Asymmetric Distribution Measure for Few-shot Learning

Wenbin Li[1] , **Lei Wang**[2] , **Jing Huo**[1*] , **Yinghuan Shi**[1] , **Yang Gao**[1] and **Jiebo Luo**[3]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, China
[2]University of Wollongong, Australia
[3]University of Rochester, USA

liwenbin@nju.edu.cn, leiw@uow.edu.au, {huojing, syh, gaoy}@nju.edu.cn, jluo@cs.rochester.edu

Abstract

The core idea of metric-based few-shot image classification is to directly measure the relations between query images and support classes to learn transferable feature embeddings. Previous work mainly focuses on image-level feature representations, which actually cannot effectively estimate a class's distribution due to the scarcity of samples. Some recent work shows that local descriptor based representations can achieve richer representations than image-level based representations. However, such works are still based on a less effective instance-level metric, especially a symmetric metric, to measure the relation between a query image and a support class. Given the natural asymmetric relation between a query image and a support class, we argue that an asymmetric measure is more suitable for metric-based few-shot learning. To that end, we propose a novel *Asymmetric Distribution Measure (ADM)* network for few-shot learning by calculating a joint local and global asymmetric measure between two multivariate local distributions of a query and a class. Moreover, a task-aware *Contrastive Measure Strategy (CMS)* is proposed to further enhance the measure function. On popular *mini*ImageNet and *tiered*ImageNet, ADM can achieve the state-of-the-art results, validating our innovative design of asymmetric distribution measures for few-shot learning. The source code can be downloaded from https://github.com/WenbinLee/ADM.git.

1 Introduction

Few-shot learning (FSL) for image classification has gained considerable attention in recent years [Vinyals *et al.*, 2016; Finn *et al.*, 2017; Sung *et al.*, 2018; Lee *et al.*, 2019], which attempts to learn a classifier with good generalization capacity for new unseen classes with only a few samples. Because of the scarcity of data, it is almost impossible to directly train

---

*Corresponding author

a conventional supervised model (*e.g.,* a convolutional neural network) from scratch by only using the few available samples. Therefore, transfer learning shall be a natural way to learn transferable knowledge to boost the target few-shot classification. Along this way, a variety of methods have been proposed, which can be roughly divided into three categories: *data-augmentation based* methods [Antoniou *et al.*, 2017; Schwartz *et al.*, 2018; Xian *et al.*, 2019], *meta-learning based* methods [Ravi and Larochelle, 2017; Jamal and Qi, 2019; Lee *et al.*, 2019] and *metric-based* methods [Vinyals *et al.*, 2016; Sung *et al.*, 2018; Li *et al.*, 2019b]. Metric-based FSL methods have achieved significant successes and attracted increasing attention due to their simplicity and effectiveness. In this work, we focus on this kind of methods.

The basic idea of metric-based FSL methods is to learn a transferable deep embedding network by directly measuring the relations between query images and support classes. Thus, two key issues are involved in such a kind of methods, *i.e.,* feature representations and relation measure. For feature representations, traditional methods such as ProtoNet [Snell *et al.*, 2017] and RelationNet [Sung *et al.*, 2018] generally adopt image-level global feature representations for both query images and support classes. However, due to the scarcity of samples in each class, the distribution of these image-level global features cannot be reliably estimated. Recently, CovaMNet [Li *et al.*, 2019b] and DN4 [Li *et al.*, 2019a] introduce deep local descriptors into FSL and attempt to utilize the distribution of local descriptors to represent each support class, which have been verified to be more effective than using the image-level global features.

On the relation measure, the existing methods including the above methods usually adopt an instance-level metric, where the query image is taken as one single instance (*i.e.,* an image-level feature representation) or a set of instances (*i.e.,* a set of local feature descriptors). For example, in ProtoNet, the Euclidean distance is chosen to calculate the distance between a query instance and the prototype (*i.e.,* mean vector) of each support class. Also, CovaMNet proposes a covariance metric function to measure a local similarity between each local descriptor of a query image and a support class. Afterwards, it aggregates all the local similarities to obtain a global similarity as the relation between this query image and this class.

However, these existing methods have only considered the distributions of the support classes while neglecting the natural distribution of the local descriptors from a query image. Moreover, the instance-level metric they employ can only capture a kind of local relations (*i.e.,* local similarities) between the query images and support classes. We argue that the distribution associated with a query image is equally important and a *distribution-level measure* shall be designed to capture the global-level relations between a query and a class. More importantly, we observe that the existing methods usually adopt a symmetric metric function (*i.e.,* $M(a, b) = M(b, a)$) to calculate a symmetric relation between a query and a class. For instance, both the Euclidean distance used in ProtoNet and the cosine similarity adopted in CovaMNet and DN4 are symmetric functions. However, we highlight that there is naturally an asymmetric relation between a query image and a certain class. In particular, when each image is represented by a set of deep local descriptors, the distribution of the descriptors in one query image is only comparable to part of the distribution of the descriptors extracted from a support class. Therefore, we argue that an asymmetric measure is more suitable for the metric-based FSL to capture the asymmetric relations.

To this end, we develop a novel *Asymmetric Distribution Measure (ADM)* network for metric-based FSL. First, we represent each image as a set of deep local descriptors (instead of a single image-level global feature) and consider characterizing both a query image and a support class from the perspective of local descriptor based distributions (*e.g.,* a Gaussian distribution with mean vector and covariance matrix). Second, we employ an asymmetric Kullback–Leibler (KL) divergence measure to align the distribution of a query with the distribution of a support class to capture the global distribution-level asymmetric relations. Third, to further improve the metric by taking the context of the task into consideration, we propose a task-aware *Contrastive Measure Strategy (CMS)*, which can be used as a plug-in to any measure functions. Finally, inspired by the successful image-to-class measure (an asymmetric measure as a whole) introduced in DN4 [Li *et al.*, 2019a] which mainly captures the asymmetric relations via individual local descriptor based cosine similarity measures, we combine the whole distribution based KL divergence measure with the image-to-class measure together to simultaneously capture the global and local relations.

The main contributions of this work are as follows:

- We propose a pure distribution based method for metric-based FSL and show that an asymmetric measure is more suitable for this kind of FSL methods.

- We simultaneously combine the global relations (*i.e.,* the KL divergence measure) and the local relations (*i.e.,* the image-to-class measure) together to measure the complete asymmetric distribution relations between a query and a class.

- We propose an adaptive fusion strategy to adaptively integrate the global and local relations.

- We design a task-aware contrastive measure strategy (CMS) as a plug-in to further enhance the adopted measure functions.

2 Related Work

We first briefly review metric-based FSL methods in the literature, and then introduce related work that inspires our work.

The first metric-based FSL method was proposed in [Koch *et al.*, 2015], which adopts a Siamese neural network to learn transferable and discriminative feature representations. In [Vinyals *et al.*, 2016], a Matching Net which directly compares a query image with a support class was presented, where a subsequently widely used episodic training mechanism was also proposed. After that, [Snell *et al.*, 2017] proposed a ProtoNet, which represents a support class by a prototype, *i.e.,* the mean vector of all sample in this class. Then a specific metric, *i.e.,* Euclidean distance, was used to perform the final classification. Recently, based on ProtoNet, an infinite mixture prototypes (IMP) network was proposed [Allen *et al.*, 2019], where each support class was represented by a set of adaptive prototypes. In addition, to avoid choosing a specific metric function, RelationNet [Sung *et al.*, 2018] proposed to learn a metric through a deep convolutional neural network to measure the similarity between queries and support classes.

The above methods are all based on image-level feature representations. Due to the scarcity of samples in each class in FSL, the distribution of each class cannot be reliably estimated in a space of image-level features. Some recent work, such as CovaMNet [Li *et al.*, 2019b] and DN4 [Li *et al.*, 2019a] shows that the rich local features (*i.e.,* deep local descriptors) can achieve better representations than the image-level features, because the local features can be regarded as a natural data augmentation operation. CovaMNet employs the second-order covariance matrix of the extracted deep local descriptors to represent each support class and designs a covariance-based metric to measure the similarities between a query image and a support class. Different from CovaMNet, DN4 argues that the pooling of local features into a compact image-level representation will lose considerable discriminative information. Therefore, DN4 proposes to directly use the raw local descriptor sets to represent both query images and support classes, and then employs a cosine-based image-to-class measure to perform the relation measure.

Inspired by CovaMNet and DN4, our ADM also takes the raw and rich deep local descriptors to represent an image. Compared with CovaMNet, the key difference is that CovaMNet only considers the distribution associated with a support class but neglect the distribution associated with a query image, while we consider both. Another important difference is that both CovaMNet and DN4 employ a cosine similarity function (*i.e.,* a symmetric instance-level metric) to calculate a series of local relations between a query image and a certain class. In contrast, our ADM can capture the complementary global relations by using an extra distribution-level measure. In addition, we observe that the relation between a query image and a certain class is actually asymmetric, *i.e.,* a query image is only commensurate with a sample in an image class when it is viewed as a set. Therefore, we argue that an asymmetric measure shall be considered for metric-based FSL to reflect this property.
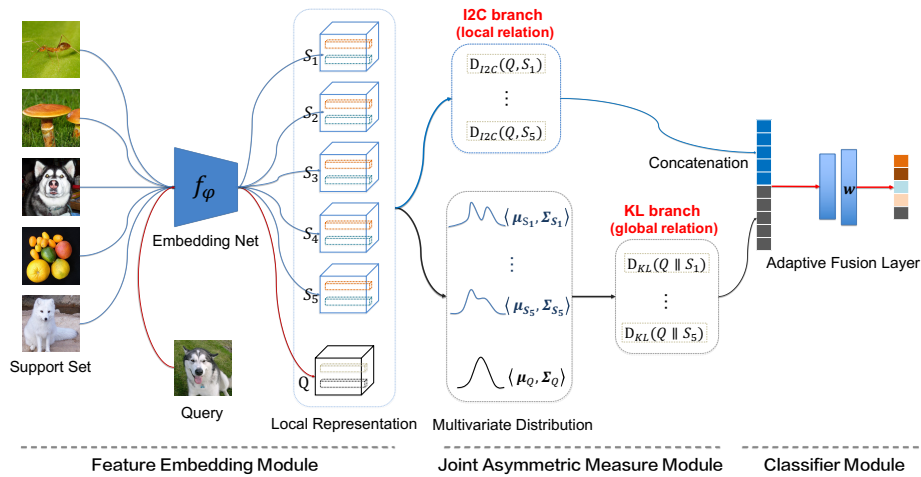
Figure 1: Architecture of the proposed Asymmetric Distribution Measure (ADM) network for a 5-way 1-shot task, which consists of three modules, *i.e.,* a feature embedding module, a joint asymmetric measure module and a classifier module.

3 Preliminary

Problem formulation. Under the few-shot setting, there are usually three sets of data, *i.e.,* a support set $\mathcal{S}$, a query set $\mathcal{Q}$ and an auxiliary set $\mathcal{A}$. In particular, $\mathcal{S}$ and $\mathcal{Q}$ share the same label space, which are corresponding to the training and test sets respectively in the general classification task. If $\mathcal{S}$ contains C classes with K (*e.g.,* 1 or 5) samples per class, we call this classification task C-way K-shot. However, $\mathcal{S}$ only has a few samples in each class, making it almost impossible to train a deep neural network effectively. Therefore, the auxiliary set $\mathcal{A}$ is generally introduced to learn transferable knowledge to tackle this problem. Also, $\mathcal{A}$ enjoys more classes and more samples per class than $\mathcal{S}$, but has a disjoint label space from $\mathcal{S}$.

Episodic training. To learn a classifier that can generalize well, an episodic training mechanism [Vinyals *et al.*, 2016] is normally adopted in the training stage of the metric-based FSL methods. Specifically, in each episode, a new task simulating the target few-shot task is randomly constructed from $\mathcal{A}$. Each simulated task consists of two subsets, $\mathcal{A}_\mathcal{S}$ and $\mathcal{A}_\mathcal{Q}$, which are akin to $\mathcal{S}$ and $\mathcal{Q}$, respectively. At each iteration, one episode (task) is adopted to train the current model. Basically, tens of thousands of episodes (tasks) will be randomly sampled to train this model. Once the training process is completed, we can predict the labels of $\mathcal{Q}$ using the trained model based on $\mathcal{S}$.

4 Methodology

As illustrated in Figure 1, our ADM model mainly consists of three components: a feature embedding module, a joint asymmetric measure module, and a classifier module. The first module learns feature embeddings and produces rich deep local descriptors for an input image. Afterwards, the distributions of each query image and each support class can be represented at the level of deep local descriptors. The second module defines a joint asymmetric distribution measure between a query's distribution and a support class's distribution by con-

sidering both the asymmetric local and global relations. As for the last module, we adaptively fuse the local and global relations together by a jointly learned weight vector, and then adopt a non-parametric nearest neighbor classifier as the final classifier. These three modules are jointly trained from scratch in an end-to-end manner.

4.1 Feature Embedding with Local Descriptors

As have been shown by some recent work [Li *et al.*, 2019b; Li *et al.*, 2019a], local descriptor based feature representations are much richer than image-level features and can alleviate the scarcity issue of samples in FSL. Following the above work, we employ the rich and informative local descriptors to represent each image as well.

To this end, we design a feature embedding module $f_\varphi(\cdot)$, which can extract deep local descriptors for input images. Specifically, given an image X, $f_\varphi(X)$ will be a $c \times h \times w$ three-dimensional (3D) tensor, which can be seen as a set of c-dimensional local descriptors

$$f_\varphi(X) = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{c \times n}, \qquad (1)$$

where $\boldsymbol{x}_i$ is the i-th local descriptor and $n = h \times w$ is the total number of local descriptors for image X. These local descriptors can be seen as the local representations of the spatial local patches in this image. Basically, for each query image, we use the extracted n local descriptors to estimate its distribution in the space of $\mathbb{R}^c$. As for each support class, all the local descriptors of all the images in this class will be used together to estimate its distribution in the space of $\mathbb{R}^c$. Since the local descriptors can capture the local subtle information, they can benefit more for the final image recognition.

4.2 Our Asymmetric Distribution Measure (ADM)

Kullback–Leibler divergence based distribution measure. Assuming that the distribution of local descriptors extracted from an image or a support class is a multivariate Gaussian, a query image's distribution can be denoted by $Q = \mathcal{N}(\boldsymbol{\mu}_Q, \boldsymbol{\Sigma}_Q)$, and a certain support class's distribution can be

expressed by $S = \mathcal{N}(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S)$, where $\boldsymbol{\mu} \in \mathbb{R}^c$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{c \times c}$ indicate the mean vector and covariance matrix of a specific distribution, respectively. Thus, Kullback-Leibler (KL) divergence [Duchi, 2007] between Q and S can be defined as:

$$D_{\mathrm{KL}}(Q\|S) = \frac{1}{2}\Big(\mathrm{trace}(\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_Q) + \ln\big(\frac{\det\boldsymbol{\Sigma}_S}{\det\boldsymbol{\Sigma}_Q}\big) \\ + (\boldsymbol{\mu}_S - \boldsymbol{\mu}_Q)^\top \boldsymbol{\Sigma}_S^{-1}(\boldsymbol{\mu}_S - \boldsymbol{\mu}_Q) - c\Big), \quad (2)$$

where $\mathrm{trace}(\cdot)$ is the trace operation of matrix, $\ln(\cdot)$ denotes logarithm with the base of e, and $\det$ indicates the determinant of a square matrix. As seen, Eq.(2) takes both the mean and covariance into account to calculate the distance between two distributions.

Typically, since the KL divergence measure is asymmetric, $D_{\mathrm{KL}}(Q\|S)$ mainly matches the distribution of Q to the one of S, which is essentially different from $D_{\mathrm{KL}}(S\|Q)$. One important advantage of using Eq.(2) is that it can naturally capture an asymmetric relation between a query image and a support class, forcing the query images to be close to the corresponding true class when used in our network training.

To further show the advantage of using an asymmetric measure, we purposely introduce a symmetric distribution metric function, *e.g.,* 2-Wasserstein distance [Olkin and Pukelsheim, 1982], whose formulation is defined as follows,

$$D_{\mathrm{wass}}(Q, S)^2 = \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_S\|_2^2 + \\ \mathrm{trace}\Big(\boldsymbol{\Sigma}_Q + \boldsymbol{\Sigma}_S - 2\big(\boldsymbol{\Sigma}_Q^{\frac{1}{2}}\boldsymbol{\Sigma}_S\boldsymbol{\Sigma}_Q^{\frac{1}{2}}\big)^{\frac{1}{2}}\Big), \quad (3)$$

However, due to the square rooting of matrices, the calculation of the above distance function is time consuming and the optimization of this function is difficult. Therefore, in the literature [Berthelot *et al.*, 2017; He *et al.*, 2018], an approximation function is normally employed

$$D_{\mathrm{wass}}(Q, S)^2 = \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_S\|_2^2 + \|\boldsymbol{\Sigma}_Q - \boldsymbol{\Sigma}_S\|_F^2, \quad (4)$$

where the first term calculates the squared Euclidean distance between two mean vectors and the second term is a squared Frobenius norm of the difference between two covariance matrices. The comparison and analysis between 2-Wasserstein distance and KL divergence will be detailed in Section 5.5.

Image-to-Class based distribution measure. The above KL divergence measure can capture the global distribution-level relation between a query image and a support class. Nevertheless, the local relations are not taken into consideration yet. According to a deep analysis of DN4 [Li *et al.*, 2019a], we observe that there may be two implicit reasons for the success of DN4. One reason is that the local descriptor based measure (*i.e.,* local relations) it used enjoys a stronger generalization ability than the image-level feature based measure. The other key reason is that the image-to-class measure used in DN4 is asymmetric on the whole, which aligns well with our argument of the necessity of the asymmetric measure. Therefore, such an asymmetric image-to-class measure is also introduced into our model to capture the local-level relations between a query and a support class. However, the difference in our work lies that the indispensable global relation is also measured by an asymmetric distribution-level measure (*i.e.,* KL divergence).

To be specific, given a query image Q and a support class S, which will be represented as $f_\varphi(Q) = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{c \times n}$ and $f_\varphi(S) = [f_\varphi(X_1), \ldots, f_\varphi(X_K)] \in \mathbb{R}^{c \times nK}$, respectively, where K is the number of shots in S. Thus, the image-to-class (I2C) similarity measure can be formulated as

$$D_{\mathrm{I2C}}(Q, S) = \sum_{i=1}^{n} \mathrm{Topk}\Big(\frac{f_\varphi(Q)^\top \cdot f_\varphi(S)}{\|f_\varphi(Q)\|_F \cdot \|f_\varphi(S)\|_F}\Big), \quad (5)$$

where $\mathrm{Topk}(\cdot)$ means selecting the k largest elements in each row of the correlation matrix between Q and S, *i.e.,* $\frac{f_\varphi(Q)^\top \cdot f_\varphi(S)}{\|f_\varphi(Q)\|_F \cdot \|f_\varphi(S)\|_F}$. Typically, k is set as 1 in our work.

Classification with an adaptive fusion strategy. Since two types of relations have been calculated, *i.e.,* global-level relations calculated by the KL divergence measure and local-level relations produced by the I2C measure, a fusion strategy shall be designed to integrate these two parts. To tackle this issue, we adopt a learnable 2-dimensional weight vector $\boldsymbol{w} = [w_1, w_2]$ to implement this fusion. It is worth noting that because the KL divergence indicates dissimilarity rather than similarity, we use the negative of this divergence to obtain a similarity. Specifically, the final fusion similarity between a query Q and a class S can be defined as follows

$$D(Q, S) = -w_1 \cdot D_{\mathrm{KL}}(Q\|S) + w_2 \cdot D_{\mathrm{I2C}}(Q, S). \quad (6)$$

As seen in Figure 1, for a 5-way 1-shot task and a specific query Q, the output of the I2C branch or KL branch is a 5-dimensional similarity vector. Next, we concatenate these two vectors together to get a 10-dimensional vector. After that, we apply a 1D convolution layer with the kernel size of 1×1 along with a dilation value of 5. In this way, we can obtain a weighted 5-dimensional similarity vector by learning a 2-dimensional weights $\boldsymbol{w}$. Additionally, a Batch Normalization layer is also added before the 1D convolution layer to balance the scale of the two parts of similarities. Finally, a non-parametric nearest neighbor classifier is performed to obtain the final classification results. Also, a cross-entropy loss is used to learn the entire network.

4.3 Our Contrastive Measure Strategy (CMS)

To make the distribution measure more discriminative, we further propose an alternative task-aware *Contrastive Measure Strategy (CMS)* by introducing additional contrastive information. Specifically, for a support set $\mathcal{S} = \{S_1, \cdots, S_C\}$, where C is the number of classes in $\mathcal{S}$, we construct a *distribution-level triplet* $\langle Q, S_i, S_i' \rangle$. In this triplet, Q denotes a query's distribution, S_i is the distribution of one class we want to match Q with, and S_i' indicates the entire distribution of the remaining classes $S_j|_{j=1}^C (j \neq i)$. In this way, we can define the contrastive KL divergence measure as follows

$$D_{\mathrm{KL}}^{\mathrm{con}}(Q\|S_i) = D_{\mathrm{KL}}(Q\|S_i) - D_{\mathrm{KL}}(Q\|S_i'). \quad (7)$$

The advantage of using the above contrastive measure function over merely using $D_{\mathrm{KL}}(Q\|S_i)$ in Eq.(2) is that the context of the entire support classes is taken into consideration. In this way, we can take a whole view of the entire task when measuring the relation between Q and each individual class S_i, making the measure function more discriminative. This will be experimentally demonstrated shortly.

Method	Type	Measure	miniImageNet 5-way Acc (%)		tieredImageNet 5-way Acc (%)	
			1-shot	5-shot	1-shot	5-shot
ProtoNet[‡] [NeurIPS 2017]	Symmetric	Instance-level	$48.45_{\pm0.96}$	$66.53_{\pm0.51}$	$48.58_{\pm0.87}$	$69.57_{\pm0.75}$
RelationNet [CVPR 2018]	Symmetric	Instance-level	$50.44_{\pm0.82}$	$65.32_{\pm0.70}$	$54.48_{\pm0.93}$	$71.31_{\pm0.78}$
Wass (Ours)	Symmetric	Distribution-level	$50.27_{\pm0.62}$	$67.50_{\pm0.52}$	$52.76_{\pm0.71}$	$73.58_{\pm0.57}$
Wass-CMS (Ours)	Symmetric	Distribution-level	$50.80_{\pm0.64}$	$68.36_{\pm0.50}$	$53.48_{\pm0.68}$	$73.95_{\pm0.56}$
KL (Ours)	Asymmetric	Distribution-level	$\mathbf{52.94_{\pm0.63}}$	$\mathbf{69.38_{\pm0.51}}$	$\mathbf{55.59_{\pm0.70}}$	$\mathbf{74.21_{\pm0.56}}$
KL-CMS (Ours)	Asymmetric	Distribution-level	$\mathbf{53.10_{\pm0.62}}$	$\mathbf{69.73_{\pm0.50}}$	$\mathbf{56.54_{\pm0.70}}$	$\mathbf{74.83_{\pm0.56}}$

Table 1: Ablation study on both miniImageNet and tieredImageNet. The second column refers to whether the measure function adopted is symmetric or not. The third column indicates which kind of measure function is employed, i.e., instance-level or distribution-level. For each setting, the best and second best methods are highlighted.

5 Experimental Result

5.1 Datasets

All experiments are conducted on both miniImageNet [Vinyals et al., 2016] and tieredImageNet [Ren et al., 2018].

miniImageNet is widely used in FSL, which is a small subset of ImageNet [Deng et al., 2009]. It contains 100 classes with 600 images in each class. We use the same splits as in [Ravi and Larochelle, 2017], which takes 64, 16 and 20 classes for training, validation and test, respectively.

tieredImageNet is also a mini-version of ImageNet. Different from miniImageNet, tieredImageNet has a larger number of classes (608 classes) and more images for each class (1281 images per class). On this dataset, we strictly follow the splits used in [Ren et al., 2018], which takes 351, 97 and 160 classes for training, validation and test, respectively.

For both miniImageNet and tieredImageNet, the resolution of all the images is resized to 84×84.

5.2 Network Architecture

It can be easily verified that adopting a deeper network for embedding or using pre-trained weights will provide higher accuracy. Following the previous works [Snell et al., 2017; Sung et al., 2018; Li et al., 2019b; Li et al., 2019a], we adopt the same embedding network with four convolutional blocks, i.e., Conv-64F, to make a fair comparison with other methods. Specifically, each of the first two blocks contains a convolutional layer (with 64 filters of size 3×3), a batch-normalization layer, a Leaky ReLU layer and a max pooling layer. The last two blocks adopt the same architecture but without pooling layers. The reason for only using two pooling layers is that we need richer local descriptors to represent the distributions of both queries and classes. For example, in a 5-way 1-shot setting, when the size of the input image is 84×84, we can only obtain 25 local descriptors for each image (class) by adopting four pooling layers. It is clearly insufficient to represent a distribution with a feature dimensionality of 64. In contrast, using the adopted network architecture with two pooling layers, we obtain 441 local descriptors for each image (class).

5.3 Implementation Details

Both 5-way 1-shot and 5-way 5-shot classification tasks are conducted to evaluate our methods. We use 15 query images per class in each single task (75 query images in total) in both training and test stages. In particular, we employ the episodic training mechanism [Vinyals et al., 2016] to train our models from scratch without pre-training. In the training stage, we use the Adam algorithm [Kingma and Ba, 2014] to train all the models for 40 epoches. In each epoch, we randomly construct 10000 episodes (tasks). Also, the initial learning rate is set as 1×10^{-3} and multiplied by 0.5 every 10 epoches. During test, 1000 tasks are randomly constructed to calculate the final results, and this process is repeated five times. The top-1 mean accuracy is taken as the evaluation criterion. At the same time, the 95% confidence intervals are also reported.

5.4 Comparison Methods

Since our methods belong to the metric-based FSL methods, we will mainly compare our methods with metric-based methods, such as Matching Net [Vinyals et al., 2016], ProtoNet [Snell et al., 2017], RelationNet [Sung et al., 2018], IMP [Allen et al., 2019], CovaMNet [Li et al., 2019b] and DN4 [Li et al., 2019a]. Moreover, representative meta-learning based FSL methods are also listed for reference, including Meta LSTM [Ravi and Larochelle, 2017], MAML [Finn et al., 2017], SNAIL [Mishra et al., 2017], MTL [Sun et al., 2019], TAML-Entropy [Jamal and Qi, 2019], and MetaOptNet-RR [Lee et al., 2019]. Note that meta-learning based methods are essentially different from metric-based methods at two aspects. The first aspect is that an additional parameterized meta-learner is usually learned in meta-learning based methods while the metric-based methods do not have. The second aspect is that during test, meta-learning based methods will fine-tune the model (or classifier) to obtain the final classification results while metric-based methods do not need fine-tuning.

Most results of these compared methods are quoted from their original work or the relevant reference. Some methods are not in the same setting with our method, such as ProtoNet, so we use the results of their modified versions to ensure fair comparison. For some recent meta-learning based methods, such as SNAIL, MTL and TAML-Entropy, we only report their results with a similar embedding network, e.g., Conv-32F, which has the same architecture with Conv-64F but has 32 filters in each convolutional block.

5.5 Ablation Study

In this section, we first verify the validity of our argument on asymmetric measure for metric-based FSL. Next, based on

Method	Venue	Embed.	Type	Para.	*mini*ImageNet 5-way Acc (%)		*tiered*ImageNet 5-way Acc (%)	
					1-shot	5-shot	1-shot	5-shot
Meta LSTM	ICLR'17	*Conv-32F*	Meta	-	$43.44_{\pm0.77}$	$60.60_{\pm0.71}$	-	-
MAML	ICML'17	*Conv-32F*	Meta	-	$48.70_{\pm1.84}$	$63.11_{\pm0.92}$	$51.67_{\pm1.81}$	$70.30_{\pm1.75}$
SNAIL	ICLR'18	*Conv-32F*	Meta	-	45.10	55.20	-	-
MTL	CVPR'19	*Conv-32F*	Meta	-	$45.60_{\pm1.80}$	$61.20_{\pm0.90}$	-	-
TAML-Entropy	CVPR'19	*Conv-32F*	Meta	-	$49.33_{\pm1.80}$	$66.05_{\pm0.85}$	-	-
MetaOptNet-RR	CVPR'19	*Conv-64F*	Meta	-	$\mathbf{53.23_{\pm0.59}}$	$69.51_{\pm0.48}$	$54.63_{\pm0.67}$	$72.11_{\pm0.59}$
Matching Nets	NeurIPS'16	*Conv-64F*	Metric	113 kB	$43.56_{\pm0.84}$	$55.31_{\pm0.73}$	-	-
ProtoNet[‡]	NeurIPS'17	*Conv-64F*	Metric	113 kB	$48.45_{\pm0.96}$	$66.53_{\pm0.51}$	$48.58_{\pm0.87}$	$69.57_{\pm0.75}$
RelationNet	CVPR'18	*Conv-64F*	Metric	228 kB	$50.44_{\pm0.82}$	$65.32_{\pm0.70}$	$54.48_{\pm0.93}$	$71.31_{\pm0.78}$
IMP	ICML'19	*Conv-64F*	Metric	113 kB	$49.6_{\pm0.8}$	$68.1_{\pm0.8}$	-	-
CovaMNet	AAAI'19	*Conv-64F*	Metric	113 kB	$51.19_{\pm0.76}$	$67.65_{\pm0.63}$	$54.98_{\pm0.90}$	$71.51_{\pm0.75}$
DN4	CVPR'19	*Conv-64F*	Metric	113 kB	$51.24_{\pm0.74}$	$\mathbf{71.02_{\pm0.64}}$	$53.37_{\pm0.86}$	$74.45_{\pm0.70}$
KL	Ours	*Conv-64F*	Metric	113 kB	$52.94_{\pm0.63}$	$69.38_{\pm0.51}$	$55.59_{\pm0.70}$	$74.21_{\pm0.56}$
KL-CMS	Ours	*Conv-64F*	Metric	113 kB	$53.10_{\pm0.62}$	$69.73_{\pm0.50}$	$\mathbf{56.54_{\pm0.70}}$	$74.83_{\pm0.56}$
ADM	Ours	*Conv-64F*	Metric	113 kB	$\mathbf{54.26_{\pm0.63}}$	$\mathbf{72.54_{\pm0.50}}$	$56.01_{\pm0.69}$	$\mathbf{75.18_{\pm0.56}}$

Table 2: The mean accuracies of the 5-way 1-shot and 5-shot tasks on both *mini*ImageNet and *tiered*ImageNet, with 95% confidence intervals. The third column refers to which kind of embedding network is employed. The fifth column shows the total number of parameters used by each method. [‡] results are obtained by reimplementing in the same setting. For each setting, the best and second best methods are highlighted.

two distribution-level measure functions, we evaluate the effectiveness of the proposed CMS strategy. Specifically, both the 2-Wasserstein distance (Wass for short) and KL divergence (KL for short) are performed on *mini*ImageNet and *tiered*ImageNet. Also, the contrastive versions using our proposed CMS are named as Wass-CMS and KL-CMS, respectively. Moreover, two instance-level symmetric metric based methods, ProtoNet and RelationNet, are taken as baselines.

As seen in Table 1, compared to symmetric metric based methods, such as ProtoNet, RelationNet and Wass, the proposed asymmetric measure can obtain superior results. For example, on the *mini*ImageNet, KL gains 4.49%, 2.50% and 2.67% over these methods on the 1-shot task, respectively. This verifies that an asymmetric measure is more suitable for metric-based FSL.

We can also see that the proposed CMS strategy can indeed improve the performance of distribution-based measure functions, especially on the 1-shot setting. For instance, on the *tiered*ImageNet, Wass-CMS achieves 0.72% improvement over Wass, and KL-CMS obtains 0.95% improvement over KL on the 1-shot task. This shows that the task-aware CMS strategy does enhance the distribution-based measure functions, thanks to taking a whole view of the entire task.

5.6 Comparison with the State of the Art

Experimental results on the comparison with the state-of-the-art methods are reported in Table 2, where two types of FSL methods (*i.e.,* both meta-learning based and metric-based) are compared. Since our methods are metric-based methods, we will mainly compare our methods with other metric-based ones. Moreover, the total number of parameters of each method is also shown in the fifth column.

From Table 2, it can be seen that the proposed ADM (I2C+KL+Fusion) outperforms all the other metric-based and meta-learning based methods on both 1-shot and 5-shot settings. For example, on the *mini*ImageNet, our ADM obtains 10.7%, 5.81%, 3.82%, 4.66%, 3.07% and 3.02% improve-

ments over Matching Nets, ProtoNet, RelationNet, IMP, CovaMNet and DN4 on the 1-shot task, respectively. Moreover, on the *tiered*ImageNet, our ADM achieves 5.61%, 3.87%, 3.67%, 0.73% improvements over ProtoNet, RelationNet, CovaMNet and DN4 on the 5-shot task, respectively. This verifies the effectiveness and superiority of our proposed ADM, owing to the integration of both local and global asymmetric relations.

The proposed KL and KL-CMS are also very competitive with the state-of-the-art methods. Specifically, on the 1-shot setting, KL and KL-CMS can obtain significantly improvements over the existing metric-based methods. For instance, on the *mini*ImageNet, KL/KL-CMS gains 9.38%/9.54%, 4.49%/4.65%, 2.5%/2.66%, 3.34%/3.5%, 1.75%/1.91% and 1.7%/1.86% improvements over Matching Nets, ProtoNet, RelationNet, IMP, CovaMNet and DN4, respectively. It verifies that such kind of distribution-based asymmetric measure is more suitable for metric-based FSL.

6 Conclusion

In this study, we provide a new perspective for metric-based FSL by considering the asymmetric nature of the similarity measure and design a novel *Asymmetric Distribution Measure (ADM)* network to address this task. Furthermore, to make full use of the context of the entire task, we propose a *Contrastive Measure Strategy (CMS)* to learn a more discriminative distribution metric space. Extensive experiments on two benchmark datasets verify the effectiveness and advantages of both local asymmetric relations and global asymmetric relations in metric-based FSL.

Acknowledgements

References

[Allen *et al.*, 2019] Kelsey R Allen, Evan Shelhamer, Hanul Shin, and Joshua B Tenenbaum. Infinite mixture prototypes for few-shot learning. *ICML*, 2019.

[Antoniou *et al.*, 2017] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv*, 2017.

[Berthelot *et al.*, 2017] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv*, 2017.

[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, 2009.

[Duchi, 2007] John Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 3, 2007.

[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, 2017.

[He *et al.*, 2018] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[Jamal and Qi, 2019] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *CVPR*, pages 11719–11727, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

[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.

[Lee *et al.*, 2019] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, pages 10657–10665, 2019.

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

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

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

[Olkin and Pukelsheim, 1982] Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[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*, 2018.

[Schwartz *et al.*, 2018] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *NeurIPS*, pages 2850–2860, 2018.

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

[Xian *et al.*, 2019] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, pages 10275–10284, 2019.