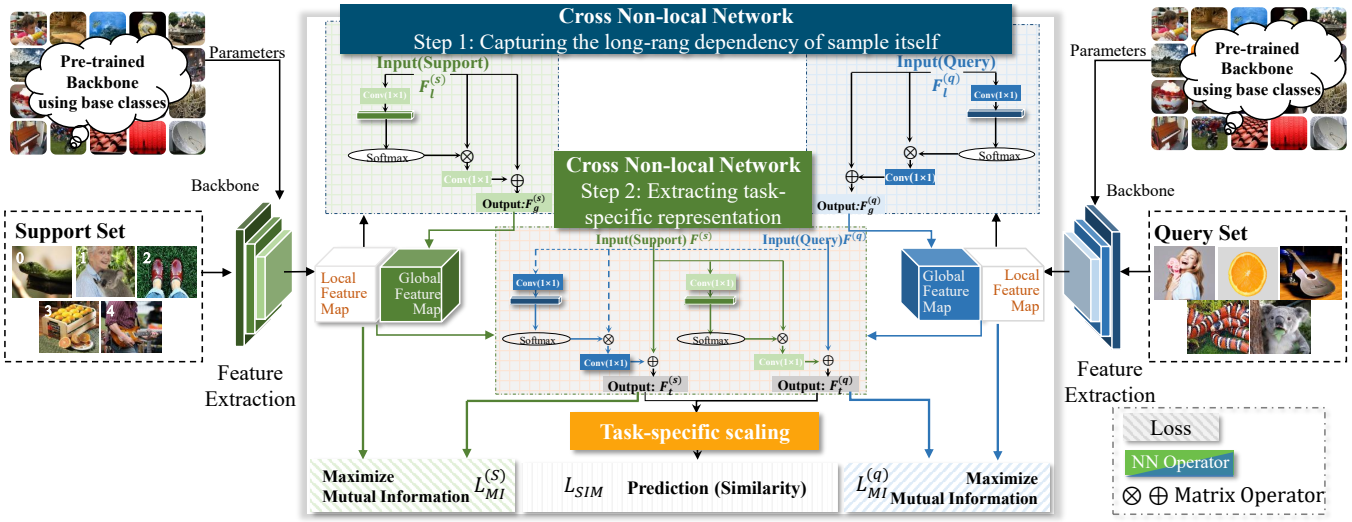


Figure 2: The overview of our proposed framework for few-shot learning. At first, we extract the local feature map F_l by the pre-trained backbone. Here we use the *ResNet12*. Secondly, we extract task-specific representation by our proposed Cross Non-Local Neural Network. We capture the long-range dependency of the samples themselves and concatenate the global and local features at step one. Then we refine the representation with the global context of the current task. In addition, we add a task-specific scaling to deal with the multi-scale and task-specific features. At last, we compute the cross-entropy loss and the mutual information between the original and refined features for optimation.

attention for machine translation (Vaswani et al. 2017) and Non-local neural Network for video classification (Wang et al. 2018). The core concept of the non-local neural network is the Non-local means (Buades, Coll, and Morel 2005), and it computes interactions between any two positions for capturing the long-range dependencies, i.e., it computes the response at a position as a weighted sum of the features at all position. (Cao et al. 2019) find that the attention maps for the different positions are almost the same, and they propose a simplified non-local block by explicitly using the same attention map for all positions. However, all of them only extract the global features of the current samples. To obtain the task-specific representation in few-shot learning, we propose a Cross Non-Local Neural Network (CNL), and it refines the features based on the global information of all samples in the current task. To maintain the important features during the refinement, we maximize the mutual information between the local features and task-specific features as a constraint for training the CNL.

Method

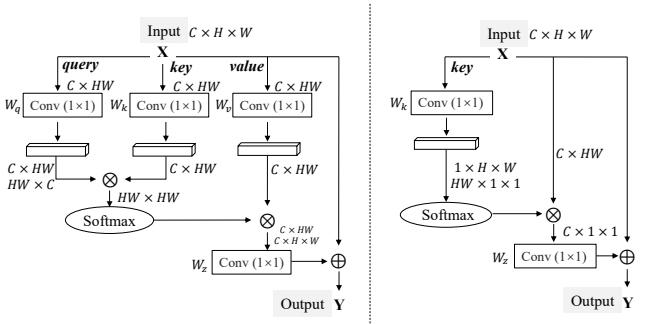
In this section, we start with the problem definition in this work, then give a brief overview of the proposed framework. Secondly, we revisit the non-local neural network for capturing long-range dependency and detail our proposed Cross Non-Local Neural Network for obtaining task-specific representation. Finally, we introduce the mutual information-based constraint and the task-specific scaling.

Problem Definition and Learning Paradigm

We define the N-way M-shot few-shot classification as problem $D(\tau)$, which selects $M \times N$ labeled samples as support set $\mathbf{S} = (x_i, y_i)_{i=1}^{M \times N}$ from N classes and selects Q unlabeled samples as query set $\mathbf{Q} = (x_j, y_j)_{j=1}^{Q}$ from the same N classes in each episode. The objective of $D(\tau)$ is to predict the category of the sample in the query set $\mathbf{Q}$ based on the support set $\mathbf{S}$. The processes of training and testing have the same setting, while classes of them do not overlap. In this work, we train the model to follow a two-stage learning paradigm. Firstly, we take the classes in the training set as base classes and do an image classification with supervision for learning a universal feature extractor. Then, we train the few-shot model to recognize the novel class based on only a few labeled samples in a meta-learning scenario.

Overview of the Framework

The overview of our framework is illustrated in Figure 2. For different queries, the model aims to focus on different regions of the samples in support set via a global understanding of the current task. Hence it is not enough to only focus on the local feature map extracted by the backbone. The model needs to understand the global context of the current task. To tackle this problem, we propose to capture the long-range dependency in few-shot tasks and fuse the local and global features for understanding the task more deeply. We propose the Cross Non-Local Neural Network (CNL) for extracting the task-specific and dynamic representation. In order to prevent losing primary information during the refinement, we maximize the mutual information between the local features and refined features and take it as a constraint for training CNL. To deal with the multi-scale and task-specific features well, we add a task-specific scaling. In the following sections, we will introduce the details of our proposed method.



(a) Standard Non-local Block (b) Simplified non-local block

Figure 3: The architectures of a standard non-local block and a simplified non-local block. "⊗" denotes matrix multiplication and "⊕" denotes broadcast element-wise addition.

Capturing Long-Range Dependency

In few-shot learning, the model should have a global understanding of the samples and current tasks for extracting task-specific features and building a good feature space. Hence it is essential to capture the long-range dependency of the samples and the current task. Although RNN (Recurrent Neural Networks) can alleviate this problem, the serial computational process suffers from losing information and cannot handle the relationships over long distances. Meanwhile, CNNs concentrate on the local areas more due to their limited receptive field. And stacking convolution layers deeply for capturing long-range dependency is inefficient and hard to deliver information between distant positions. Non-local neural network (Wang et al. 2018), inherited from non-local means (Buades et al. 2005), is proposed to tackle this problem, which strengthens the features at a position via aggregating information from other positions. To better understand our approach, we revisit the non-local block in this section firstly. Then we detail the architecture of our proposed Cross Non-Local Neural Network (CNL), which aims to extract task-specific and context-aware representation dynamically.

Revisiting the Non-Local Block The non-local block (Wang et al. 2018) captures long-range dependencies directly by computing the interaction as shown in Figure 3(a), rather than relying on neighboring points in a small-window. Consider an input feature map $x \in \mathbb{R}^{C\times H\times W}$, where C, H and W is the number of channel, height and width. The non-local block is:

$$y_i \triangleq x_i + \frac{1}{\mathcal{C}(x)} \sum_{j=1}^{H\cdot W} f(x_i, x_j) g(x_j). \tag{1}$$

Here, $f(x_i, x_j) = \frac{exp(<W_q x_i, W_k x_j>)}{\sum_{m=1}^{H\cdot W} exp(<W_q x_i, W_k x_m>)}$. i is the index of an output position and j enumerates all positions. In addition, f calculates the similarity of x_i and x_j, linear transformation function $g(x_j)$ calculates the representation of the feature map at position j and $\mathcal{C}(x)$ is the normalization factor.

(Cao et al. 2019) simplifies the non-local block because they find that the global context extracted from the non-local block are almost same for different positions. Hence they compute a global attention map for all positions as shown in Figure 3 (b). This simplified non-local block is defined as,

$$y_i \triangleq x_i + W_v \sum_{j=1}^{H\cdot W} \frac{exp(W_k x_j)}{\sum_{m=1}^{H\cdot W} exp(W_k x_m)} x_j. \tag{2}$$

However, the non-local block only focuses on the global context of the samples themselves and cannot extract dynamic features based on the global information of different queries.

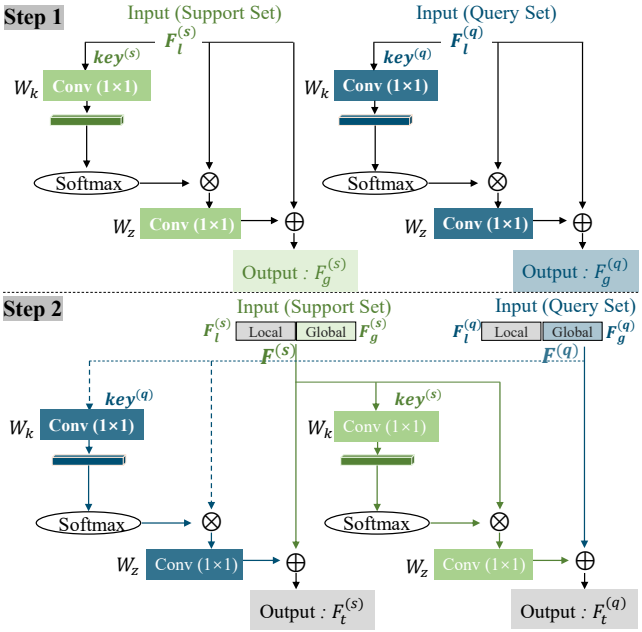


Figure 4: The architecture of Cross Non-Local Neural Network. At step one, it captures the long-range dependency for the samples in support and query set, respectively. At step two, it extracts the task-specific features.

Cross Non-Local Neural Network

For different queries in few-shot tasks, the model should pay different attention to the same samples in the support set. In addition to the global context of the sample itself, it is essential to consider the relationship between the specific query and the samples in support set for improving the performance and adaptability. To address this problem, we propose the Cross Non-Local Neural Network, as shown in Figure 4, which not only captures the long-range dependency of the sample itself and also captures the long-range dependency between the sample and the related task. It aims to strengthen the features of samples in the support set via aggregating information from all positions of themselves and the query. Likewise, we refine the features map of the query according to all the positions of themselves and the samples in the support set. CNL is divided into two steps to extract the task-specific representation.

At step one, we refine the local features of all the samples extracted by the pre-trained backbone with the global context

of themselves in Equation 3. It aims at capturing the long-range dependency for each sample, and it is task-agnostic. To avoid losing the local information, we take a concatenation operator for the feature map F_l (extracted by the backbone) and the refined feature map F_g, which is defined as F.

$$F_{g_i} = F_{l_i} + W_z \sum_{j=1}^{H \cdot W} \frac{exp(W_k F_{l_j})}{\sum_{m=1}^{H \cdot W} exp(W_k F_{l_m})} F_{l_j}. \tag{3}$$

where, W is transforming matrixes.

Secondly, to obtain the task-specific representation F_t, we augment the features of the samples in the support set via aggregating information from all positions of the current query. And we strengthen the features of the current query with the global context of the samples in the support set. It is defined as,

$$F_{t_i}^{(s)} = F_i^{(s)} + W_z \sum_{j=1}^{H \cdot W} \frac{exp(W_k F_j^{(q)})}{\sum_{m=1}^{H \cdot W} exp(W_k F_m^{(q)})} F_j^{(q)}, \tag{4a}$$

$$F_{t_i}^{(q)} = F_i^{(q)} + W_z \sum_{j=1}^{H \cdot W} \frac{exp(W_k F_j^{(s)})}{\sum_{m=1}^{H \cdot W} exp(W_k F_m^{(s)})} F_j^{(s)}. \tag{4b}$$

In this part, we make full use of local and global features. In addition to considering the relationship between distant positions, we refine the features with the global context of the current task, which is adaptive and dynamic.

MI-based Constraint for Refining

To retain more discriminative local information and reduce the redundancy in the refining process, we maximize the mutual information between the representation F_t extracted CNL and the local representation F_l extracted from the backbone based on the principle of InfoMax (Linsker 1988). The purpose is raising the dependency between F_t and F_l and making F_t contain more frequently occurring patterns in F_l. In other words, let F_t highlight the important regions while retaining common and primary information of the initial local features F_l. In addition, MI-based constraint can avoid the over-smoothing in CNL and reduce redundancy by using smaller embedding space to express rich details.

Mutual Information The MI is defines as Equation 5.

$$I(X;Z) \triangleq H(X) - H(X|Z), \tag{5a}$$

$$I(X;Z) \triangleq D_{\mathrm{KL}}(\mathbf{P}_{XZ} || \mathbf{P}_X \otimes \mathbf{P}_Z), \tag{5b}$$

$$= \sum_{x,z} p(x,z) \log \frac{p(x|z)}{p(x)} \tag{5c}$$

Here $\mathbf{P}_X$ and $\mathbf{P}_Z$ are two probability distributions corresponding to random variables X and Z, $H(.)$ is the Shannon entropy, and $H(X|Z)$ is the conditional entropy of X given Z. $D_{\mathrm{KL}}(.)$ is KL-divergence.

However, it is difficult to compute the mutual information for high-dimensional and continuous variables. Hence, we adopt the infoNCE (Oord, Li, and Vinyals 2018) as the lower bound on MI to estimate the mutual information, which is based on Noise-Contrastive Estimation (Gutmann and Hyvärinen 2010). We use it for maximizing the mutual information between the original local features F_l extracted from the pre-trained backbone and the refined task-specific representation F_t extracted by the Cross Non-Local Neural Network.

Assume set $H = \{f_{l_i}\}_{i=1}^{V}$ contains a positive and $V-1$ negative samples, the MI-based constraint is defined as,

$$L_{\mathrm{MI}} = -\mathbb{E}_H \left[\log \frac{g_d(F_l, F_t)}{\sum_{F_{l_j} \in H} g_d(F_{l_j}, F_t)} \right] \tag{6}$$

Intuitively, function $g_d(f_l, f_t) \propto \frac{p(f_l|f_t)}{p(f_l)}$ calculates whether F_l and F_t matches.

$$\begin{aligned} L_{\mathrm{MI}} &= -\mathbb{E}_H \left[\log \frac{\frac{p(F_l|F_t)}{p(F_l)}}{\frac{p(F_l|F_t)}{p(F_l)} + \sum\limits_{F_{l_j} \sim H_{Neg}} \frac{p(F_{l_j}|F_t)}{p(F_{l_j})}} \right] \\ &\geq \mathbb{E}_H \left[\log \left(V \times \frac{p(F_l)}{p(F_l|F_t)} \right) \right] = -I(F_l, F_t) + log(V). \end{aligned} \tag{7}$$

As Equation 7, we can maximize the mutual information between F_l and F_t by minimizing loss L_{MI}.

Task-Specific Scaling

The features extracted by the proposed Cross Non-Local Neural Network is multi-scale because it contains local and global features. In addition, the features of the same sample vary across different few-shot tasks. Many existing metric-based methods use the cosine similarity to measure the similarity between the samples in the support set and the query for prediction. However, it cannot measures the angle between two vectors exactly without eliminating the offsets in patterns, and it is not sensitive to the absolute value. To address this problem, we propose a task-specific scaling to deal with the multi-scale and context-aware features. The potential expression varies across different samples in different few-shot tasks. Degenerating to a simple Location-scale distribution family, it is different on translations (e.g., first-order moment, expectation) and degrees of stability (e.g., variance or second-order moment). Hence it is effective to do normalization separately based on task-specific scaling.

Firstly, we obtain the task representation r by computing the mean of all the sample features in the current task. Then we compute the mean of r as the scale for the samples, defined as μ_r. This scale is different for the same samples in the support set when the query is different. Therefore, it can deal with the multi-scale and task-specific features extracted by CNL. This task-specific scaling is defined as,

$$r = \frac{1}{M \times N + 1} (F_t^{(q)} + \sum_{i=1}^{M \times N} F_{t_i}^{(s)}), \tag{8a}$$

$$F_{t_i}^{(s)} = F_{t_i}^{(s)} - \mu_r, \quad F_{t_i}^{(q)} = F_{t_i}^{(q)} - \mu_r, \tag{8b}$$

$$\hat{y} = d(F_t^{(s)}, F_t^{(q)}) = 1 - \frac{< F_t^{(s)}, F_t^{(q)} >}{\|F_t^{(s)}\|_{\ell_2} \|F_t^{(q)}\|_{\ell_2}}. \tag{8c}$$

Here, we use the cosine similarity with task-specific scaling.

Methods	Backbone	MiniImagenet		TieredImagenet		Method Category
		1-shot	5-shot	1-shot	5-shot	
SNAIL (Nikhil et al. 2018)	*ResNet12*	55.71 ± 0.99	68.88 ± 0.92	-	-	Memory Network
TADAM (Oreshkin et al. 2018)	*ResNet12*	58.50 ± 0.30	76.70 ± 0.30	-	-	
MTL (Sun et al. 2019)	*ResNet12*	61.2 ± 1.8	75.5 ± 0.8	-	-	Meta Learning
MetaOptNet (Lee et al. 2019)	*ResNet12*	62.64 ± 0.82	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53	
LEO (Rusu et al. 2019)	*WRN-28-10*	61.76 ± 0.08	77.69 ± 0.12	66.33 ± 0.05	81.44 ± 0.09	
MatchNet(Vinyals et al. 2016)	*ResNet12*	63.08 ± 0.80	75.99 ± 0.60	68.50 ± 0.92	80.60 ± 0.71	Metric-based approaches
ProtoNet (Snell et al. 2017)	*ResNet12*	60.37 ± 0.83	78.02 ± 0.57	65.65 ± 0.92	83.40 ± 0.65	
GCR (Li et al. 2019a)	*4Conv*	53.21 ± 0.40	72.34 ± 0.32	-	-	
TapNet (Sung et al. 2019)	*ResNet12*	61.65 ± 0.15	76.36 ± 0.10	63.08 ± 0.15	80.26 ± 0.12	
CTM (Li et al. 2019b)	*ResNet18*	64.12 ± 0.82	80.51 ± 0.13	68.41 ± 0.39	84.28 ± 1.73	
DeepEMD (Zhang et al. 2020)	*ResNet12*	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58	
Ours	*ResNet12*	**67.96** ± 0.98	**83.36** ± 0.51	**73.42** ± 0.95	**87.72** ± 0.75	

Table 1: Average classification accuracy (%) with 95% confidence interval of 1000 5-way few-shot tasks on MiniImageNet and TieredImagenet datasets. Our proposed method outperforms the state-of-the-art methods on both two datasets.

Objective Function

In this paper, the objective function of our proposed framework is defined as,

$$L(\theta) = \lambda L_{\text{SIM}} + L_{\text{MI}} + \frac{\gamma}{2n}\sum_{\theta}\theta^2, \tag{9a}$$

$$L_{\text{SIM}} = \sum y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i). \tag{9b}$$

Here, L_{MI} is the MI-based constraint, L_{SIM} is the cross-entropy loss between the ground truth and the prediction, the last term in Equation 9(a) is the ℓ_2 regularization, and the tuning parameters λ and γ balance corresponding components in the loss function.

Experiments

In this section, we will introduce the implementation of our method and evaluate our proposed method, including the following parts: (i) The performances of the proposed framework on two public datasets; (ii) Whether capturing long-range dependencies can help the model construct a better feature space, and whether Cross Non-Local Neural Network can fuse the features of samples in support and query set for obtaining task-specific representation; (iii) Are the MI-based constraint and task-specific scaling effective?

Implementation Details

In this work, we adopt the *ResNet12* (He et al. 2016) as the backbone for extracting the local features. We train the model in two stages. First, we take an image classification with supervision as the pre-training task for training the *ResNet12* (He et al. 2016) with the samples of base classes. This is done by performing an AdaptiveAvgPool2d operation on the features of layer 4 in *ResNet12* and input it into a full-connection layer for classifying. Secondly, we train the proposed few-shot method based on the pre-trained backbone following the principle proposed by (Vinyals et al. 2016), where the processes of testing and the training have the same condition. We set λ as 1 in the experiments.

Dataset

Mini-Imagenet. Mini-Imagenet is a subset of ImageNet (Deng et al. 2009) for few-shot classification, proposed by (Vinyals et al. 2016). It contains 100 classes with 600 images per class, which are divided into 64, 16, 20 for training/ validation/ testing.

TieredImagenet. TieredImagenet (Ren et al. 2018) is also a subset of ImageNet but with more classes and images (779,165) compared to the Mini-Imagenet, which enlarges the domain difference between training and testing. There are 608 classes from 34 super-classes, which are divided into 20, 6, 8 for training/ validation/ testing.

Comparisons with State-of-the-Art Methods

We compare our method with state-of-the-art methods on Mini-Imagenet and TieredImagenet datasets. Our proposed algorithm achieves a consistent improvement over the other methods in 5-way 1-shot and 5-way 5-shot tasks. Comparative results are listed in Table 1. We can observe that our method achieve new state-of-the-art performances on the two datasets in both 5-way 1-shot and 5-way 5-shot setting. Compared with the latest metric-based approaches with the same backbone, our method can build better feature space with the task-specific representation. The Cross Non-Local Neural Network with the task-specific scaling and MI-based constraint is effective, especially in 1-shot tasks. The results proved that the global understanding of the current task and different attention on the same samples based on different queries could improve the model's performance.

Ablation Analysis

In this section, we evaluate the effectiveness of the key components in our method, which contains a Cross Non-Local Neural Network, task-specific scaling, and the MI-based constraint.

The detailed results in Table 2 illustrate that each critical component plays a pivotal part in our method. Firstly, we take the pre-trained *ResNet12* with cosine similarity as the baseline. Here, we take the features extracted by pre-trained

Dataset	MiniImagenet	TieredImagenet
	5-way 1-shot	
Local info	63.08 ± 0.62	69.34 ± 0.81
+Task-specific Scaling	66.61 ± 0.85	70.97 ± 0.85
Local "‖" Global info	64.20 ± 0.98	71.21 ± 0.85
+Task-specific Scaling	66.52 ± 0.98	72.22 ± 0.81
Task-Specific info	64.97 ± 0.98	72.02 ± 0.85
+Task-specific Scaling	67.83 ± 0.99	72.98 ± 0.85
+MI constraint	**67.96** ± 0.98	**73.42** ± 0.95

Table 2: Ablation experiments for evaluating the effectiveness of each component in our method, such as Cross Non-Local Neural Network, task-specific scaling, and the MI-based constraint. The local information is extracted by the pre-trained *ResNet12*, the global information is extracted by a non-local network and the task-specific information is extracted by our proposed CNL. "‖" denotes the concatenating operation.

ResNet12 as the local features of the samples, and the performance is improved significantly by adding the task-specific scaling into cosine similarity. Next, we extract the global feature by a non-local network, and then we concatenate both the local and global features to do the same task. It outperforms the baseline only with the local features. Moreover, better performance can be achieved by adding the task-specific scaling. This result shows that capturing long-range dependency is effective in few-shot tasks. At last, we extract the task-specific representation by Cross Non-Local Neural Network and adding the task-specific scaling into cosine similarity. We have achieved the the-state-of-art performances on both two datasets in the 5-way 1-shot tasks. This demonstrates Cross Non-Local Neural Network is useful for having a global understanding of the current tasks, and it can pay different attention to the same samples based on different queries. And the task-specific scaling can deal with the multi-scale and task-specific features well. In addition, the MI-based constraint is effective in the refinement process.

Effect of Cross Non-Local Neural Network We compare multiple methods for capturing long-range dependency. The fully-connected layer uses learned weights to capture the long-range dependency instead of a function used in a non-local neural network, and it needs fixed-size input. The non-local neural network only focuses on the global context of the current sample. Our proposed Cross Non-Local Neural Network can capture the long-range dependency between the current sample and other samples in its related task, which can extract dynamic and task-specific representation. The results in Table 3 demonstrate that our proposed Cross Non-Local Neural Network with task-specific scaling outperforms other methods.

Effect of Task-Specific Scaling There are some useful metric functions in few-shot learning, such as cosine similarity, Euclidean Distance, Adaptive Margin, and so on. As Table 4 shown, adding the task-specific scaling into the cosine similarity improves the performance significantly. Notably, it outperforms other metric functions and has achieved new

Method	MiniImageNet 5-way 1-shot	Input Size
Pre-trained *ResNet12* + Cosine Similarity		
Fully-connected layer	64.33± 0.77	Fixed
Bilinear	63.23± 0.76	Fixed
Dilated Convolution	59.89± 0.98	Unfixed
Non-local (Wang et al. 2018)	65.12± 0.97	Unfixed
GCNet (Cao et al. 2019)	65.84± 0.98	Unfixed
CNL with task-specific scaling	**67.83** ± 0.99	Unfixed

Table 3: Comparison of the methods for capturing long-range dependency in 5-way 1-shot setting.

Metric	MiniImageNet	
	5-way	10-way
Backbone: Pre-trained *ResNet12*		
Euclidean(Snell et al. 2017)	60.08	47.09
Dot (Chen et al. 2019)	59.41	44.08
EMD(Zhang et al. 2020)	65.91	49.66
Adjusted Cosine	65.75	50.74
Cosine (Vinyals et al. 2016)	63.08	44.34
+Task-specific scaling	**66.61**	**51.13**

Table 4: Metric comparison for 5-way 1-shot and 10-way 1-shot classification on MiniImagenet. These methods use the same pre-trained *ResNet12* to extract local features for building feature space.

the-state-of-the-art performance in 5-way 1-shot tasks on the MiniImageNet dataset. After further analysis, the reason is that most of the existing methods focus on dealing with the features with the same scale and range. Adding task-specific scaling can deal with multi-scale and task-specific features better. In addition, this scaling improves performance without higher computational complexity, and it also can be added to other metric functions.

Conclusion

Capturing the long-range dependency and extracting task-specific representation can improve the performances of few-shot algorithms. Our proposed Cross Non-Local Neural Network with task-specific scaling and MI-based constraint can help the model focus on different regions of the same images based on different queries. The task-specific scaling can deal with the multi-scale and task-specific features well, which is effective and also can be used in other metric-based approaches. Comprehensive experimental results demonstrate that our method has achieved new state-of-the-art performances, and each essential part plays a vital role. In the future, we plan to apply our method in other tasks, such as visual question answering and image retrieval.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2018AAA0100503 & 2018AAA0100500),

the National Natural Science Foundation of China (No. 61773167), the Science and Technology Commission of Shanghai Municipality (No. 19511120200 & 18DZ2270800), and the Open Fund of PDL.

References

Alfassy, A.; Karlinsky, L.; Aides, A.; Shtok, J.; Harary, S.; Feris, R.; Giryes, R.; and Bronstein, A. M. 2019. Laso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6548–6557.

Allen, K. R.; Shelhamer, E.; Shin, H.; and Tenenbaum, J. B. 2019. Infinite Mixture s for Few-Shot Learning. *arXiv preprint arXiv:1902.04552* .

Bertinetto, L.; Henriques, J. F.; Valmadre, J.; Torr, P.; and Vedaldi, A. 2016. Learning feed-forward one-shot learners. In *Advances in neural information processing systems*, 523–531.

Buades; Antoni; Coll, B.; and Morel, J. M. 2005. A non-local algorithm for image denoising. In *IEEE Computer Society Conference on Computer Vision & Pattern Recognition*.

Buades, A.; Coll, B.; and Morel, J.-M. 2005. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, 60–65. IEEE.

Cao, Y.; Xu, J.; Lin, S.; Wei, F.; and Hu, H. 2019. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 0–0.

Chelsea; Finn; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 1126–1135.

Chen, W.; Liu, Y.; Kira, Z.; Wang, Y. F.; and Huang, J. 2019. A Closer Look at Few-shot Classification. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. *International Conference on Machine Learning* .

Geng, R.; Li, B.; Li, Y.; Sun, J.; and Zhu, X. 2020. Dynamic Memory Induction Networks for Few-Shot Text Classification. *Association for Computational Linguistics (ACL)* .

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 297–304.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2019. Cross Attention Network for Few-shot Classification. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 4005–4016.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Jamal, M. A.; and Qi, G.-J. 2019. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11719–11727.

Kim, J.; Kim, T.; Kim, S.; and Yoo, C. D. 2019. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11–20.

Krähenbühl, P.; and Koltun, V. 2011. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, 109–117.

Krizhevsky; Alex; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10657–10665.

Li, A.; Huang, W.; Lan, X.; Feng, J.; Li, Z.; and Wang, L. 2020. Boosting Few-Shot Learning With Adaptive Margin Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12576–12584.

Li, A.; Luo, T.; Xiang, T.; Huang, W.; and Wang, L. 2019a. Few-Shot Learning With Global Class Representations. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 9714–9723.

Li, H.; Eigen, D.; Dodge, S.; Zeiler, M.; and Wang, X. 2019b. Finding Task-Relevant Features for Few-Shot Learning by Category Traversal. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 1–10.

Linsker, R. 1988. Self-organization in a perceptual network. *Computer* 21(3): 105–117.

Liu, W.; Rabinovich, A.; and Berg, A. C. 2015. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579* .

Nikhil; Mishra; Rohaninejad, M.; Chen, X.; and Abbeel, P. 2018. A Simple Neural Attentive Meta-Learner. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* .

Oreshkin; Boris; López, P. R.; and Lacoste, A. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, 721–731.

Ravi, S.; and Larochelle, H. 2017. Optimization as a Model for Few-Shot Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-Learning for Semi-Supervised Few-Shot Classification. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Rusu, A. A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; and Hadsell, R. 2019. Meta-Learning with Latent Embedding Optimization. In *International Conference on Learning Representations*.

Snell; Jake; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 4077–4087.

Sun, Q.; Liu, Y.; Chua, T.-S.; and Schiele, B. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 403–412.

Sung; Yoon, W.; Seo, J.; and Moon, J. 2019. TapNet: Neural Network Augmented with Task-Adaptive Projection for Few-Shot Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 7115–7123.

Triantafillou; Eleni; Zemel, R.; and Urtasun, R. 2017. Few-shot learning through an information retrieval lens. In *Advances in Neural Information Processing Systems*, 2255–2265.

Tsutsui; Satoshi; Fu, Y.; and Crandall, D. 2019. Meta-Reinforced Synthetic Data for One-Shot Fine-Grained Visual Recognition. In *Advances in Neural Information Processing Systems*, 3057–3066.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vinyals; Oriol; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, 3630–3638.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-Shot Image Classification with Differentiable Earth Mover's Distance and Structured Classifiers. In *Proceedings of the IEEECVF Conference on Computer Vision and Pattern Recognition*, 12203–12213.

Zhang, R.; Che, T.; Ghahramani, Z.; Bengio, Y.; and Song, Y. 2018. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, 2365–2374.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.