Few-shot Visual Learning with Contextual Memory and Fine-grained Calibration

Yuqing Ma[1] , **Wei Liu**[1] , **Shihao Bai**[1] , **Qingyu Zhang**[1] , **Aishan Liu**[1] ,
Weimin Chen[3] and **Xianglong Liu**[*1,2]
[1]State Key Lab of Software Development Environment, Beihang University, China
[2]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, China
[3]NetEase Fuxi AI Lab, Hangzhou, China
{mayuqing, 16061175, 16061167, zqy1723, liuaishan}@buaa.edu.cn,
chenweimin@corp.netease.com, xlliu@nlsde.buaa.edu.cn

Abstract

Few-shot learning aims to learn a model that can be readily adapted to new unseen classes (concepts) by accessing one or few examples. Despite the successful progress, most of the few-shot learning approaches, concentrating on either global or local characteristics of examples, still suffer from weak generalization abilities. Inspired by the inverted pyramid theory, to address this problem, we propose an inverted pyramid network (IPN) that intimates the human's coarse-to-fine cognition paradigm. The proposed IPN consists of two consecutive stages, namely global stage and local stage. At the global stage, a class-sensitive contextual memory network (CCMNet) is introduced to learn discriminative support-query relation embeddings and predict the query-to-class similarity based on the contextual memory. Then at the local stage, a fine-grained calibration is further appended to complement the coarse relation embeddings, targeting more precise query-to-class similarity evaluation. To the best of our knowledge, IPN is the first work that simultaneously integrates both global and local characteristics in few-shot learning, approximately imitating the human cognition mechanism. Our extensive experiments on multiple benchmark datasets demonstrate the superiority of IPN, compared to a number of state-of-the-art approaches.

1 Introduction

Deep learning methods have shown the powerful learning capability in the past decade. The standard deep learning models [Ji *et al.*, 2013; Zagoruyko and Komodakis, 2016] mainly contain millions of parameters and heavily rely on the huge amount of training data. Largely different from the deep learning model, the human cognition system exhibits remarkable abilities to infer the novel concepts effortlessly from only one or a few examples and reliably recognize them later on. To acquire the similar strong generalization ability, few-shot learning has been introduced to learn a model that can be readily adapted to new unseen classes (concepts) by accessing only one or few examples. A variety of few-shot learning methods have been proposed recently, which can be roughly divided into optimization-based, generation-based, and metric-based methods.

Optimization-based methods train a desired meta learner over a variety of learning tasks and optimize it for the best performance on a distribution of tasks, including potentially unseen tasks. To accomplish this task, usually there needs an across-task meta-learner that identifies how to update the parameters of the learner's model. [Rusu *et al.*, 2018] introduced a data-dependent meta-learning approach which learns a low-dimensional latent generative representation of model parameters, and performs gradient-based adaptation in this space. In [Sun *et al.*, 2019], a novel meta-transfer learning method is proposed which combines the advantages of meta learning and transfer learning to transfer large scale pre-trained DNN weights for solving few-shot learning tasks.

Generation-based methods attempt to augment few-shot data with a generative meta-learner or learn to predict classification weights for novel classes. [Qiao *et al.*, 2018] proposed a novel approach that can adapt a pre-trained neural network to novel categories by directly predicting the parameters from the activations without training. In [Gidaris and Komodakis, 2019], a Denoising Autoencoder network is used to refine a set of initial classification weights to make them more discriminative with respect to the classification task at hand.

Metric-based methods have achieved considerable success by learning to compare the support and query samples in a shared feature space. The early study of [Vinyals *et al.*, 2016] introduced the episodic training mechanism into few-shot learning and utilized a bidirectional LSTM to encode each support sample in the context of the whole support set, and matched the query sample to the support sample through an attention mechanism. The following typical methods such as [Sung *et al.*, 2018] attempted to embed the samples by simply summing each support class in an element-wise manner.

Most of the previous few-shot learning approaches concentrated on the abstract global information for each sample. This is consistent with the human cognition system, which usually first makes a coarse recognition from the global perspective, *e.g.*, shape, size, structure, *etc*. However, in practice when the object is too difficult to be distinguished from the global perspective, humans will further resort to the detailed local features. This recognition scheme follows the coarse-

*Corresponding Author

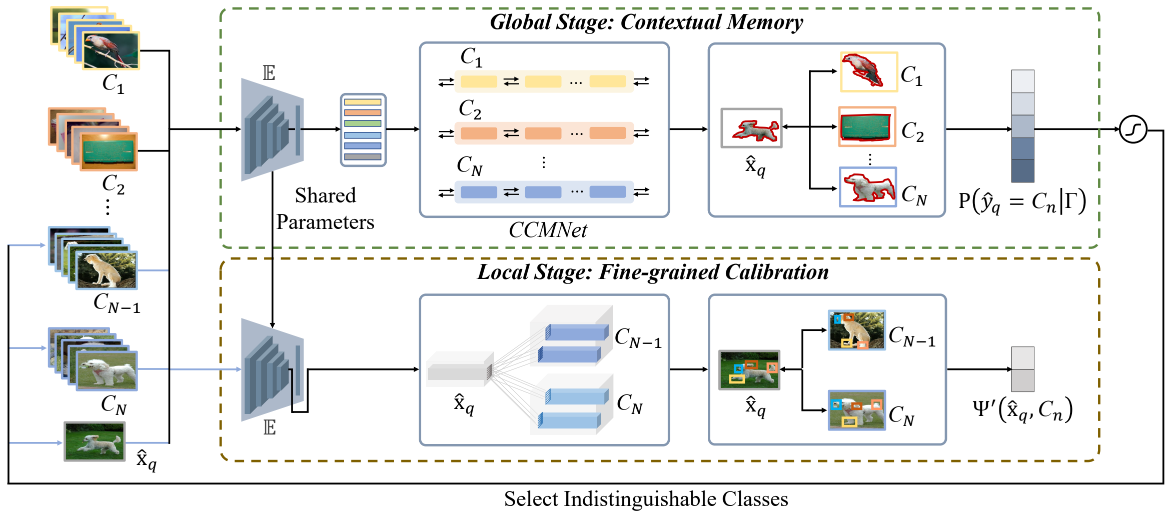

Figure 1: The IPN framework.

to-fine theory of visual perception, inspired by the Gestalt dictum that the whole is prior to the parts (the whole arises before the parts). To compensate for the weakness of the previous methods, recently a few studies [Li *et al.*, 2019b; Li *et al.*, 2019c] have been proposed to extract the subtle difference from the local perspective, and achieved satisfying performance on fine-grained dataset. Unfortunately, only focusing on either global or local characteristics, these methods are still far from strong generalization abilities over datasets with various classes.

To address this problem, in this paper we propose an inverted pyramid network (IPN) to intimate the human's coarse-to-fine cognition paradigm, inspired by the reverse-hierarchy theory [Ahissar and Hochstein, 2004]. Reverse-hierarchy theory, also known as inverse pyramid theory, suggests that the magnocellular stream provides the fast "coarse" initial sweep, while slow parvocellular signals representing "fine" analysis are processed in a later time window. Following this paradigm, the proposed IPN consists of two consecutive stages, namely global stage and local stage. At the global stage, we propose a class-specific contextual model with a memory mechanism (CCMNet) to learn the discriminative global support-query relation embeddings. Specifically, CCMNet sequentially processes the query sample and one support sample of a specific class at each time step, and learns the discriminative relation embedding between the support and query sample based on the contextual information. Besides, the information flow of the classical GRU is further modified to preserve the long-term dependencies using fewer parameters, enabling the strong sensitivity to the contextual information. At the local stage, to compensate for the weakness of the globally predicted query-to-class similarity, the fine-grained calibration can be further appended by simply comparing the query with its nearest patches, and thus targets more precise query-to-class similarity evaluation.

To the best of our knowledge, IPN is the first work that simultaneously integrates both global and local characteristics in few-shot learning, approximately imitating the human cognition mechanism. Extensive experiments conducted on two commonly-used few-shot datasets *mini*ImageNet and *tiered*ImageNet further verify the superiority of our IPN model. Especially, even using our CCMNet alone can achieve 66% 1-shot accuracy and nearly 83% 5-shot accuracy on *mini*ImageNet, outperforming most state-of-the-art approaches under the same setting. Moreover, by further applying the fine-grained calibration, our two-stage framework can consistently obtain accuracy gains (up to 5.6%), on different datasets and under different few-shot settings.

2 The Inverted Pyramid Network

We develop a novel two-stage few-shot learning architecture named Inverse Pyramid Network (IPN), inspired by the coarse-to-fine theory of human visual perception, meaning that the whole arises before the parts. Therefore, the proposed IPN consists of two consecutive stages, namely global stage and local stage. At the global stage, a class-sensitive contextual memory network is proposed to progressively capture the global relations such as the similarities of shape, structure *etc.*, between support samples from a class and query in an online setting. After that, fine-grained calibration will be conducted to further compare the local discriminative parts for indistinguishable classes.

Next, we first introduce the preliminary of the few-shot setting, then present the Class-sensitive Contextual Memory Network at the global stage and the fine-grained calibration at the local stage, and finally demonstrate the inference process of the proposed model on novel classes.

2.1 Preliminary

Let $\mathcal{S}$ denote a support set, which contains N different image classes $(\mathcal{C}_1, \ldots, \mathcal{C}_N)$ and K (K is small, *e.g.*, $K = 5$) labeled samples per class. Given a query set $\mathcal{Q}$, few-shot learning aims to classify each unlabeled sample in $\mathcal{Q}$ according to the set $\mathcal{S}$. This setting is also called N-way K-shot classification. We adopt episodic training which is commonly em-

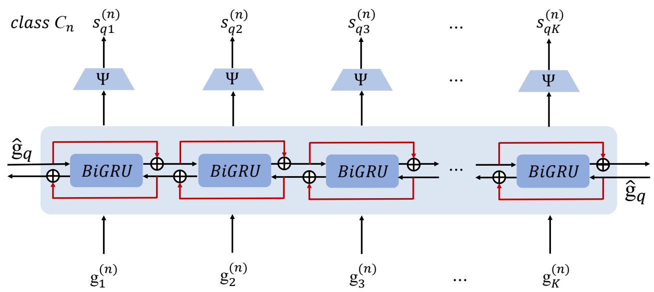

Figure 2: Class-sensitive Contextual Memory Network. The red arrow represents the skip link.

ployed in the literature as an effective approach to learn the transferable knowledge from a relatively large labeled training dataset with a set of classes $\mathcal{C}_{train}$ which has a disjoint class label space with the test dataset with novel classes $\mathcal{C}_{test}$, namely $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \varnothing$.

More concretely, in episodic training, a small subset of N classes are sampled from $\mathcal{C}_{train}$ to construct an N-way K-shot problem as follows: a task Γ contains a support set $\mathcal{S}$ and a query set $\mathcal{Q}$, where $\mathcal{S} = \{\mathbf{x}_k^{(1)}\}_{k=1}^K \cup ... \cup \{\mathbf{x}_k^{(N)}\}_{k=1}^K$ and $\mathcal{Q} = \{(\hat{\mathbf{x}}_q, \hat{y}_q)\}_{q=1}^T$. Here, $\mathbf{x}_k^{(n)}$ denotes the k-th sample of class $\mathcal{C}_n$ in the support set. T is the number of query samples, and $\hat{\mathbf{x}}_q$, $\hat{y}_q \in \{\mathcal{C}_1, ..., \mathcal{C}_N\}$ are the q-th query data and its ground truth label, respectively. In each episode, the model is trained to minimize the loss of its predictions of $\mathcal{Q}$ through learning the labeled support set $\mathcal{S}$.

2.2 Global Stage: Contextual Memory

Given a sample $\mathbf{x}$ of a task Γ, the embedding network $\mathbb{E}$ is first utilized to extract the global feature representations $\mathbf{g} = \mathbb{E}(\mathbf{x}) \in \mathbf{R}^{1\times D}$. Based on the feature representations, we propose the class-sensitive contextual memory network (CCMNet) to capture the query-to-class relations, by fully exploring the context of a class. Initializing the hidden state with query features, CCMNet takes a support sample from a specific class as input and captures the support-query relation at each time step. Through learning relation embeddings sample by sample in a class-sensitive contextual environment, CCMNet enjoys a better understanding about the query-to-class similarity and thus achieves comparatively accurate classification performance.

Class-sensitive Contextual Memory Network

For the proposed CCMNet, we adopt the classical GRU [Cho *et al.*, 2014] and modify its information flow to selectively absorb the information from past experience. GRU is a recurrent gating mechanism, where the reset gate mainly determines how much of the past information needs to be removed, and the update gate determines how much of the past information (from previous time steps) needs to be memorized and passed to the future.

We further modify the information flow of classical GRU to fit the task of relation learning from the contextual information. For one thing, we initialize the hidden state with the query representations, and feed the global feature representation from the same class into the memory network at every time step. Thus, we capture the relation between the query sample and support sample at each step and make the prediction given the class-sensitive contextual environment. For another, we add a skip link in each time step to further consider experience from the previous two time steps, in avoid of catastrophic forgetting and to better learn the contextual information across the class.

Figure 2 shows the whole structure of CCMNet with the modified GRU module. Specifically, when learning the relations between the query sample $\hat{\mathbf{x}}_q$ and the n-th class, we initialize the hidden state $\mathbf{h}_0^{(n)} = \hat{\mathbf{g}}_q$. And in the time step k, we feed the k-th sample denoted as $\mathbf{g}_k^{(n)}$ from class $\mathcal{C}_n$ into the CCMNet, and the hidden state updating in our class-sensitive GRU is conducted as follows:

$$\mathbf{z}_k^{(n)} = \sigma(\mathbf{W}_z\mathbf{g}_k^{(n)} + \mathbf{U}_z(\mathbf{h}_{k-1}^{(n)} + \mathbf{h}_{k-2}^{(n)})) \quad (1)$$

$$\mathbf{r}_k^{(n)} = \sigma(\mathbf{W}_r\mathbf{g}_k^{(n)} + \mathbf{U}_r(\mathbf{h}_{k-1}^{(n)} + \mathbf{h}_{k-2}^{(n)})) \quad (2)$$

$$\tilde{\mathbf{h}}_k^{(n)} = \phi(\mathbf{W}_h\mathbf{g}_k^{(n)} + \mathbf{U}_h(\mathbf{r}_k^{(n)} \odot (\mathbf{h}_{k-1}^{(n)} + \mathbf{h}_{k-2}^{(n)}))) \quad (3)$$

$$\mathbf{h}_k^{(n)} = \mathbf{z}_k^{(n)} \odot (\mathbf{h}_{k-1}^{(n)} + \mathbf{h}_{k-2}^{(n)}) + (1 - \mathbf{z}_k^{(n)}) \odot \tilde{\mathbf{h}}_k^{(n)} \quad (4)$$

where $\mathbf{h}_k^{(n)}$ is the updated hidden state, with $\mathbf{h}_{k-2}^{(n)}$ a skip link to $\mathbf{h}_{k-1}^{(n)}$, considering previous two time steps to involve more class-sensitive contextual information. When $k = 1$, we denote $\mathbf{h}_{k-2}^{(n)} = \hat{\mathbf{g}}_q$. $\mathbf{W}_z$, $\mathbf{U}_z$, $\mathbf{W}_r$, $\mathbf{U}_r$, $\mathbf{W}_h$, and $\mathbf{U}_h$ are all learnable parameters, σ and ϕ are the sigmoid activation function and the tanh activation function, respectively. $\mathbf{z}_k^{(n)}$ and $\mathbf{r}_k^{(n)}$ represent the update gate and the reset gate.

The entire iterations explore the relations between $\hat{\mathbf{g}}_q$ and the category $\mathcal{C}_n$ by traversing all samples of $\mathcal{C}_n$ in the support set. Due to the special gating update mechanism, the hidden state $\mathbf{h}_k^{(n)}$ after iterative update retains the common features of the query sample and category $\mathcal{C}_n$, while irrelevant interference information is forgotten. Therefore, $\mathbf{h}_k^{(n)}$ can be used as a relation embedding to measure the query-to-class similarity. Besides, through the skip link, not only the gradient vanishing problem in the back-propagation procedure is alleviated, but also effectively mitigates the occurrence of catastrophic forgetting of earlier data and transmits more information to the current step.

To further learn the contextual information and eliminate the influence of sequence order, we adopt the bidirectional mechanism. We concatenate the output hidden states $\overrightarrow{\mathbf{h}}_k^{(n)}$, $\overleftarrow{\mathbf{h}}_{K-k+1}^{(n)}$ from two opposite directions together as the final relation embedding:

$$\overline{\mathbf{h}}_k^{(n)} = [\overrightarrow{\mathbf{h}}_k^{(n)}, \overleftarrow{\mathbf{h}}_{K-k+1}^{(n)}] \quad (5)$$

where $[\cdot, \cdot]$ denotes the concatenation operation. As a result, for the n-th class, we can obtain a set of relation embeddings $\{\overline{\mathbf{h}}_k^{(n)}\}_{k=1}^K$.

Learning at the Global Stage

For each relation embedding, we could learn a similarity score $\mathbf{s}_{qk}^{(n)} = \Psi(\overline{\mathbf{h}}_k^{(n)})$ where Ψ is a similarity measure, and $\mathbf{s}_{qk}^{(n)} \in [0, 1]$. Intuitively, the larger the score is, the higher the probability that they belong to the same category is. Thus, for each training episode, we could compare the similarity score with the ground truth label and compute the loss function:

$$\ell = \sum_{q=1}^{T}\sum_{n=1}^{N}\sum_{k=1}^{K} \delta_{qk} \log \mathbf{s}_{qk}^{(n)} + (1-\delta_{qk})\log(1-\mathbf{s}_{qk}^{(n)}) \quad (6)$$

where δ_{qk} is defined as:

$$\delta_{qk} = \begin{cases} 1, & \text{if} \quad \hat{y}_q = \mathcal{C}_n \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

2.3 Local Stage: Fine-grained Calibration

As mentioned above, it is hard for the model to distinguish similar categories given the compact global representations. Therefore, we need to compare the different local details to calibrate the uncertain global prediction. Under such consideration, we reuse the backbone network $\mathbb{E}$ to further gain a better understanding of the input samples and mine the distinguishable characteristics from local perspective.

Here we adopt the method presented in DN4 [Li *et al.*, 2019b]. Given a sample $\mathbf{x}$, we view the features output by the last convolutional layer of $\mathbb{E}$, as a set of local patch features $[\mathbf{p}_1, \cdots, \mathbf{p}_M]$ where $\mathbf{p}_j$ is the j-th local patch feature. For each local feature $\mathbf{p}_j$ of query sample, we find its L-nearest neighbors $\mathbf{p'}_j^l|_{l=1}^L$ in the local feature space of support samples from the same class $\mathcal{C}_n$. Then we calculate the cosine similarity between $\mathbf{p}_j$ and each $\mathbf{p'}_j$, and sum the $M \times L$ similarities as the query-to-class similarity:

$$\Psi'(\hat{\mathbf{x}}_q, \mathcal{C}_n) = \sum_{j=1}^{M}\sum_{l=1}^{L} \cos(\mathbf{p}_j, \mathbf{p'}_j^l) \quad (8)$$

$$\cos(\mathbf{p}_j, \mathbf{p'}_j^l) = \frac{\mathbf{p}_j^\top \mathbf{p'}_j^l}{\|\mathbf{p}_j\| \cdot \|\mathbf{p'}_j^l\|} \quad (9)$$

Note that the fine-grained calibration process is non-parametric and computation effective. Following the reverse-hierarchy cognition paradigm, it is very simple and flexible to be appended to the global stage on demand, forming an effective two-stage coarse-to-fine few-shot learning framework, *i.e.*, our Inverse Pyramid Network (IPN) model.

2.4 Inference on Novel Class

In the testing stage, given a query sample $\hat{\mathbf{x}}_q$, we first use the CCMNet to generate relation embeddings and compute the similarity scores for support-query pairs. Then the global query-to-class similarity can be expressed as:

$$\mathrm{P}(\hat{y}_q = \mathcal{C}_n|\Gamma) = \frac{\exp\left(\sum_k \mathbf{s}_{qk}^{(n)}\right)}{\sum_{n'}\left(\exp\left(\sum_k \mathbf{s}_{qk}^{(n')}\right)\right)} \quad (10)$$

In practice some classes of the task are too similar and indistinguishable from global perspective, and thus the global query-to-class similarities are very close. In this case, we should further conduct the fine-grained calibration. Specifically, assuming that $\mathcal{C}_i$ and $\mathcal{C}_j$ are the two categories with the highest query-to-class similarities, we calculate the prediction reliability τ of the task as:

$$\tau = \mathrm{P}(\hat{y}_q = \mathcal{C}_i|\Gamma)/\mathrm{P}(\hat{y}_q = \mathcal{C}_j|\Gamma) \quad (11)$$

We set a reliability threshold τ_0 and compare it with the prediction reliability τ. If $\tau \geq \tau_0$, we consider $\mathcal{C}_i$ as the final prediction of the query sample directly. Otherwise we resort to the fine-grained calibration to further obtain the more precise query-to-class similarity through which we make the final prediction.

3 Experiments

In this section, we evaluate our IPN with state-of-the-art few-shot approaches on widely used datasets.

3.1 Experimental Settings

Datasets

We employ the widely used datasets in prior studies, including *mini*ImageNet dataset [Vinyals *et al.*, 2016] and *tiered*ImageNet dataset [Ren *et al.*, 2018]. The *mini*ImageNet dataset consists of 100 classes, each of which contains 600 images of size 84×84, while the *tiered*ImageNet contains 608 classes with 77915 images in total. The classes of *tiered*ImageNet are grouped into 34 higher-level nodes based on WordNet hierarchy [Deng *et al.*, 2009], and is further partitioned into disjoint sets of training, testing, and validation nodes, ensuring a distinct distance between training and testing classes thus making the classification more challenging. For both datasets, we adopt the common splits as previous work.

Model Architectures

We use the recently common-used feature embedding architecture WRN-28 [Zagoruyko and Komodakis, 2016] as backbone. WRN-28 whose output is a 640-dimensional feature vector is a 28-layer wide residual network with width factor 10. We pre-train the WRN-28 network by optimizing the accuracy of the multi-classes classification on the whole training set of *mini*ImageNet or *tiered*ImageNet, and then freeze the parameters during the training phase. The similarity measure Ψ is implemented as Multi-Layer Perceptions (MLPs) consisting of 3 fully-connected layers.

Implementation Details

Standard data augmentations including random crop, left-right flip, and color jitter are applied in the training stage. The mini-batch size for all experiments is 20. The number of training iterations on *mini*ImageNet and *tiered*ImageNet are 100K and 200K. We use the validation set to select the training episodes with the best accuracy. We use Adam optimizer with an initial learning rate of 0.001, and reduce the learning rate by half every 15K and 30K iterations, respectively on *mini*ImageNet and *tiered*ImageNet. The weight decay is set to 10^{-6}. When conducting finegrained calibration at local stage, the prediction reliability threshold τ_0 is set to 1.5, and the number of nearest neighbors L is set to 3. As presented in [Kim *et al.*, 2019; Liu *et al.*, 2018], most few-shot approaches adopted two kinds of transductive inference methods to improve the classification performance. In our CCMNet, we

Models	Backbone	1-shot	5-shot
Optimization-based			
mLSTM [Ravi and Larochelle, 2017]	Conv4	43.44 ± 0.77	60.60 ± 0.71
MAML [Finn *et al.*, 2017]	Conv4	48.70 ± 1.84	63.10 ± 0.92
Meta-SGD [Li *et al.*, 2017]	Conv4	50.47 ± 1.87	64.03 ± 0.94
SNAIL [Mishra *et al.*, 2017]	ResNet-12	55.71± 0.99	68.88 ± 0.92
REPTILE [Nichol *et al.*, 2018]	Conv4	49.97 ± 0.32	65.99 ± 0.58
LEO [Rusu *et al.*, 2018]	WRN-28	61.76 ± 0.08	77.59 ± 0.12
MTL [Sun *et al.*, 2019]	ResNet-12	61.20 ± 1.80	75.50 ± 0.80
Generation-based			
PLATIPUS [Finn *et al.*, 2018]	Conv4	50.13 ± 1.86	-
VERSA [Gordon *et al.*, 2018]	Conv4	53.40 ± 1.82	67.37 ± 0.86
LwoF [Gidaris and Komodakis, 2018]	ResNet	55.45 ± 0.89	70.13 ± 0.68
Param_Predict [Qiao *et al.*, 2018]	WRN-28	59.60 ± 0.41	73.74 ± 0.19
wDAE [Gidaris and Komodakis, 2019]	WRN-28	61.07 ± 0.15	76.75 ± 0.11
Metric-based			
Matching Net [Vinyals *et al.*, 2016]	Conv4	43.56 ± 0.84	55.31 ± 0.73
GNN [Garcia and Bruna, 2017]	Conv4	50.33 ± 0.36	66.41 ± 0.63
Prototypical Net [Snell *et al.*, 2017]	Conv4	49.42 ± 0.78	68.20 ± 0.66
Relation Net [Sung *et al.*, 2018]	Conv4	50.40 ± 0.80	65.30 ± 0.70
TPN [Liu *et al.*, 2018]	Conv4	53.75 ± 0.86	69.43 ± 0.67
TADAM [Oreshkin *et al.*, 2018]	ResNet-12	58.50 ± 0.30	76.70 ± 0.30
CovaMNet [Li *et al.*, 2019c]	Conv4	51.19 ± 0.76	67.65 ± 0.63
DN4 [Li *et al.*, 2019b]	Conv4	51.24 ± 0.74	71.02 ± 0.64
EGNN [Kim *et al.*, 2019]	Conv4	-	76.37 ± 0.30
TapNet [Yoon *et al.*, 2019]	ResNet-12	61.65 ± 0.15	76.36 ± 0.10
CTM [Li *et al.*, 2019a]	ResNet-18	62.05 ± 0.55	78.63 ± 0.06
CCMNet	WRN-28	**66.30 ± 0.48**	**82.89 ± 0.39**
Ours	WRN-28	**67.42 ± 0.45**	**83.98 ± 0.35**

Table 1: Few-shot image classification accuracies of 5-way 1-shot and 5-shot tasks on *mini*ImageNet.

simply concatenate the query features together as the initialization of the hidden state and learn the relation embeddings simultaneously.

Evaluation Protocols

On both datasets, we conduct 5-way 1-shot and 5-shot experiments which are standard few-shot learning settings. For evaluation, each episode was formed by randomly sampling 1 query for each of 5 classes. We report the mean accuracy (%) of 10000 randomly generated episodes as well as the 95% intervals on test set.

3.2 Comparison with State-of-the-Art

We first investigate the performance of our model, compared to state-of-the-art few-shot approaches, respectively on *mini*ImageNet and *tiered*ImageNet. These approaches are particularly divided into optimization-based, generation-based and metric-based. Table 1 and Table 2 list the few-shot classification accuracies of the 5-way 1-shot and 5-shot tasks along with the specifications of the backbone embedding models for feature extraction, where “Conv-4” indicates the 4-layer convolutional neural network. We use bold fonts for the two best results.

From Table 1, we can observe that as the shots increase, all the methods perform better, which is adhere to our intuition. Moreover, a deeper embedding network will lead to a better classification performance compared to methods equipped with Conv4, achieving above 70% accuracy on 5-shot setting. In most cases, our IPN model significantly outperforms others under the same experimental setting, achieving 67.42% and 83.98% accuracy on 5-way 1-shot and 5-shot setting, respectively. Though as presented in [Chen *et al.*, 2019], as the backbone gets deeper, the gap among different methods drastically reduces, our IPN model can consistently gain nearly 5% improvements on 5-shot setting over the second best approach CTM, confirming the superiority of IPN mainly owning to the inverted pyramid paradigm. The similar trend can also be observed in Table 2. The proposed IPN shows comparable results with the state-of-the-arts, achieving 73.18% and 86.59% accuracy on 5-way 1-shot and 5-shot setting, respectively.

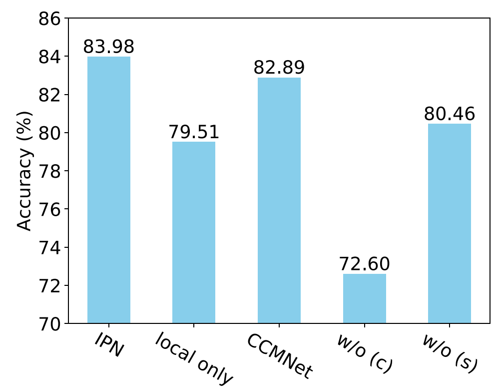

Figure 3: Ablation study on *mini*ImageNet. It shows few-shot classification results of the proposed IPN, IPN without the global stage (denoted local only), the proposed CCMNet, CCMNet without contextual information, and CCMNet without skip links.

3.3 Ablation Study

Figure 3 demonstrates the effects of each component of the proposed IPN framework on *mini*ImageNet. It respectively shows the few-shot classification results of the proposed IPN,

Models	Backbone	1-shot	5-shot
Optimization-based			
MAML [Finn *et al.*, 2017]	Conv4	51.67 ± 1.81	70.30 ± 0.08
Meta-SGD [Li *et al.*, 2017]	Conv4	62.95 ± 0.03	79.34 ± 0.06
REPTILE [Nichol *et al.*, 2018]	Conv4	52.36 ± 0.23	71.03 ± 0.22
LEO [Rusu *et al.*, 2018]	WRN-28	66.33 ± 0.05	81.44 ± 0.09
Generation-based			
LwoF [Gidaris and Komodakis, 2018]	Conv4	50.90 ± 0.46	66.69 ± 0.36
wDAE [Gidaris and Komodakis, 2019]	WRN-28	$\mathbf{68.18 \pm 0.16}$	$\mathbf{83.09 \pm 0.12}$
Metric-based			
Matching Net [Vinyals *et al.*, 2016]	Conv4	54.02 ± 0.00	70.11 ± 0.00
GNN [Garcia and Bruna, 2017]	Conv4	43.56 ± 0.84	55.31 ± 0.73
Prototypical Net [Snell *et al.*, 2017]	Conv4	53.31 ± 0.89	72.69 ± 0.74
Relation Net [Sung *et al.*, 2018]	Conv4	54.48 ± 0.93	71.32 ± 0.70
TPN [Liu *et al.*, 2018]	Conv4	57.53 ± 0.96	72.85 ± 0.74
EGNN [Kim *et al.*, 2019]	Conv4	-	80.15 ± 0.30
TapNet [Yoon *et al.*, 2019]	ResNet-12	63.08 ± 0.15	80.26 ± 0.12
CTM [Li *et al.*, 2019a]	ResNet-18	64.78 ± 0.11	81.05 ± 0.13
CCMNet	WRN-28	67.54 ± 0.50	82.40 ± 0.31
Ours	WRN-28	$\mathbf{73.18 \pm 0.43}$	$\mathbf{86.59 \pm 0.33}$

Table 2: Few-shot image classification accuracies of 5-way 1-shot and 5-shot tasks on *tiered*Imagenet.

IPN without the global stage (denoted local only), the proposed CCMNet, CCMNet without contextual information, and CCMNet without skip links. Since we directly adopt the DN4 model proposed in [Li *et al.*, 2019b] to compare local fine-grained details at our local stage, we simply replace the backbone of DN4 model with ours and re-train the model under the same setting. As for removing contextual information from CCMNet, at each time step, we assign previous hidden state $\mathbf{h}_{k-1}^{(n)}, \mathbf{h}_{k-2}^{(n)}$ with query features and thus capture the query-to-class relations without passing contextual messages.

As it could be seen from Figure 3, only considering global characteristics or local ones, the performance decline by nearly 1% and 4%, respectively. As shown in Table 1 and Table 2, comparing the performance of CCMNet and our full model, accuracy gains of fine-grained calibration is comparatively obvious on the more challenging *tiered*ImageNet with various classes and disjoint higher-level semantic hierarchy, confirming that fine-grained calibration is effective and significative in real-world scenarios. We further investigate the effectiveness of each part of CCMNet. Without contextual information, the performance significantly drops by 10%. Without the skip link, it witnesses more than 2% decline. In contrast, with contextual information and skip link, the proposed CCMNet achieves 66.30% 1-shot accuracy and 82.89% 5-shot accuracy on *mini*ImageNet, which outperforms most few-shot approaches. Thus, we can conclude that, with help of the CCMNet at global stage and fine-grained calibration at local stage, our IPN framework enjoys strong power to achieve the best performance compared to others.

Figure 4 shows t-SNE visualizations of relation embeddings for the proposed IPN. The model is trained under 10-way 5-shot. Circles indicate the relation embeddings of a query sample and numerous support samples from 10 classes. Different colors denote different class labels of support samples. Intuitively, if two support samples are similar, they share the similar relation with the same query sample. As the Figure 4 depicted, there are obvious 10 clusters of relations. Each cluster is compact and separates from other clusters, proving that the proposed IPN model could learn the discriminative relation embeddings.

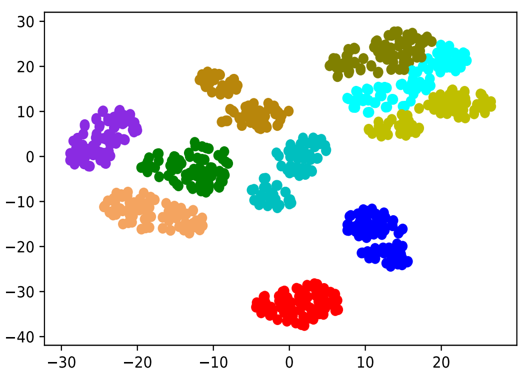

Figure 4: t-SNE visualization of relation embeddings. Circles indicate the relation embeddings of a query sample and numerous support samples from 10 classes. Different colors denote different class labels of support samples.

4 Conclusion

In this paper, we proposed a two-stage inverted pyramid network (IPN) for few-shot learning inspired by the inverted pyramid theory, which is the first work integrating both global and local characteristics in few-shot learning. At the global stage, the CCMNet is introduced to predict the query-to-class similarity from the global perspective. Then at the local stage, a fine-grained calibration is further appended to compensate for the weakness of the global prediction. Extensive experiments conducted on several widely-used datasets demonstrate that IPN outperforms other state-of-the-art few-shot approaches by a large margin.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61872021, 61690202), and Beijing Nova Program of Science and Technology (Z191100001119050).

References

[Ahissar and Hochstein, 2004] Merav Ahissar and Shaul Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8(10):457–464, 2004.

[Chen *et al.*, 2019] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification, 2019.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[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 *2009 IEEE Computer Vision and Pattern Recognition*, 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 *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[Finn *et al.*, 2018] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, pages 9516–9527, 2018.

[Garcia and Bruna, 2017] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.

[Gidaris and Komodakis, 2018] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[Gidaris and Komodakis, 2019] Spyros Gidaris and Nikos Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. *arXiv preprint arXiv:1905.01102*, 2019.

[Gordon *et al.*, 2018] Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018.

[Ji *et al.*, 2013] Zhong Ji, Jing Wang, Yuting Su, Zhanjie Song, and Shikai Xing. Balance between object and background: Object-enhanced features for scene image classification. *Neurocomputing*, 120:15–23, 2013.

[Kim *et al.*, 2019] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 11–20, 2019.

[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] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 1–10, 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 *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8642–8649, 2019.

[Liu *et al.*, 2018] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*, 2018.

[Mishra *et al.*, 2017] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

[Nichol *et al.*, 2018] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[Oreshkin *et al.*, 2018] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018.

[Qiao *et al.*, 2018] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 7229–7238, 2018.

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

[Ren *et al.*, 2018] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[Rusu *et al.*, 2018] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[Sun *et al.*, 2019] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 403–412, 2019.

[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 *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 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 *Advances in neural information processing systems*, pages 3630–3638, 2016.

[Yoon *et al.*, 2019] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. *arXiv preprint arXiv:1905.06549*, 2019.

[Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.