Cooperative Bi-path Metric for Few-shot Learning

Zeyuan Wang\textsuperscript{1}, Yifan Zhao\textsuperscript{1}, Jia Li\textsuperscript{1,3,4}, Yonghong Tian\textsuperscript{2,4}

\textsuperscript{1}State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University, Beijing, 100191, China
\textsuperscript{2}School of Electronics Engineering and Computer Science, Peking University, Beijing, 100871, China
\textsuperscript{3}Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, 100191, China
\textsuperscript{4}Peng Cheng Laboratory, Shenzhen, 518066, China

ABSTRACT

Given base classes with sufficient labeled samples, the target of few-shot classification is to recognize unlabeled samples of novel classes with only a few labeled samples. Most existing methods only pay attention to the relationship between labeled and unlabeled samples of novel classes, which do not make full use of information within base classes. In this paper, we make two contributions to investigate the few-shot classification problem. First, we report a simple and effective baseline trained on base classes in the way of traditional supervised learning, which can achieve comparable results to the state-of-the-art. Second, based on the baseline, we propose a cooperative bi-path metric for classification, which leverages the correlations between base classes and novel classes to further improve the accuracy. Experiments on two widely used benchmarks show that our method is a simple and effective framework, and a new state of the art is established in the few-shot classification field.

CCS CONCEPTS

• Computing methodologies → Supervised learning by classification; Dimensionality reduction and manifold learning; • Information systems → Similarity measures.

KEYWORDS

few-shot learning, image classification, metric learning, locally linear embedding

1 INTRODUCTION

Image recognition provides an intuitive and fundamental way to understand the visual world, which is also deeply rooted in many advanced kinds of research, including autonomic production and artificial intelligence. As the crucial part of visual content, image recognition has witnessed a great progress with the proposals of large dataset \cite{1} and deep learning techniques \cite{2, 3}. However, the notable performance of these methods heavily relies on the large amount of manually annotated datasets, which are labour-intensive and sometimes inaccessible, e.g., ImageNet \cite{1} with over 15 million annotations. Remarkably, humans and other animals seem to have the potential to recognize one identity with very less related knowledge. Therefore, the few-shot learning (FSL) problem with limited seen knowledge forces the model to make a typical generalization of each class, which is a more realistic setting in some extreme industrial applications.

By training a model on base classes that contain sufficient labeled samples, the goal of few-shot learning is to make the model generalize well on the novel classes which do not intersect with the base classes, i.e., correctly classifying unlabeled samples (query samples) according to a small number of labeled samples (support samples). To make the conditions of the training phase match those of the testing phase, Matching Networks \cite{4} firstly suggested that both training and testing should adopt episodic procedure, which came from meta-learning. Models will meet many few-shot learning tasks in both the training and testing phases. Each task consists of several classes, and each class contains a few support samples and several query samples. Leading by the pioneer work \cite{4}, many subsequent researches \cite{5–17} followed this episodic learning and achieved notable improvements. However, some recent works \cite{18–21} did not follow this factitious sampling setting, but directly trained model in the way of traditional supervised learning. Thus a natural concern arises, is episodic training necessary for few-shot learning? Keeping this in our mind, we first make extensive experimental analyses on two commonly used benchmarks \cite{4, 12}. Counter-intuitively, without the always-used episodic training process, state-of-the-art performance can also be achieved through adequate training strategy with all the samples in base classes. This finding not only brings us a rethinking of this conventional setting but also can be considered as a high-performance baseline for few-shot learning.

To start from another perspective, metric learning \cite{22–33} is a major genre in the field of few-shot learning. This kind of method classifies query samples by learning a feature extractor on the base classes, extracting features of samples of the novel classes during testing, and measuring the distance or similarity between labeled support samples and unlabeled query samples. However, most of
Our main contribution is three-fold: 1) We make extensive experimental analyses of the conventional episodic training and full-supervision training for the few-shot classification problem, and propose a new high-performance baseline based on this experimental finding. 2) We propose a novel Cooperative Bi-path Metric learning approach, which first exploits the base classes as an intermediary for facilitating the classification process. 3) We make extensive experimental analyses to demonstrate our findings. Compared with the existing methods, our proposed approach achieves new state-of-the-art results on two widely used benchmarks, i.e., minilimageNet [4] and tieredImageNet [12].

2 RELATED WORK
There are two main branches of methods in the few-shot learning field, one is meta-learning based methods and the other is metric-learning based methods. The former adopts the episodic training procedure in the training stage and expects to learn the common attributes between different tasks through this procedure, namely the appropriate hyperparameters for models. The latter focuses on how to better extract features from samples and classify samples according to the extracted features, that is, the network is expected to learn a function that can properly measure the similarity between samples.

Meta-learning based methods. The purpose of meta-learning, or learning to learn [34–36], is to train a meta-learner to learn task-agnostic knowledge (or hyper-parameters), which can assist the training of learners on different tasks. Meta-learning based methods [5–7, 10, 18, 19] are a major branch in the field of few-shot learning. For example, Meta-LSTM [6] trained an LSTM-based [37] meta-learner to discover good initialization for learner’s parameters, as well as a mechanism for updating the learner’s parameters by a small sample set. Similarly, Meta-SGD [5] trained a meta-learner to produce learner’s initialization, update direction, and learning rate, but in a single meta-learning process. Also, MAML [7] aimed to find the appropriate initial parameters of the learner, so that the learner could converge rapidly with a few samples. Similar to MAML [7], Reptile [10] was a meta-learning method for finding neural network initialization parameters but simply performed SGD on each task without computing gradient twice as MAML [7] did. Gidaris et al. [18] and Qiao et al. [19] both used base classes to train a feature extractor in the first stage, then used the base classes training a classification weight generator (meta-learner) in the second stage, and used the classification weight generator and support samples to generate classification weight vectors (parameters of the learner) for novel classes during the testing phase. As we can see, all the above meta-learning methods require training a meta-learner and more or less fine-tuning on novel classes, which undoubtedly increases the complexity of the methods. However, our baseline is trained in the way of supervised learning, which can achieve comparable performance to the state of the art, meanwhile greatly simplifies the training process.

Metric-learning based methods. Another branch in the few-shot learning field is metric-learning based methods [4, 8, 9, 11, 12, 14], which focus on embedding samples into a metric space so that the samples can be classified according to similarity or distance between each other. Matching Networks [4] used LSTM [37] and variants to extract feature vectors from support samples and query samples, and then classified query samples by calculating cosine similarities between support and query samples. Prototypical Networks [9] took the average of the feature vectors of support samples within a class as the prototype of the class, and assigned the query samples to the nearest prototype in Euclidean distance, meanwhile its convolutional neural network was trained end-to-end. Triantafillou
et al. [8] proposed an information retrieval-inspired approach viewing each batch point as a query that ranked the remaining ones and defined a model to optimize mean Average Precision over these rankings. Relation Network [11] adapted the same convolutional neural network as Prototypical Networks [9] to extract the features of support samples and query samples, but features of support samples and query samples were concatenated and input into a non-linear relation module to obtain classification scores. Ren et al. [12] presented a new problem: semi-supervised few-shot classification, in which support samples consisted of labeled and unlabeled samples, and proposed several novel extensions on Prototypical Networks [9]. Also, TADAM [14] learned a task-dependent metric with Metric Scaling, Task Conditioning, and Auxiliary Task Co-training for few-shot classification. Although the above metric-learning based methods could achieve admirable results on few-shot classification, they only considered the direct relationship between query samples and support samples when predicting the labels of query samples of novel classes, and ignored the relationship between novel classes and base classes. After observing this, we propose a novel Cooperative Bi-path Metric to utilize the relationship between the support samples, query samples, and base classes to assist in the classification of queries, which can further improve the accuracy of classification.

3 METHOD

3.1 Problem Definition

Given a training set \( D_{\text{base}} \) containing samples of base classes \( C_{\text{base}} \), the goal of few-shot learning is to train a model with \( D_{\text{base}} \) to achieve high accuracy on the classification tasks obtained by sampling on test set \( D_{\text{novel}} \), which contains samples of novel classes \( C_{\text{novel}} \). The base classes are totally different from novel classes, that is, \( C_{\text{base}} \cap C_{\text{novel}} = \emptyset \). And each task is composed of a support set \( D_{\text{support}} \) with labeled samples and a query set \( D_{\text{query}} \) with unlabeled samples. For a N-way K-shot task, \( D_{\text{support}} = \left\{ (x^{(i)}, y^{(i)}) \right\}_{i=1}^{N \times K} \) contains \( N \) classes and \( K \) support samples for each class. A model is trained to predict the labels of the query samples in \( D_{\text{query}} \) as accurately as possible according to \( D_{\text{support}} \).

3.2 A Strong Baseline for Few-shot Learning

Is episodic training necessary? In the view of episodic training, a lot of works [4–17] made the procedure of training and testing consistent to achieve higher performance. During training phase, tasks were sampled, then loss function of the model was calculated based on the tasks, and the network parameters were updated through the backpropagation. However, some works [18–21] did not follow this setting, but trained classification networks in the way of traditional supervised learning.

To make fair comparisons with the previous works [13, 14, 16, 17, 20], we adopt ResNet-12 [3] as the baseline’s backbone and use the training set \( D_{\text{base}} \) to optimize it in a fully-supervised manner. However, it is interesting to find that a network with only a supervised training strategy can also get superior classification results on novel classes (elaborated in Sect. 4). This indicates the basic knowledge learnt from base classes can not be further improved by the episodic learning and motivates the proposal of our strong baseline.

**Baseline for few-shot learning.** For the problem of few-shot learning, we advocate three meaningful cues in constructing a strong baseline: 1) Data augmentation: following prior work [17], we use horizontal flip, random crop and random erasing [38] as data augmentation. 2) Temperature in learning: inspired by [39], we also introduce a hyper-parameter called temperature, which was first applied in model distillation to change the smoothness of distribution after softmax normalization and the value of cross-entropy. 3) Dense classification: instead of embedding the image features as a vector, we apply dense classification loss [20] to regularize our model, i.e., all the local feature vectors of the feature map before the last fully-connected layer are classified through the fully-connected layer without average pooling. For each training sample \( (x, y) \in D_{\text{base}} \), the proposed baseline with loss \( \mathcal{L} \) has the following form:

\[
\mathcal{L} = - \frac{1}{r} \sum_{i=1}^{r} \log \frac{\exp \left( t \left( f^{(i)} \cdot p^{(y)} + b_{y} \right) \right)}{\sum_{j=1}^{|C_{\text{base}}|} \exp \left( t \left( f^{(i)} \cdot p^{(j)} + b_{j} \right) \right)},
\]

where \( t \) is the temperature hyperparameter, \(|·|\) is the cardinality of a set. \( f^{(i)} \in \mathbb{R}^c \) is the local vector at position \( i \) of the training sample’s feature map \( F \in \mathbb{R}^{c \times r} \) with channel dimension \( c \) and spatial resolution \( r \). \( p^{(j)} \in \mathbb{R}^c \) is the parameter vector for class \( j \) in the fully-connected layer’s parameter matrix \( P \in \mathbb{R}^{c \times |C_{\text{base}}|} \). \( b_{j} \) is the bias for class \( j \) in the fully-connected layer’s bias vector \( b \in \mathbb{R}^{|C_{\text{base}}|} \).

During testing phase, for a N-way K-shot task, a query sample in \( D_{\text{query}} \) is assigned to the class \( \hat{y} \) with maximum classification score \( \phi^{(n)} \):

\[
\hat{y} = \arg \max_n \left( \phi^{(n)} \right).
\]

Classification score \( \phi^{(n)} \) for novel class \( n \) is defined as:

\[
\phi^{(n)} = \cos \left( q, s^{(n)} \right).
\]

And \( \cos (\cdot, \cdot) \) is cosine similarity between two vectors. \( q \) and \( s^{(n)} \) are feature vectors of query sample and class \( n \) respectively:

\[
q = \text{GAP} (Q),
\]

\[
s^{(n)} = \frac{1}{K} \sum_{k=1}^{K} \text{GAP} \left( s^{(n,k)} \right).
\]

And \( \|·\| \) is the \( L_2 \) norm of a vector. \( Q \) and \( s^{(n,k)} \) are feature maps of the query sample and the \( k \)-th support sample of class \( n \) in \( D_{\text{support}} \). \( \text{GAP}(\cdot) \) is the global average pooling on a feature map \( F \) defined as:

\[
\text{GAP} (F) = \frac{1}{r} \sum_{i=1}^{r} f^{(i)},
\]

where \( f^{(i)} \in \mathbb{R}^c \) is the local vector at position \( i \) of the feature map \( F \in \mathbb{R}^{c \times r} \).
3.3 Cooperative Bi-path Metric

As shown in Eq. (2) and (3), the previously proposed baseline in this paper as well as previous methods [4, 9, 11, 14–17, 20] simply classify query samples only according to support samples. The main drawback is that the prior knowledge on base classes is not fully exploited, which can also be complementary to classification decision. Thus a natural thought arises: the similarity distributions on base classes of support samples and query samples within the same class should also be similar, as illustrated in Fig. 1, and this information is useful for classifying query samples.

Starting from this point, we propose a novel method namely Cooperative Bi-path Metric as the classification criterion during the testing phase, which is shown in Fig. 2. Cooperative Bi-path Metric utilizes base classes as an intermediate way to assist with the classification of query samples. Our proposed metric measures the similarity by two individual paths: inductive similarity $\phi$ and transductive similarity $\psi$. Most existing methods regard the former one as the only classification criterion, as shown in the lower half of Fig. 2, which calculates the inductive similarity $\phi$ (e.g. cosine similarity) between the support set and the query set. While Cooperative Bi-path Metric not only measures the inductive similarity $\phi$ but also uses base classes as an agent to calculate the transductive similarity $\psi$ between the support set and the query set, as shown in the upper part of Fig. 2. Firstly, it calculates the similarity distribution $\rho_{\text{support}}$ and $\rho_{\text{query}}$ of support set and query set on base classes, and then calculates the similarity between $\rho_{\text{support}}$ and $\rho_{\text{query}}$ i.e., the transductive similarity $\psi$ between support set and query set. The final classification score $\psi$ during the test phase is a weighted sum of $\phi$ and $\psi$:

$$
\hat{y} = \arg \max_n \psi^{(n)},
$$

$$
\psi^{(n)} = \alpha \phi^{(n)} + (1 - \alpha) \psi^{(n)},
$$

$$
\phi^{(n)} = \sigma(\rho_{\text{query}}^{(n)} - \rho_{\text{support}}^{(n)}).
$$

And $\psi^{(n)}$ is Cooperative Bi-path Metric’s final classification score for novel class $n$. $\phi^{(n)}$ is defined in Eq. (3). $\alpha$ is a hyperparameter to adjust the weight between $\phi^{(n)}$ and $\psi^{(n)}$. $\sigma(\cdot)$ is a similarity function that measures the similarity between two distributions, and it can be cosine similarity or negative Euclidean distance and so on. $\rho_{\text{query}}$ and $\rho_{\text{support}}$ can be formally represented as:

$$
\rho_{\text{query}} = \sigma'(q, B),
$$

$$
\rho_{\text{support}}^{(n)} = \sigma'(s^{(n)}, B),
$$

where $q$ and $s^{(n)}$ are defined in Eq. (4). $\sigma'(\cdot)$ is another similarity function that measures the similarity between a vector and each column of a matrix, while it can be similar to or different from $\sigma(\cdot)$. $B$ is a feature matrix of base classes $\mathcal{C}_{\text{base}}$ which can be formally represented as:

$$
B = \begin{bmatrix} b^{(1)} & \ldots & b^{(|\mathcal{C}_{\text{base}}|)} \end{bmatrix}.
$$

And $b^{(i)}$ is the feature vector of base class $i$, which is defined as:

$$
b^{(i)} = \frac{1}{M^{(i)}} \sum_{j=1}^{M^{(i)}} \text{GAP} \left( P^{(i,j)} \right),
$$

where $M^{(i)}$ is the number of samples of base class $i$. GAP ($\cdot$) is global average pooling defined in Eq. (5). $P^{(i,j)}$ is the feature map of the $j$-th sample of base class $i$ in $\mathcal{D}_{\text{base}}$.

As can be seen from the above, Cooperative Bi-path Metric is a nonparametric (model-free) method, if we do not consider the selection of similarity functions $\sigma$, $\sigma'$ and weight hyperparameter $\alpha$. It does not introduce additional network parameters or change the training process, only additionally takes the similarity distributions of the support samples and the query samples on base classes into consideration. We can just simply append Cooperative Bi-path Metric to any trained models. However, in this way, the classification of query samples depends not only on a small number of support samples but also on the information provided by base classes, thus increasing the robustness of the model when support samples are insufficient.

3.4 Revisiting Few-shot Learning with LLE

According to Eq. (6) and (7) in Section 3.3, each base class makes equal contribution to $\phi^{(n)}$, while $\rho_{\text{query}}$ and $\rho_{\text{support}}^{(n)}$ are linear on all base classes without focusing on some specific classes. Thus there arises a concern: for each query sample, different base vectors should contribute differently based on the correlations in the latent space. For example, the *walker foxhound* from base classes should be prominent when querying the golden retriever sample.

**Cooperative Bi-Path metric with LLE.** We replace $\rho_{\text{query}}$ and $\rho_{\text{support}}^{(n)}$ with nonlinear $\tilde{\rho}_{\text{query}}$ and $\tilde{\rho}_{\text{support}}^{(n)}$ through using local linear embedding (LLE) [40]. Compared with the conventional dimensionality reduction methods such as PCA and LDA which focus on sample variance, LLE focuses on maintaining the local linear characteristics of samples when reducing sample dimensionality. LLE assumes that each sample can be represented by linearly combining its $k$ nearest neighbors, and the weight coefficient of the linear relationship before and after dimensionality reduction remains unchanged. It can be seen that LLE has some selectivity in the dimensionality reduction process, which meets our expectation that the samples should focus on some specific base classes. The process of Cooperative Bi-path Metric with LLE is shown in Alg. 1.

In Alg. 1, once $\tilde{\rho}_{\text{query}}$ and $\tilde{\rho}_{\text{support}}^{(n)}$ are obtained, query samples are assigned to novel class $\hat{y}$ with maximum classification score. Besides, Cooperative Bi-path Metric with LLE increases nonlinearity between base classes and novel samples through reducing dimensionality with LLE, thus different base classes can make different influence on classifying different query samples by the process of finding their $k$ nearest neighbors.

4 EXPERIMENTS

4.1 Experiment Setting

**Datasets.** We conduct experiments on two widely-used benchmarks, i.e., miniImageNet [4] and tieredImageNet [12]. MiniImageNet is a subset of ImageNet [1], with 100 classes and 600 images
Algorithm 1: Cooperative Bi-path Metric with LLE

\textbf{Input:} \( B = [b^{(1)}, \ldots, b_{|C_{\text{base}}|}] \in \mathbb{R}^{c \times |C_{\text{base}}|}; \) feature matrix of base classes, \( q \in \mathbb{R}^c; \) feature vector of query sample, \( s^{(n)}_{\text{query}} \in \mathbb{R}^c; \) feature vector of novel class \( n, k: \) number of nearest neighbors to be considered for each sample, \( c': \) dimensionality after reduction, \( \alpha: \) weight hyperparameter introduced in Section 3.3

\textbf{Output:} \( \hat{s}^{(n)}; \) final classification score of the query sample for novel class \( n \)

\begin{algorithmic}[1]
\For {\( i \in 1, \ldots, |C_{\text{base}}| \)}
\State Find the \( k \) nearest neighbors of \( b^{(i)} \) in Euclidean distance: \( N^{(i)} = KNN \left( b^{(i)}, k \right) = [b^{(i)}, \ldots, b^{(i,k)}] \in \mathbb{R}^{c \times k}; \)
\State Calculate the local covariance matrix:
\[ C^{(i)} = \left( B^{(i)} - N^{(i)} \right)^{\top} \left( B^{(i)} - N^{(i)} \right) \in \mathbb{R}^{k \times k}, \] where \( B^{(i)} = [b^{(i)}, \ldots, b^{(i)}] \in \mathbb{R}^{c \times k}; \)
\State Calculate the weight coefficient vector:
\[ \mathbf{w}^{(i)} = \left( C^{(i)} - 1 \right)_{C^{(i)^{-1}}}^{-1} \in \mathbb{R}^k, \]
where \( 1_{C^{(i)^{-1}}} \) is the inverse of the matrix \( C^{(i)}; \)
\EndFor
\State Calculate matrix:
\[ M = \left( I - \frac{1}{|C_{\text{base}}|} - W \right) \left( I - \frac{1}{|C_{\text{base}}|} - W \right)^{\top} \in \mathbb{R}^{|C_{\text{base}}| \times |C_{\text{base}}|}, \] where \( I - \frac{1}{|C_{\text{base}}|} \in \mathbb{R}^{|C_{\text{base}}| \times |C_{\text{base}}|} \) is a identity matrix and \( W \)’s each element \( W_{j,i} = \left( \mathbf{w}^{(i)} \right)_{j} \) if \( b^{(j)} \) is the \( k \)-th neighbor of \( b^{(i)}; \) otherwise.
\State Obtain dimensionality reduced feature matrix:
\[ \tilde{B} = \left( \tilde{b}^{(1)}, \ldots, \tilde{b}^{(|C_{\text{base}}|)} \right) \in \mathbb{R}^{c \times |C_{\text{base}}|}, \] where \( i \)-th row vector in \( \tilde{B} \) is the eigenvector corresponding to the \((i + 1)\)-th smallest eigenvalue of the matrix \( M; \)
\For {\( q \)}
\State Find the \( k \) nearest neighbors of \( q \) in \( \tilde{B} \) as step 2:
\[ N_{\text{query}} = KNN \left( \tilde{b}^{(i)}_{\text{query}}, \ldots, \tilde{b}^{(k)}_{\text{query}} \right) \in \mathbb{R}^{c \times k}; \]
\State Find \( k \) corresponding dimensionality reduced vectors of \( N_{\text{query}} \) in \( \tilde{B}: \)
\[ \tilde{N}_{\text{query}} = [\tilde{b}^{(1)}_{\text{query}}, \ldots, \tilde{b}^{(k)}_{\text{query}}] \in \mathbb{R}^{c \times k}; \]
\EndFor
\For {\( s^{(n)} \)}
\State Calculate the dimensionality reduced feature vector:
\[ \tilde{s}^{(n)} \in \mathbb{R}^{c'} \] of novel class \( n \) as step 9-12;
\EndFor
\State Calculate nonlinear \( P_{\text{query}}, P_{\text{support}} \) and \( \tilde{q}^{(n)} \) according to Eq. (6) and (7) with dimensionality reduced \( \tilde{B}, \tilde{q}, \tilde{s}^{(n)} \) and weight hyperparameter \( \alpha; \)
\end{algorithmic}
classes, especially in cases where the labeled support samples are severely insufficient.

Cooperative Bi-path Metric with LLE can slightly improve the accuracy compared with the vanilla version in the 5-way 5-shot setting, and it is about 0.9% higher than the state of the art, which illustrates the importance of how to adaptively utilize different base classes for classifying different samples of novel classes. However, the information of base classes is underused in existing methods, we regard this as a promising but underappreciated direction in few-shot learning.

It is worth noting that Cooperative Bi-path Metric classifies query samples based on the same trained backbone of the baseline, and the accuracy improvement over the baseline is stable and not subject to the randomness of the training procedure. Compared to the baseline, Cooperative Bi-path Metric does not introduce additional network parameters or model updating processes, just changes the classification criterion during the testing phase. It is a computationally lightweight method, can be easily integrated into other trained models, such as the Prototypical Networks [9] or Matching Networks [4].

To further illustrate the effectiveness of traditional supervised learning, we also compare our baseline on tieredImageNet [12] with existing methods. As we can see from Tab. 2, our baseline also gets the best results in both settings. Notably, the baseline and the proposed Cooperative Bi-path Metric (as well as the version with LLE) set up a new state of the art in the field of few-shot learning.

### 4.3 Rethinking Few-shot Training Mode

**Is episodic training necessary for few-shot learning?** To make the training mode and testing mode consistent as well as to get better performance, the previous few-shot learning methods adopted the episodic training process. More specifically, similar to the test phase, they also sampled many \( N \)-way \( K \)-shot tasks during the training phase to train a model with the cross-entropy loss of query samples over \( N \) classes in each task. We call this kind of loss few-shot loss. Different from the episodic training mode with few-shot loss, the other training mode does not only focus on the classes within a task but adopts the traditional supervised learning, using the whole training set \( \mathcal{D}_{\text{base}} \) to train a feature extractor and a \( |\mathcal{C}_{\text{base}}| \)-way fully-connected layer. In this mode, the entire network is trained with the cross-entropy loss of samples over all classes in the training set. We call this kind of loss global loss. To study the impact of different training modes on the model’s accuracy, we train TADAM [14] and our proposed baseline in different modes. For convenience, in the experiments involving global loss, we also organize training data in the form of tasks, but only calculate query samples’ global loss. For both modes, during the testing phase, the feature extractor is used to extract the features of all the samples within a task, and query samples are classified into the class with the maximum inductive similarity.

It can be seen from Tab. 3 that for TADAM [14], global loss alone is similar to few-shot loss alone in 1-shot setting, but it is much better than few-shot loss in 5-shot. And it is interesting to find that...
using both global loss and few-shot loss works worse than global loss alone in 5-shot, i.e., the introduction of few-shot loss reduces the performance of global loss. For baseline++, we can also find that global loss is better than few-shot loss in both settings, and the gap between themselves is larger than that for TADAM [14]. Through this experiment, we realize that the episodic training mode is not necessary and its capacity is limited. We also believe that using as much global information as possible to train an efficient feature extractor is important for few-shot learning.

During both training and testing phases, TADAM [14] uses Euclidean distance to measure the distance between samples and classes, while baseline++ uses cosine similarity. We follow the same training strategy as TADAM [14], but the produced results are slightly different from the reported ones.

4.4 Bag of Tricks for Strong Baseline
To explore the influence of the three tricks (data enhancement, temperature, and dense classification) on the baseline, we conduct ablation experiments on miniImageNet. The experimental results are shown in Tab. 4, from which we can find that these tricks can greatly improve the accuracy of the baseline, especially data enhancement.

To further explore the impact of different values of temperature \( t \) on the baseline’s accuracy, we conduct experiments with different values of \( t \), and the corresponding results are shown in Fig. 3. It can be seen from Fig. 3 that the accuracy curves are upward convex, and the highest accuracies are obtained when \( t \) is 0.6 in both settings. All the models proposed in this paper adopt 0.6 as the value of \( t \) on miniImageNet and 0.7 on tieredImageNet.

4.5 Variants of Cooperative Bi-path Metric
Variants of vanilla Cooperative Bi-path Metric. As we can see from the Tab. 4, Cooperative Bi-path Metric, as well as the version with LLE, can further improve classification accuracy by utilizing base classes during testing phase. However, we found that the specific implementation of different details of Cooperative Bi-path Metric has a considerable impact on the performance, and different variants of Cooperative Bi-path Metric have different accuracy. For the vanilla Cooperative Bi-path Metric, it needs to consider the specific implementation of these five details: (i) should cosine similarity or Euclidean distance be used as the similarity function \( \sigma_f \) to calculate \( \rho_{\text{query}} \) and \( \rho_{\text{support}}^{(n)} \) for query samples and support samples? (ii) after obtaining \( \rho_{\text{query}} \) and \( \rho_{\text{support}}^{(n)} \), whether softmax is applied on \( \rho_{\text{query}} \) and \( \rho_{\text{support}}^{(n)} \) to obtain normalized similarity distribution (all components add up to 1)? (iii) should cosine similarity or Euclidean distance or KL divergence be used as the similarity function \( \sigma \) to calculate the transductive similarity \( \phi^{(n)} \) between \( \rho_{\text{query}} \) and \( \rho_{\text{support}}^{(n)} \)? (iv) how to balance the inductive similarity \( \phi^{(n)} \) and the transductive similarity \( \phi^{(n)} \) by weight hyperparameter \( \alpha \)? For (i)-(iii), we enumerate and experiment with all possible combinations. For each combination, the highest accuracy is found through changing \( \alpha \) in range [0, 1] with a interval of 0.05. All results are shown in Tab. 5. It can be found that under the setting of 1-shot, Cooperative Bi-path Metric can greatly improve the accuracy (compared with 64.07), and the value of \( \alpha \) is quite small, i.e., the transductive similarity \( \phi^{(n)} \) plays a major role. However, the result under 5-shot is not ideal. The improvement of accuracy is limited (compared with 80.47), and the value of \( \alpha \) is large, so inductive similarity \( \phi^{(n)} \) is still in the dominant place.

---

**Table 4: Influence of the three tricks on the baseline on 5-way classification on miniImageNet benchmark. DA: data augmentation; \( t \): temperature; DC: dense classification.**

| DA | \( t \) | DC | 1-shot | 5-shot |
|----|----|----|--------|--------|
| ✓ | ✓ | ✓ | 51.68 ± 0.44 | 69.53 ± 0.36 |
| ✓ | ✓ | ✓ | 59.72 ± 0.44 | 77.67 ± 0.34 |
| ✓ | ✓ | ✓ | 52.09 ± 0.44 | 69.80 ± 0.37 |
| ✓ | ✓ | ✓ | 54.00 ± 0.45 | 70.74 ± 0.36 |
| ✓ | ✓ | ✓ | 64.07 ± 0.45 | 80.47 ± 0.33 |

**Table 5: Accuracy of the variants of vanilla Cooperative Bi-path Metric on 5-way classification on miniImageNet benchmark. CS: cosine similarity; ED: Euclidean distance; KL: KL divergence. The best results are bolded.**

| \( \sigma_f \) | softmax | \( \sigma \) | 1-shot | 5-shot |
|-----|-------|-------|--------|--------|
| -   | -     | -     | -      | -      |
| No  | CS    | 0.15  | 64.60 ± 0.46 | 0.75 | 80.48 ± 0.33 |
| ED  | 0.80  | 64.52 ± 0.46 | 1.00 | 80.47 ± 0.33 |
| CS  | 0.05  | 64.77 ± 0.46 | 0.35 | 80.50 ± 0.33 |
| Yes | ED    | 0.05  | 64.75 ± 0.46 | 0.65 | 80.49 ± 0.33 |
| KL  | 0.05  | 64.75 ± 0.46 | 0.50 | 80.49 ± 0.33 |
| No  | CS    | 0.20  | 64.62 ± 0.46 | 0.85 | 80.48 ± 0.33 |
| ED  | 0.95  | 64.20 ± 0.45 | 1.00 | 80.47 ± 0.33 |
| CS  | 0.10  | 64.43 ± 0.45 | 0.70 | 80.48 ± 0.33 |
| Yes | ED    | 0.10  | 64.39 ± 0.45 | 0.80 | 80.48 ± 0.33 |
| KL  | 0.10  | 64.45 ± 0.45 | 0.70 | 80.49 ± 0.33 |
Table 6: Accuracy of the variants of Cooperative Bi-path Metric with LLE on 5-way classification on miniImageNet benchmark. The dimensionality after reduction $c'$ is fixed as 63. The best results are bolded.

| $L_2$ | softmax | $\sigma$ | 1-shot | 5-shot |
|-------|---------|----------|--------|--------|
|       |         | $\sigma_i$ | $k$ | $\alpha$ | Acc. | $k$ | $\alpha$ | Acc. |
| No CS | 8       | 0.95      | 24  | 0.95      | 80.68 ± 0.32 |
| Yes ED | 5       | 0.00      | 63  | 0.95      | 80.49 ± 0.32 |
| Yes ED | 4       | 0.25      | 23  | 0.10      | 80.55 ± 0.32 |
| No CS | 5       | 0.00      | 22  | 0.95      | 80.51 ± 0.32 |
| Yes ED | 7       | 0.95      | 23  | 0.95      | 80.64 ± 0.32 |
| No CS | 6       | 0.95      | 22  | 0.95      | 80.63 ± 0.32 |
| Yes ED | 5       | 0.30      | 23  | 0.70      | 80.51 ± 0.32 |
| No CS | 7       | 0.95      | 23  | 0.95      | 80.64 ± 0.32 |
| Yes ED | 8       | 0.00      | 23  | 0.95      | 80.51 ± 0.32 |

Figure 4: The accuracy curves for different values of $\alpha$ on 5-way classification on miniImageNet benchmark.

Influence of weight hyperparameter. In order to further explore the influence of weight hyperparameter $\alpha$ on classification accuracy (especially when $\alpha = 0$, only transductive similarity $\phi^{(n)}$ works), we report the accuracy curves with different values of $\alpha$, as shown in Fig. 4. It can be seen from Fig. 4, when $\phi^{(n)}$ is the majority, the accuracy is higher, $\phi^{(n)}$ contributes more to the classification than $\phi^{(n)}$. We attribute part of the reason why $\phi^{(n)}$ takes a larger part to that the magnitude of $\phi^{(n)}$ and $\phi^{(n)}$ is different. The accuracy at $\alpha = 0$ is less than the accuracy at $\alpha = 1$, which means that classification result based on $\phi^{(n)}$ alone is less accurate than that based on $\phi^{(n)}$ alone. However, the highest points of the accuracy curves are not obtained at $\alpha = 1$, so $\phi^{(n)}$ can further improve the classification accuracy based on $\phi^{(n)}$.

5 CONCLUSIONS AND FUTURE WORK

In this work, we contribute to few-shot learning with a concise and effective baseline as well as a novel metric named Cooperative Bi-path Metric.

First, we train a simple network in the way of traditional supervised learning as the baseline, which achieves comparable results to the state of the art. This shows that episodic training mode is not necessary, and an effective feature extractor to capture discriminative features of samples is fundamental for few-shot learning.

Second, we propose Cooperative Bi-path Metric to change the criterion of classification. It uses samples' similarity distribution on base classes to assist classification decisions while existing methods did not take full use of such information of base classes. Experiments show that it can further boost the model’s performance and achieve a new state of the art in the field of few-shot image classification, indicating that using base classes to classify samples during the testing phase looks like a promising direction for future research. However, Cooperative Bi-path Metric is handcrafted and somewhat straightforward. A natural direction for improving it is training an additional convolutional neural network end to end to measure the transductive similarity. We leave this for future work.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (No. 2017YFB1002400), the National Natural Science Foundation of China (No. 61922006, No. 61532003, No. 61825101 and No. U1611461), and the Beijing Nova Program (No. Z18110006218063).

REFERENCES

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. IJCV 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In NeurIPS. 1106–1114. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In CVPR. 770–778. https://doi.org/10.1109/CVPR.2016.90

[4] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In NeurIPS. 3630–3638. http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning

[5] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. 2017. Meta-SGD: Learning to Learn Quickly for Few Shot Learning. CoRR abs/1707.09835 (2017).
