Front. Comput. Sci., 2022, 16(5): 165331
https://doi.org/10.1007/s11704-021-1188-9

RESEARCH ARTICLE

Improving meta-learning model via meta-contrastive loss

Pinzhuo TIAN, Yang GAO (✉)

Department of Computer Science and Technology, Nanjing University, Jiangsu 210023, China

Abstract Recently, addressing the few-shot learning issue with meta-learning framework achieves great success. As we know, regularization is a powerful technique and widely used to improve machine learning algorithms. However, rare research focuses on designing appropriate meta-regularizations to further improve the generalization of meta-learning models in few-shot learning. In this paper, we propose a novel meta-contrastive loss that can be regarded as a regularization to fill this gap. The motivation of our method depends on the thought that the limited data in few-shot learning is just a small part of data sampled from the whole data distribution, and could lead to various bias representations of the whole data because of the different sampling parts. Thus, the models trained by a few training data (support set) and test data (query set) might misalign in the model space, making the model learned on the support set can not generalize well on the query data. The proposed meta-contrastive loss is designed to align the models of support and query sets to overcome this problem. The performance of the meta-learning model in few-shot learning can be improved. Extensive experiments demonstrate that our method can improve the performance of different gradient-based meta-learning models in various learning problems, e.g., few-shot regression and classification.

Keywords meta-learning, few-shot learning, metaregularization, deep learning

1 Introduction

Nowadays, deep models sweep many fields, such as classification, segmentation, and object detection. However, most of them require large-scale annotated training data to achieve promising performance. This predicament poses a great challenge to apply the existing deep learning models to some real environments, for example, medical image analysis intrinsically lacking data. Recent literature formulates this problem as a few-shot learning problem, i.e., expecting the deep models can generalize to new concepts with only a few labeled samples. Generally, the learning model for few-shot learning needs some prior knowledge to achieve this goal. Much recent research in few-shot learning notice that a kind of technology, named meta-learning, can automatically learn cross-task meta-knowledge as the prior knowledge to help the learning algorithm perform on new tasks. It makes the meta-learning framework is fit for few-shot learning. Hence, much literature has implemented this idea and achieve great success in few-shot learning [1−3].

As we know, regularization techniques are helpful to improve the learning algorithm and are widely adopted in the machine learning community. For example, penalizing the ℓ_p-norm of the weights in feature selection [4], Kullback-Leibler (KL) divergence to constrain the approximate posterior in Bayesian Learning [5], dropping out random units or filters in deep models [6], and so on. However, back to few-shot learning, rare research exploits appropriate regularization techniques to improve the performance of meta-learning methods in the few-shot settings.

To overcome this problem, in this paper, we focus on how to develop a meta-regularization to regularize the learned meta-knowledge for improving the gradient-based meta-learning method in few-shot learning. In the machine learning community, regularization techniques are always built on human prior knowledge on the learning tasks. Our method also utilizes the understanding of good meta-knowledge in few-shot learning to design an appropriate meta-regularization. Considering that in the few-shot learning task, we usually contact a training dataset (a.k.a., support set) containing a few annotated data and a test dataset (a.k.a., query set) containing numerous test data. The limited data cause a bias representation of the whole dataset because it is usually sampled from a small part of the data distribution. This problem leads to the data discrepancy between support and query set. Hence, the model learned from support set can not fit the query set well. Figure 1 shows the overview of the problem of biased representation. *Effective meta-knowledge* should help the learning model trained by the support set generalize well on the query set. Based on this cogitation, we assume that the models trained by the support and query set should be aligned to each other in the model space with the help of meta-knowledge. For example, in the 2-way classification task, the classification hyperplanes learned respectively by support and query set should be aligned to make the hyperplane learned by the support set fit the query set well, otherwise, the classification hyperplane of the support set can not generalize well on the query set. Thus, we focus on designing a meta-regularization to improve the ability of meta-knowledge on

Received April 20, 2021; accepted June 9, 2021

E-mail: gaoy@nju.edu.cn

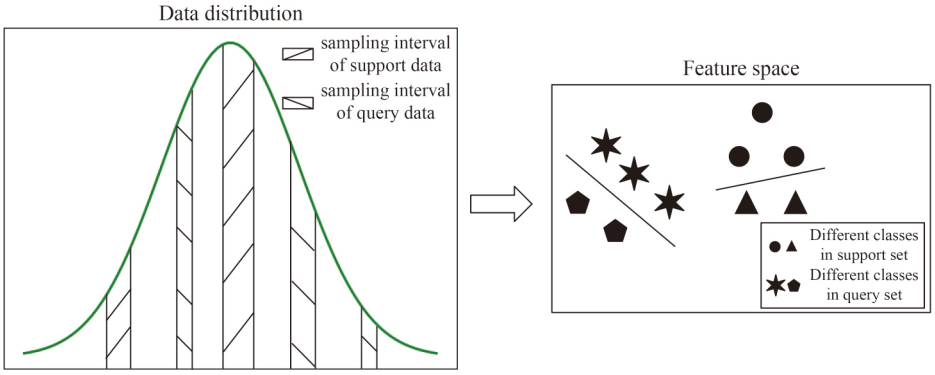

Fig. 1 The motivation of our method. Too limited data in few-shot learning leads to data discrepancy of support and query set, because they are sampled from different parts. This problem causes the model learned by the support set can not generalize well on the query set

aligning the models of support and query set in the model space. Inspired by recent contrastive learning algorithms [7,8], we consider designing a meta-contrastive loss to maximize agreement of the models trained by the support and query set to achieve this goal and further improve the meta-learning model in few-shot learning.

We conduct abundant experiments to demonstrate the effectiveness of our method. The experiments show that the proposed method is model agnostic and can be easily inserted into the hierarchical gradient-based meta-learning framework as a regularization to improve the different meta-learning models in few-shot learning. Moreover, compared with other techniques which can also improve the gradient-based model, e.g., scale factor, our method can be applied to many learning problems, i.e., classification and regression.

The contributions of this paper are summaries as follow:

- We consider the fact that too limited data in few-shot learning task can lead to the data discrepancy between support and query set. Under this consideration, we propose a novel meta-contrastive loss to improve the performance of gradient-based meta-learning models in few-shot learning by helping the learned meta-knowledge to eliminate this discrepancy.
- Compared with the traditional contrastive loss in unsupervised learning, our method focuses on the task level and deals with how to align the parameter matrix in the model space. However, the traditional contrastive loss aims to align the feature vector in the feature space.
- Extensive experiments show that our method can improve the performance of various gradient-based meta-learning models and work well in few-shot classification and regression.

2 Related work

2.1 Meta-learning for few-shot learning

Recently, meta-learning framework is widely used to overcome the few-shot learning problem. Meta-learning method in few-shot learning can be broadly divided into three categories, metric-based method, model-based method, and gradient-based method.

- Metric-based method: The motivation of this method can be introduced as learning to comparison. For example, Matching networks [9] use recurrent neural network with attention block as the embedding model to learn how to evaluate the similarity between examples in Euclidean space. Prototypical network [2] represents each category by a prototype (a.k.a., mean embedding of the examples) and utilize Euclidean distance to measure the similarity between test images and the prototypes. Different from the predefined metric space, Sung et al. [10] use a neural network to automatically learn the metric function. However, it is difficult to design an appropriate metric to measure the similarity between the data in some learning problem, which restricts metric-based method to the classification task.
- Model-based method: This method usually adopts an extra memory to store the past experience or an elaborate system to lead optimization of the learning model in the low data region. Ravi et al. [11] used recurrent neural network as a high-level model to direct the updating of the learner in the specific task. Santoro et al. [12] used an external memory-augmented neural network to save the seen examples and leveraged them to predict the results with a few examples. This kind of method is usually very complex and difficult to train.
- Gradient-based method: Gradient-based method usually utilizes the hierarchical architecture to learn meta-knowledge. Model-agnostic meta learning (MAML) [1] aims to learn a good initialization for the task-specific learning model. In the new task, the task-specific learner can be obtained by a few gradient steps from this initialization. However, there are still many limitations of MAML. Many works [13,14] are proposed to improve it. Besides MAML, some gradient-based approaches [15,16] built on bilevel optimization framework [17] aim to learn a cross-task representation as meta-knowledge to help learn new tasks. Almost all the gradient-based methods adopt an inner-loop learning process, causing that how to efficiently optimizing the gradient-based approaches becomes a problem.

Compared with metric-based and model-based methods,

gradient-based approaches attract more attention in few-shot learning, because they can be applied to many learning problems and the time consumption is acceptable in the few-shot setting. In this work, we focus on how to improve the performance of the gradient-based method in the few-shot setting.

2.2 Meta-regularization

Regularization is a useful technique in machine learning to explicitly design to improve the learning algorithm. In this part, we firstly introduce some existing regularizations in deep models. Dropout [6] is a regularization to randomly drop units (along with their connections) from the neural network during training to prevent units from co-adapting too much. Early stopping is another popular regularization method in deep learning due to both its effectiveness and simplicity [4]. In the meta-learning models, Balaji et al. [18] proposed a meta-regularization to achieve good cross-domain generalization. Tseng et al. [19] used gradient dropout to mitigate the over-fitting problem in meta-learning, however, it is customized to the MAML-based methods.

2.3 Contrastive learning

Contrastive learning is a hot topic in the machine learning community and at the core of several recent works on unsupervised learning [20]. Much research in contrastive learning uses contrastive losses to measure the similarities of sample pairs in a representation space to learn the data representation. This representation can be applied to many downstream tasks. Some literature owes the success of their methods to maximization of mutual information between latent representations [21]. However, it is not explicit if the success of contrastive approaches is determined by the mutual information, or by the specific form of the contrastive loss [22].

3 Preliminary

In this section, we describe the problem definition of few-shot learning. Then, the hierarchical framework used in gradient-based method is introduced. Our method aims to improve the performance of the meta-learning model built on this framework.

3.1 Problem definition

The standard supervised learning problem considers learning a function $x \mapsto \hat{y}$ by a set of training data (x_i, y_i) indexed by i and sampled from a task $\mathcal{T}$. In few-shot learning, we usually have a set of tasks $\mathcal{S} = \{\mathcal{T}^i\}_{i=1}^{T}$ as the training examples to learn the prior knowledge, which can help the task-specific learner effectively learn in the new task with a few labeled data. Similar to the standard supervised learning problem, each task $\mathcal{T}^i$ is made of a training dataset (a.k.a., support set) S_s^i, and a testing dataset (a.k.a., query set) S_q^i. However, the number of training data in support set is very small in few-shot learning.

According to the different learning problems, the form of the training task is different. For example, in few-shot classification, the learning task $\mathcal{T}^i$ is described as N-way K-shot classification task, indicating N categories given K samples per category, i.e., $S_s^i = \{(x_j^i, y_j^i)\}_{j=1}^{N\times K}$. The query set S_q^i also consists of the same N categories and each category has Q examples. As for few-shot regression, the learning task $\mathcal{T}^i$ can be considered as a K-shot regression task, i.e., the support set S_s^i and the query set S_q^i consist of K training data and Q test data, respectively.

3.2 Hierarchical gradient-based method

In few-shot learning, the purpose of meta-learning is to learn useful meta-knowledge over $\mathcal{S}$ as the prior knowledge to help the learner learn in the low-data region. Hence, how to learn useful meta-knowledge over $\mathcal{S}$ is crucial for the meta-learning model. Current gradient-based meta-learning methods are highly related to the hierarchical architecture. This architecture can be optimized as a bilevel optimization problem [17]. Following [23], the two-level meta-learning framework can be defined as:

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}^{\text{meta}}(\theta, w^i(\theta); S_q^i), \tag{1}$$

$$\text{s.t.} \quad w^i(\theta) = \min_{\theta} \mathcal{L}(w; \theta, S_s^i), \tag{2}$$

where $\mathcal{L}^{\text{meta}}$ and $\mathcal{L}$ refer to the functions of meta loss (as the outer objective in bilevel optimization) and task loss (as the inner objective in bilevel optimization). In fact, $\mathcal{L}^{\text{meta}}$ and $\mathcal{L}$ usually adopt the same loss function.

Note that the inner part (a.k.a., base learner) Eq. (2) aims to learn a task-specific learner for every single task with the support set S_s^i, while, the upper part (a.k.a., meta-learner) Eq. (1) learns meta-knowledge from how to improve these base-learners with the query sets. In this way, the learned meta-knowledge can help learn unseen tasks.

Under this formulation, MAML can be written as:

$$\min_{\theta} \frac{1}{T} \sum_{i=1}^{T} \mathcal{L}(\theta, w^i; S_q^i), \tag{3}$$

$$\text{s.t.} \quad w^i = \theta - \alpha \nabla_{\theta} \mathcal{L}(\theta; S_s^i), \tag{4}$$

where α is the stepsize. $\theta - \alpha \nabla_{\theta} \mathcal{L}(\theta; S_s^i)$ means one step of the inner updating, and aims to obtain a base-learner for task $\mathcal{T}^i$. When encountering a new task $\mathcal{T}^j$, the task-specific predictor can be easily obtained in a single (or a few) inner gradient step from the initial θ.

Different from MAML, MetaOpt [15] and R2D2 [16] use support vector machine [24] and ridge regression as the base-learner in Eq. (2), respectively. Both of them want to learn a cross-task meta-representation, which can improve the base-learner of a new task in the low-data region.

4 Method

Although gradient-based meta-learning methods achieve success in few-shot learning, little literature pays attention to develop meta-regularization for the meta-learning method to improve the performance in few-shot learning.

To overcome this problem, we propose a novel meta-contrastive loss to regularize the learned meta-knowledge for better generalization. Our method is built on a prior under-

standing that the models trained by the support and query set should be aligned well in the model space in each few-shot task. In this way, the base-learner learned from a few supervised information can generalize well on the query set which may contain many unseen cases.

In order to achieve the purpose of helping meta-learning algorithm align the models trained by the support and query set, inspired by the role of the contrastive loss in contrastive learning, we develop a meta-contrastive loss to address how to align the learning models. Review contrastive learning, the contrastive loss is designed to maximize agreement between differently augmented views of the same data example in the latent space to learn representation. We can find that the feature vectors are needed to be matched in contrastive learning. In our method, we utilize this idea to align models. However, the learning model might contain multiple components, e.g., a weight matrix, not a weight vector. For example, in 5-way classification task, the parameter matrix of base-learner in MetaOpt or R2D2 includes five vectors, corresponding to five classes, i.e., $\mathbf{W}=[\mathbf{w}^1,\ldots,\mathbf{w}^5],\mathbf{W}\in\mathbb{R}^{d\times 5}$ and $\mathbf{w}^i\in\mathbb{R}^d$. d is the dimension of feature vector.

Hence, the weight vector is considered as "data" in our meta-regularization loss that is different from contrastive learning (image is the data). In our method, the weight vectors related to the same category are supposed to be aligned in the model space. Typically, suppose that we randomly sample a training task $\mathcal{T}$ from $\mathcal{S}$, the support and query set in $\mathcal{T}$ are defined as S_s and S_q, respectively. The learning models of S_s and S_q can be obtained by Eq. (2) via different meta-learning methods, which are defined by $\mathbf{W}_s$ and $\mathbf{W}_q$, respectively. Consider that $\mathbf{W}_s$ and $\mathbf{W}_q$ involve C weight vectors, i.e., $\mathbf{W}_s=[\mathbf{w}_s^1,\ldots,\mathbf{w}_s^C]\in\mathbb{R}^{d\times C},\mathbf{W}_q=[\mathbf{w}_q^1,\ldots,\mathbf{w}_q^C]\in\mathbb{R}^{d\times C}$, the positive pairs are $(\mathbf{w}_s^i,\mathbf{w}_q^i),i=1,\ldots,C$. Then the proposed meta-contrastive loss for a positive pair of examples can be written as:

$$\ell_{i,i}=-\log\frac{\exp\left(\mathrm{sim}\left(\mathbf{w}_s^i,\mathbf{w}_q^i\right)/\tau\right)}{\sum_{k=1}^{C}\mathbb{1}_{[k\neq i]}\exp\left(\mathrm{sim}\left(\mathbf{w}_s^i,\mathbf{w}_q^k\right)/\tau\right)},\tag{5}$$

where $\mathbb{1}_{[k\neq i]}\in\{0,1\}$ is an indicator function evaluating to 1 iff $k\neq i$ and τ denotes a temperature parameter. The $\mathrm{sim}(\boldsymbol{u},\boldsymbol{v})=\boldsymbol{u}^\top\boldsymbol{v}/\|\boldsymbol{u}\|\,\|\boldsymbol{v}\|$ denotes the dot product between ℓ_2 normalized $\boldsymbol{u}$ and $\boldsymbol{v}$ (i.e., cosine similarity). The final loss is computed across all positive pairs in $\mathbf{W}_s$ and $\mathbf{W}_q$, both $(\mathbf{w}_s^i,\mathbf{w}_q^i)$ and $(\mathbf{w}_q^i,\mathbf{w}_s^i)$.

In some situations, the learning models of the support and query set can be a weight vector, for example, in the 1-dimension regression. We sample a minibatch of N tasks and define the models of support and query set from the same task as a positive pair in this case. The proposed meta-contrastive loss can be written as:

$$\ell_{i,i}=-\log\frac{\exp\left(\mathrm{sim}\left(\mathbf{w}_s^i,\mathbf{w}_q^i\right)/\tau\right)}{\sum_{k=1}^{N}\mathbb{1}_{[k\neq i]}\exp\left(\mathrm{sim}\left(\mathbf{w}_s^i,\mathbf{w}_q^k\right)/\tau\right)}.\tag{6}$$

Algorithm 1 summarizes the proposed method.

Remarks We put forward some insights into our method. Let us rethink the process of training a learning model for a dataset. Suppose that we have a training dataset S, the learning model $\mathcal{M}_{\mathbf{w}}$ parameterized by $\mathbf{w}$ for S can be obtained by optimizing a loss function. With an optimization algorithm $f(\mathbf{w};S)$, such as gradient descent, the learned parameter $\mathbf{w}$ can be obtained via $\mathbf{w}=f(\mathbf{w}_{init};S)$. As we can see, there exists a one-to-one mapping between the dataset S and the learned parameter $\mathbf{w}$. From this perspective, the learned model $\mathcal{M}_{\mathbf{w}}$ can be regarded as a representation of the corresponding dataset S in the model space. Concentrating on few-shot learning, the base-learner $\mathcal{M}_{\mathbf{w}}^s$ trained by the support set S_s should also be a good representation of the query set S_q. Aligning $\mathcal{M}_{\mathbf{w}}^s$ and $\mathcal{M}_{\mathbf{w}}^q$ in the model space by the proposed meta-regularization influences the data representation in support and query set and implicitly regularize the learned meta-knowledge. Figure 2 shows a simple interpretation of our method.

Algorithm 1 Meta-contrastive loss's main learning algorithm

Require: A meta-training set $\mathcal{S}=\{\mathcal{T}^i\}_{i=1}^T$, constant τ

while not done **do**
  Randomly sample a task $\mathcal{T}$ containing a support set S_s and a query set S_q
  Train a base-learner $\mathbf{W}_s=[\mathbf{w}_s^1,\cdots,\mathbf{w}_s^C]\in\mathbb{R}^{d\times C}$ with S_s via Eq. (2)
  Train a learner $\mathbf{W}_q=[\mathbf{w}_q^1,\cdots,\mathbf{w}_q^C]\in\mathbb{R}^{d\times C}$ with S_q via Eq. (2)
  for all $i\in\{1,\ldots,C\}$ and $j\in\{1,\ldots,C\}$ **do**
    $s_{i,j}=\mathbf{w}_s^{i\top}\mathbf{w}_q^j/(\|\mathbf{w}_s^i\|\,\|\mathbf{w}_q^j\|)$
  end for
  $\mathcal{L}_{ours}=\frac{1}{2C}\sum_{i=1}^{C}[\ell(\mathbf{w}_s^i,\mathbf{w}_q^i)+\ell(\mathbf{w}_q^i,\mathbf{w}_s^i)]$ via Eq. (5)
  Compute meta loss $\mathcal{L}$ via Eq. (1)
  Optimize the meta learning method via $\mathcal{L}+\mathcal{L}_{ours}$
end while

5 Experiments

In experiments, we evaluate our method in three challenging

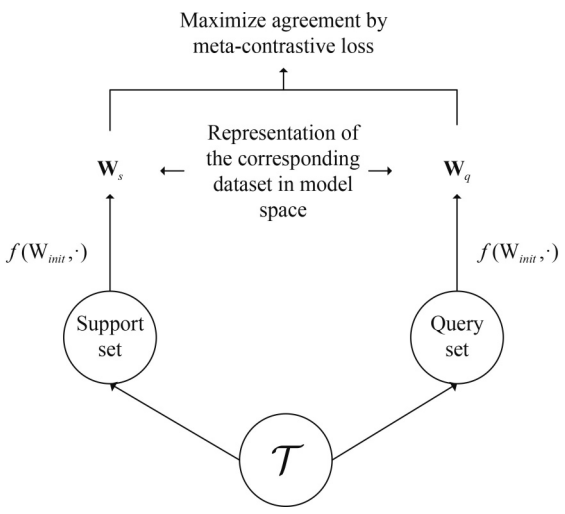

Fig. 2 A simple illustration of meta-contrastive loss. In few-shot learning, each task $\mathcal{T}$ contains a support set and a query set. In our method, we use meta-contrastive loss to align the models of support and query set to eliminate the influence of the bias representation, caused by the limited data. Compared with the traditional contrastive loss to align the feature vector, our method needs to deal with how to maximize agreement of the models with the parameter matrix

scenarios, i.e., few-shot regression, few-shot classification, and few-shot fine-grained classification. Because our method focuses on how to design an appropriate meta-regularization to improve the gradient-based meta-learning model, three state-of-the-art gradient-based meta-learning methods, i.e., ANIL [14], MetaOpt [15], and R2D2 [16] are chosen as the benchmark algorithms. To accurately show the effect of our method, all the tricks used in these benchmark methods for improving their performance to the state of the art are abandoned in our implementation.

5.1 Few-shot regression

Experimental setup. Our experimental setup of few-shot regression follows [1]. Each task involves regressing from the input to the output of a sine wave. The amplitude and phase of the training tasks are uniformly sampled from $[0.1, 5.0]$ and $[0, \pi]$, respectively. The purpose of meta-learning model is to fit the unseen sine curves with a few training data. In the training stage, the labeled datapoints are uniformly sampled from $[-5.0, 5.0]$ as S_s, and twenty labeled datapoints are given as S_q in each task also drawn from $[-5.0, 5.0]$. We adopt a neural network with one hidden layer containing 40 nodes as the feature encoder and mean-squared error (MSE) as meta loss.

Because MetaOpt and R2D2 are tailored to classification, ANIL is adopted as the baseline model in this scenario. The batchsize is set to 5. All models are trained $7,000$ iterations by Adam [25] with a learning rate of 0.001. The number of inner updating step is 5 and the inner learning rate is 0.01. During the test, we present the model with $2,000$ newly sampled tasks with 100 test points in each task.

Results. Table 1 shows the results of different models. ANIL-ours indicates that ANIL integrated with meta-contrastive loss. We can see that our method improves the generalization of ANIL on the unseen new tasks. The results also verify that our method can work well in the regression problem.

5.2 Few-shot classification

Dataset. We evaluate our method on two few-shot image classification datasets: miniImageNet [9] and tieredImageNet [26].

1. MiniImageNet consists of 100 randomly chosen classes from ImageNet [27]. These classes are randomly split into 64, 16, and 20 classes for meta-training, meta-validation, and meta-testing respectively. Each class contains 600 images.

2. TieredImageNet benchmark is a larger subset of ImageNet, composed of 608 classes grouped into 34 high-level categories. These are divided into 20 categories for meta-training, 6 categories for meta-validation, and 8 categories for meta-testing respectively, corresponding to 351, 97, and 160 classes for each split.

Experimental setup. Consider that the performance of meta-learning model in few-shot learning is influenced by the architecture of the embedding model. Two feature extractors are adopted, four-layer ConvNet in [2] and ResNet-12 in [28]. ConvNet has 4 modules with a 3×3 convolution with 64 filters, followed by a batch normalization, a ReLU nonlinearity, and a 2×2 max-pooling. ResNet-12 uses four residual blocks with 64, 128, 256, and 512 filters, respectively, and each block consists of three {3×3 convolution with k filters, batch normalization, ReLU} followed a 2×2 max-pooling layer. In ANIL, we adopt one fully connected layer containing $1,600$ hidden units as the classification head.

All the images are resized to 84×84. Adam with a learning rate of 0.001 is used as the meta-optimizer. The inner learning rate of ANIL is 0.01, and the step of inner gradient descent is 5. All the models are trained by $30,000$ iterations, and each iteration includes five training tasks. The number of query data in each training task is 10.

Results. We show the results of ANIL, MetaOpt, and R2D2 with our method or not on 5-way classification on miniImageNet and tieredImageNet in Tables 2 and 3, respectively. All the reported results are averaged over $2,000$ tasks randomly sampled from the meta-testing set. Each task contains 10 queries per category.

As seen, with meta-contrastive loss, three baselines, ANIL, MetaOpt, and R2D2 achieve better generalization on the unseen tasks. Compared with miniImageNet, tieredImageNet is more challenging for few-shot classification. Our method also can work well. Although the deeper network can improve the performance of baseline models, the effectiveness of our method to help meta-learning model align the learners trained by support and query set is obvious. In this sense, using deeper networks can not replace the effectiveness of the proposed meta-regularization.

Table 1 Mean square error of few-shot regression. Lower is better

Methods	5-shot	10-shot
ANIL	0.746 ± 0.044	0.354 ± 0.018
ANIL-ours	**0.744 ± 0.044**	**0.345 ± 0.018**

Table 2 Accuracy(%) of 5-way classification on miniImageNet

Methods	Embedding	miniImageNet 5-shot
ANIL	ConvNet	58.51 ± 0.46
ANIL-ours	ConvNet	**60.11 ± 0.46**
R2D2	ConvNet	56.79 ± 0.41
R2D2-ours	ConvNet	**61.70 ± 0.41**
MetaOpt	ConvNet	64.06 ± 0.41
MetaOpt-ours	ConvNet	**65.80 ± 0.40**
R2D2	ResNet12	58.48 ± 0.43
R2D2-ours	ResNet12	**70.04 ± 0.40**
MetaOpt	ResNet12	66.64 ± 0.41
MetaOpt-ours	ResNet12	**68.49 ± 0.42**

Table 3 Accuracy(%) of 5-way classification on tieredImageNet

Methods	Embedding	tieredImageNet 5-shot
ANIL	ConvNet	58.64 ± 0.49
ANIL-ours	ConvNet	**61.72 ± 0.50**
R2D2	ConvNet	59.53 ± 0.45
R2D2-ours	ConvNet	**64.21 ± 0.45**
MetaOpt	ConvNet	63.97 ± 0.46
MetaOpt-ours	ConvNet	**65.75 ± 0.45**
R2D2	ResNet12	60.10 ± 0.46
R2D2-ours	ResNet12	**70.21 ± 0.45**
MetaOpt	ResNet12	66.02 ± 0.46
MetaOpt-ours	ResNet12	**71.82 ± 0.47**

5.3 Few-shot fine-grained classification

Dataset. For few-shot fine-grained classification, we use the CUB-200-2011 [29] (referred to as CUB2011 hereafter) as the dataset that is widely adopted in much previous literature. This dataset contains 200 classes and 11,788 images in total. We follow the same class split proposed in [30] to construct the experiment of few-shot fine-grained classification.

Experimental setup. We use ConvNet in few-shot classification as the embedding model. The same image size is adopted. Adam with the same learning rate in few-shot classification is also used to optimize ANIL, R2D2, and MetaOpt. All the models are trained by 20,000 iterations, and each iteration includes one training task. The number of test data in each training task is also 10.

Results. Table 4 shows the results on CUB2011. All the results are averaged over 2,000 new tasks, and each task contains 10 query images per category. The proposed method can improve the performance of all the benchmark meta-learning methods, similar to few-shot classification. The experiments on different datasets also prove that the proposed method can work well in situations containing different data discrepancies.

5.4 Meta-contrastive loss versus scale factor

Previous works show that many techniques can be used to improve the performance of meta-learning models in few-shot learning. Multi-task learning and scale factor are two effective techniques widely used in many meta-learning models. In few-shot classification, we show the performance of our method under the multi-tasking training. In this part, we compare our method with scale factor in few-shot regression to exhibit the superiority of meta-contrastive loss. The learnable scale factor can improve performance by adjusting the prediction score predicted by the base-learner, which is customized for few-shot classification.

Table 5 shows the results. In few-shot regression, ANIL with scale factor even underperforms ANIL. However, integrated with our method, ANIL-ours achieves better generalization. This phenomenon demonstrates our method can be applied to many learning problems.

6 Conclusion

In this paper, we focus on how to design an appropriate meta-regularization to improve the performance of gradient-based meta-learning model in few-shot learning. Inspired by contrastive learning, we propose a meta-contrastive loss to help meta-learning model align the learners trained by the support and query set. In this way, the meta-learning model is supposed to learn better meta-knowledge. The experimental results show that our method is model agnostic and can improve different gradient-based methods in various few-shot scenarios. The analysis demonstrates that our method can be applied to different learning problems. In the future, we will aim to improve our method in the 1-shot setting.

Table 4 Accuracy(%) of 5-way classification on CUB2011

Methods	CUB2011 5-shot
ANIL	71.82 ± 0.49
ANIL-ours	**73.91 ± 0.47**
R2D2	75.73 ± 0.40
R2D2-ours	**77.23 ± 0.38**
MetaOpt	74.88 ± 0.42
MetaOpt-ours	**75.92 ± 0.41**

Table 5 Mean square error of different methods in few-shot regression

Method	Our method	Scale factor	5-shot	10-shot
ANIL			0.746	0.354
	√		0.744	0.345
		√	2.561	1.811

Acknowledgements This work was supported in part by the Science and Technology Innovation 2030 "New Generation Artificial Intelligence" Major Project (2018AAA0100905).

References

1. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. 2017, 1126−1135
2. Snell J, Swersky K, Zemel R S. Prototypical networks for few-shot learning. 2017, arXiv preprint arXiv: 1703.05175
3. Tian P, Wu Z, Qi L, Wang L, Shi Y, Gao Y. Differentiable meta-learning model for few-shot semantic segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12087−12094
4. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press, 2016
5. Kingma D P, Welling M. Auto-encoding variational bayes. 2014, arXiv preprint arXiv: 1312.6114
6. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958
7. Chen T, Kornblith S, Norouzi M, Hinton G E. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. 2020, 1597−1607
8. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 9726−9735
9. Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D. Matching networks for one shot learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016, 3637−3645
10. Sung F, Yang Y, Zhang L, Xiang T, Torr P H S, Hospedales T M. Learning to compare: relation network for few-shot learning. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 1199−1208
11. Ravi S, Larochelle H. Optimization as a model for few-shot learning. In: Proceedings of the 5th International Conference on Learning Representations. 2017
12. Santoro A, Bartunov S, Botvinick M, Wierstra D, Lillicrap T. One-shot learning with memory-augmented neural networks. 2016, arXiv preprint arXiv: 1605.06065
13. Lee H B, Lee H, Na D, Kim S, Park M, Yang E, Hwang S J. Learning to balance: bayesian meta-learning for imbalanced and out-of-distribution tasks. 2020, arXiv preprint arXiv: 1905.12917
14. Raghu A, Raghu M, Bengio S, Vinyals O. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In: Proceedings of the 33rd Conference on Neural Information Processing Systems. 2019
15. Lee K, Maji S, Ravichandran A, Soatto S. Meta-learning with

differentiable convex optimization. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. 1064: 9–10657

16. Bertinetto L, Henriques J F, Torr P H S, Vedaldi A. Meta-learning with differentiable closed-form solvers. In: Proceedings of the 7th International Conference on Learning Representations. 2019
17. Sinha A, Malo P, Deb K. A review on bilevel optimization: from classical to evolutionary approaches and applications. IEEE Transactions on Evolutionary Computation, 2018, 22(2): 276–295
18. Balaji Y, Sankaranarayanan S, Chellapp. R. MetaReg: towards domain generalization using meta-regularization. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018, 1006−1016
19. Tseng H Y, Chen Y W, Tsai Y H, Liu S, Lin Y Y, Yang M H. Regularizing meta-learning via gradient dropout. In: Proceedings of the 15th Asian Conference on Computer Vision. 2020, 218–234
20. Jaiswal A, Babu A R, Zadeh M Z, Banerjee D, Makedon F. A survey on contrastive self-supervised learning. Technologies, 2021, 9(1): 2
21. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. 2018, arXiv preprint arXiv: 1807.03748
22. Tschannen M, Djolonga J, Rubenstein P K, Gelly S, Lucic M. On mutual information maximization for representation learning. In: Proceedings of the 8th International Conference on Learning Representations. 2020
23. Franceschi L, Frasconi P, Salzo S, Grazzi R, Pontil M. Bilevel programming for hyperparameter optimization and meta-learning. In: Proceedings of the 35th International Conference on Machine Learning. 2018, 1568−1577
24. Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273–297
25. Kingma D, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations. 2015
26. Ren M, Triantafillou E, Ravi S, Snell J, Swersky K, Tenenbaum J B, Larochelle H, Zemel R S. Meta-learning for semi-supervised few-shot classification. In: Proceedings of the 6th International Conference on Learning Representations. 2018
27. Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of 26th Annual Conference on Neural Information Processing Systems. 2012, 1106−1114
28. Oreshkin B N, Rodriguez P, Lacoste A. TADAM: task dependent adaptive metric for improved few-shot learning. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018, 719–729
29. Welinder P, Branson S, Mita T, Wah C, Schroff F, Belongie S, Perona P. Caltech-UCSD birds 200. CNS-TR-2010-001. Pasadena: California Institute of Technology, 2010
30. Chen W Y, Liu Y C, Kira Z, Wang Y C F, Huang J B. A closer look at few-shot classification. In: Proceedings of the 7th International Conference on Learning Representations. 2019

Pinzhuo Tian is currently working towards his PhD degree with the State Key Lab for Novel Software Technology, Department of Computer Science Technology, Nanjing University, China. His research interests lie in machine learning, including meta-learning and transfer learning.

Yang Gao received his PhD degree from the Department of Computer Science and Technology of Nanjing University, China in 2000. He is a professor at the Department of Computer Science and Technology, Nanjing University, China. His research interests include artificial intelligence and machine learning. He has published more than 100 papers in top conferences and journals in and outside of China. He is a member of IEEE.