Research on image classification method based on improved multi-scale relational network

Wenfeng Zheng [1] , **Xiangjun Liu** [1] , **Lirong Yin** [Corresp. 2]

[1] School of Automation, University of Electronic Science and Technology of China, Chengdu, China

[2] Department of Geography and Anthropology, Louisiana State University and Agricultural and Mechanical College, Baton Rouge, Louisiana, United States

Corresponding Author: Lirong Yin
Email address: lyin5@lsu.edu

Small sample learning aims to learn information about object categories from a single or a few training samples. This learning style is crucial for deep learning methods based on large amounts of data. The deep learning method can solve small sample learning through the idea of meta-learning "how to learn by using previous experience." Therefore, this paper takes image classification as the research object to study how meta-learning quickly learns from a small number of sample images. The main contents are as follows:

After considering the distribution difference of data sets on the generalization performance of measurement learning and the advantages of optimizing the initial characterization method, this paper adds the model-independent meta-learning algorithm and designs a multi-scale meta-relational network. First, the idea of META-SGD is adopted, and the inner learning rate is taken as the learning vector and model parameter to learn together. Secondly, in the meta-training process, the model-independent meta-learning algorithm is used to find the optimal parameters of the model. The inner gradient iteration is canceled in the process of meta-validation and meta-test. The experimental results show that the multi-scale meta-relational network makes the learned measurement have stronger generalization ability, which further improves the classification accuracy on the benchmark set and avoids the need for fine-tuning of the model-independent meta-learning algorithm.

The image classification method based on improved Multi-scale relational network

6 Wenfeng Zheng[1]¶, Xiangjun Liu[1]¶, Lirong Yin[2]*

8 [1] School of Automation, University of Electronic Science and Technology of China,Chengdu
9 610054, Sichuan, P. R. China
10 [2] Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA
11 70803, United States

13 Corresponding author:
14 Lirong Yin[2]
15 Louisiana State Univ, Dept Geog & Anthropol, Baton Rouge, LA 70803 USA.
16 E-mail: yin.lyra@gmail.com, lyin5@lsu.edu(LY)

Abstract

19   Small sample learning aims to learn information about object categories from a single or a
20 few training samples. This learning style is crucial for deep learning methods based on large
21 amounts of data. The deep learning method can solve small sample learning through the idea of
22 meta-learning "how to learn by using previous experience." Therefore, this paper takes image
23 classification as the research object to study how meta-learning quickly learns from a small
24 number of sample images. The main contents are as follows:
25   After considering the distribution difference of data sets on the generalization performance
26 of measurement learning and the advantages of optimizing the initial characterization method,
27 this paper adds the model-independent meta-learning algorithm and designs a multi-scale meta-
28 relational network. First, the idea of META-SGD is adopted, and the inner learning rate is taken
29 as the learning vector and model parameter to learn together. Secondly, in the meta-training
30 process, the model-independent meta-learning algorithm is used to find the optimal parameters
31 of the model. The inner gradient iteration is canceled in the process of meta-validation and meta-
32 test. The experimental results show that the multi-scale meta-relational network makes the
33 learned measurement have stronger generalization ability, which further improves the

34   classification accuracy on the benchmark set and avoids the need for fine-tuning of the model-
35   independent meta-learning algorithm.
36   **Keywords**: less sample learning, meta-learning, multi-scale characteristics, model-independent,
37   image classification, META-SGD, Multi-scale relational network

Introduction

39   Deep learning has made significant progress in computer vision fields, but only on the
40   premise that they have a large amount of annotated data (Ni et al., 2019; Vinyals, Blundell,
41   Lillicrap, & Wierstra, 2016; Zheng, Liu, & Yin, 2021). However, it is impractical to acquire large
42   amounts of data in real life. As far as deep Learning is concerned, fitting into a more complex
43   model requires more data to have good generalization ability. Once there is a lack of data, deep
44   learning technology can make the in-sample training effect good, but the generalization
45   performance of new samples is poor. Inspired by the human ability to learn quickly from a small
46   sample, many researchers have become increasingly aware of the need to study machine learning
47   from a small sample. In recent years, small sample learning has become a very important frontier
48   research direction in deep Learning. Human beings can learn new concepts from a single or a few
49   samples and obtain extremely rich representations from sparse data. This ability is attributed to
50   people's ability to realize and control their learning process, called meta-learning (Biggs, 1985;
51   Ding et al., 2019; Ma, Zheng, Chen, & Yin, 2021; Yin et al., 2019).
52   Up to now, research on learning with fewer samples can be divided into two aspects: 1.
53   Generation model based on probabilistic reasoning; 2. Discriminant model based on meta-learning.
54   The pioneering work of low-sample Learning can be traced back to the work of Li Feifei et
55   al., in early 2000, which defined the concept of low-sample Learning: learning the new category
56   by using one or a few samples of the image of the new category (Fe-Fei, 2003). Li Fei-Fei proposed
57   a Bayesian learning framework in 2004. Andrew L. Maas et al. used the Bayesian network to
58   capture the relationship between attributes in 2009, which can deal with near-deterministic
59   relationships and include soft probabilistic relationships (Kemp & Maas, 2009; Li, Zheng, Wang,
60   Yin, & Wang, 2015). In 2016, Danilo J. Rezende et al. implemented the hierarchical Bayesian
61   programming learning framework in the deep generation model based on feedback and attention
62   principles. Compared with the shallow hierarchical Bayesian programming learning framework
63   (B. M. Lake, Salakhutdinov, & Tenenbaum, 2015; Tang, Liu, Li, et al., 2020), this method has a
64   wide range of applications. However, more data is needed to avoid over-fitting (Rezende,
65   Danihelka, Gregor, & Wierstra, 2016). In 2017, Vicarious researchers came up with a probabilistic
66   model of vision, based on how the visual cortex works, reporting up from level to level, and named
67   it the recursive Cortical network. With the same accuracy, the recursive cortical network uses only
68   one-millionth of the deep learning method's training sample. It can be used to crack various
69   variants of text captchAs after one training (George et al., 2017; Li et al., 2020).
70   Meta-learning, also known as learning to learn, refers to using previous experience to learn a
71   new task quickly, rather than thinking about the new task in isolation. In 2016, Brenden M. Lake
72   et al. emphasized its importance as the cornerstone of Artifical intelligence (B. M. Lake, Ullman,

73   Tenenbaum, & Gershman, 2017). Up to now, the research directions of using meta-learning to deal
74   with small sample learning include: 1. Memory enhancement; 2. Measure learning; 3. Learn the
75   optimizer; 4. Optimize initial characterization.
76   1)Memory enhancement
77   Memory enhancement refers primarily to the use of cyclic neural networks with memory or
78   temporal convolution to iterate over examples of a given problem, accumulating information that
79   is used to solve the problem in its hidden activation layer or external memory. Considering that
80   neural network Turing machines can carry out short-term memory through external storage and
81   carry out long-term memory through slow weight updating (Graves et al., 2016; Tang et al., 2021),
82   Adam Santoro et al proposed a memory enhancement network based on Long and short Memory
83   Network (LSTM) in 2016 (Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap, 2016; Tang, Liu,
84   Deng, et al., 2020). Next,Munkhdalai et al. proposed a meta-network (Munkhdalai & Yu, 2017),
85   which is composed of a meta-learning device and a learning device, and at the same time, a
86   memory unit is added externally. By using the gradient as the meta-information, the fast weight
87   and slow weight are generated on two time scales, and then the fast weight and slow weight are
88   combined by the layer enhancement method.
89   2)Measure Learning
90   Metric Learning refers to learning a similarity measure from data, and then using this measure
91   to compare and match samples of new unknown categories. In 2015, Gregory Koch et al. proposed
92   the twin network (Chen et al., 2020; Koch, Zemel, & Salakhutdinov, 2015). The recognition
93   accuracy of the model on Onniglot data is close to human. In 2016, Oriol Vinyals et al. proposed
94   an end-to-end and directly optimized matching network based on memory and attention (Li et al.,
95   2017; Vinyals et al., 2016). This network can quickly learn with few samples, and for new classes
96   never seen in the training process, the trained model is not changed, and the test samples can be
97   classified with a few calibration samples of each class. Flood Sung et al. proposed the relationship
98   network in 2018 (Sung et al., 2018). Learning to get similarity measure is more flexible and can
99   capture the similarity between features better than the artificially selected measure method, and
100   good results are obtained on several benchmark data sets learned with few samples.
101   3)Learn the optimizer
102   The learning optimizer refers to learning how to update the learner parameters, that is,
103   learning the updating function or updating rules of the new model parameters. In 2017, Ravi and
104   Larochelle used LSTM as a meta-learner, and took the learner's initial recognition parameters,
105   learning rate and loss gradient as LSTM states to learn the learner's initialization parameters and
106   parameter update rules (Ravi & Larochelle, 2016). Yang et al. introduced metric Learning based
107   on the method proposed by Ravi and Larochelle, and proposed a meta-metric learner (Yang, Liu,
108   Dong, & Wu, 2020). The author integrates the matching network and the method of updating rules
109   using LSTM learning parameters, and obtains a better method. However, such a structure is more
110   complex than the structure based on measurement learning alone, and each parameter of the learner
111   is updated independently in each step, which will largely limit its potential (Sung et al., 2018).
112   4)Optimize initial characterization
113   Optimize the initial representation, that is, optimize the initial representation directly. In 2017,

114 Chelsea Finn et al. proposed a model-agnostic meta-learning algorithm (MAML) (Finn, Abbeel,
115 & Levine, 2017; Zheng et al., 2017). Compared with the previous meta-learning method, this
116 method introduces no additional parameters and has no restrictions on the model structure. Instead
117 of updating a function or learning rule, it only uses gradient to update the learner weight. Andrei
118 A. Liu et al. used the entire training set to pre-train the feature extractor in the case that MAML
119 could not handle high-dimensional data well, and then used the parameter generation model to
120 capture various parameters useful for task distribution (Liu et al., 2020; Zheng, Li, Yin, & Wang,
121 2016a). Totally different from the mamL-based improvement, Ruixiang Zhang et al. applied
122 antagonistic neural networks in the field of meta-learning. This method uses the idea of generating
123 antagonistic network to expand data and improve the learning ability of the network. However, the
124 overall effect of this method is not stable, the relative cost is high, and the technology is not mature
125 enough (Zhang, Che, Ghahramani, Bengio, & Song, 2018; Zheng, Li, Yin, & Wang, 2016b).
126   The ability to learn and adapt quickly from small amounts of data is critical to AI. However,
127 the success of deep learning depends to a large extent on a large amount of tag data, and in deep
128 neural network learning, each task is isolated from the Learning, and Learning is always from
129 scratch when facing a new task. Limited data or rapid generalization in a dynamic environment
130 challenge current deep learning methods. Therefore, this paper takes the problem of image
131 classification with few samples as the research object, combines measurement learning and
132 optimization of initial representation methods, and designs a multi-scale meta-relational network.
133 Finally, the classification accuracy and training speed of the two baseline data sets with small
134 sample learning are improved.

Materials & Methods

136 **Materials:**

137 **1. Omniglot data set**

138   In 2011, the Omniglot data set was collected by Brenden Lake and his collaborators at MIT
139 through Amazon's Mechanical Turk (B. Lake, Salakhutdinov, Gross, & Tenenbaum, 2011; Zheng,
140 Li, Xie, Yin, & Wang, 2015). It consists of 50 international languages, including mature
141 international languages such as Latin and Korean, little-known local dialects, and fictional
142 character sets such as Aurek-Besh and Klingon. The number of letters in each language varies
143 widely, from about 15 to 40 letters, with 20 samples for each letter. Therefore, the Omniglot data
144 set consists of 1623 categories and 32,460 images. Figure 1 illustrates the five languages of the
145 Omniglot data set.

146

147   Figure 1 Omniglot data set example

148 **2. MiniImageNet data set**

149   MiniImageNet data set, proposed by Vinyals et al., consists of 60,000 84×84×3 color images,
150 a total of 100 categories, each with 600 samples (Vinyals et al., 2016). The distribution of data sets
151 is very different, and the image category involves animals, household goods, remote sensing
152 images, food, etc. Vinyals did not publish this data set. Ravi and Larochelle randomly selected 100
153 classes from the ImageNet data set to create a new MiniImageNet data set, which was divided into

154   training set, validation set and test set at a ratio of 64:16:20 (Ravi & Larochelle, 2016).

155   **Method**

156   **1.   Metalearning based on metric Learning**

157   In meta-learning, metric Learning refers to learning similarity measurement from a wide

158   range of task Spaces, so that the experience extracted from the previous learning tasks can be used

159   to guide the Learning of new tasks to achieve the purpose of learning how to learn. The learner

160   (meta-learner) learns the target set on each task of the training set by measuring the distance of the

161   support set, and finally learns a metric. Then for the new task of the test set, it can quickly classify

162   the target set correctly with the help of a small number of samples of the support set.

163   At present, the methods of small sample image classification based on metric Learning

164   include: twin network, matching network, prototype network and relational network. The twin

165   network is composed of two identical convolutional neural networks, and the similarity between

166   two images is calculated by comparing the loss function with the input of paired samples. The

167   other three methods, instead of using paired sample inputs, calculate the similarity between the

168   two images by setting the support set and the target set. In this paper, the improvement is made on

169   the basis of multi-scale relational network, so the main structure of this method is mainly

170   introduced below.

171   The structure of relational network (Sung et al., 2018) first obtains feature graphs of support

172   set and target set samples through embedded modules, then splices feature graphs of support set

173   and target set samples in depth direction, and finally obtains relationship score by learning splicer

174   features through relationship module, so as to determine whether support set and target set samples

175   belong to the same category.

176   The calculation formula of relational fractions is as follows:

177
$$r_{i,j} = g_\phi(C(f_\varphi(x_i), f_\varphi(x_j))), i = 1, 2, ..., C \tag{1}$$

178   Where, x_i represents the supporting set sample, x_j represents the target set sample, $f_\varphi(x)$

179   represents the embedded module, $f_\varphi(x_i)$ and $f_\varphi(x_j)$ represent the feature graph of the supporting

180   set and target set sample, and $C\left(f_\varphi(x_i), f_\varphi(x_j)\right)$ represents the feature graph splicing operator.

181   g_ϕ stands for relationship module.

182   **2.   Model-independent meta-learning algorithm**

183   According to the idea of transfer learning, when adapting to a new task, MAML only needs

184   to fine-tune the learner so that the parameter of the learner is adapted from θ to θ_i'. However,

185   different from migration learning, MAML only needs to perform one or more gradient descent

186   iterative steps on a small amount of data in new task T_i to enable the learner to converge to the

187   optimal parameter.

188   Here, θ_i' only considers the MAML that converges to the optimal parameter after an iterative

189   step of gradient descent as follows:

190                             $$\theta_i' = \theta - \alpha \nabla_\theta L_{T_i} f(\theta) \tag{2}$$

191   Where, $f(\theta)$ stands for learner, L_{T_i} stands for loss on specific task, $\nabla_\theta L_{T_i} f(\theta)$ stands for loss

192   gradient, and α stands for learner gradient update pace, namely learning rate of learner.

193   On a small amount of data in new task T_i, the learner can be converged to the optimal

194   parameter θ_i' after one step of gradient descent iteration, so MAML must find a set of initialization

195   representations that can be effectively fine-tuned according to a small number of samples. In order
196   to achieve this goal, based on the idea of meta-learning "using previous experience to quickly learn
197   new tasks", learners need to learn the learner parameter θ on different tasks. MAML defines this
198   process as meta-learning process.

199   In the Meta-learning process, for different training tasks, the optimal parameter θ_i' suitable

200   for the specific task was obtained through a step of gradient iteration, and then sampled again on

201   each task for testing, requiring $L_{T_i} f(\theta_i')$ to reach the minimum value. Therefore, MAML adopted

202   the sum of test errors on different tasks as the optimization objective of the meta-learning process,
203   as shown below:

204              $$min_\theta \sum_{T_i \sim p(T)} L_{T_i} f(\theta_i') = \sum_{T_i \sim p(T)} L_{T_i} f(\theta - \alpha \nabla_\theta L_{T_i} f(\theta)) \tag{3}$$

205   Where, $p(T)$ represents the distribution of the task set, L_{T_i} represents the loss on the specific task,

206    $\nabla_\theta L_{T_i} f(\theta)$ represents the loss gradient, and α represents the learning rate of the learner.

207   It can be seen from Equation (3-3) that the optimization goal of the meta-learning process is

208   to adopt the updated model parameter θ_i' adapted to the specific task, while the meta-optimization

209   process is ultimately executed on the learner parameters θ. The stochastic gradient descent method
210   is adopted, and the updated iteration formula of model parameter θ is shown as follows:

211                        $$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i \sim p(T)} L_{T_i} f(\theta_i') \tag{4}$$

212   Where, $p(T)$ represents the distribution of the task set, L_{T_i} represents the loss on the specific task,

213    $\nabla_\theta L_{T_i} f(\theta)$ represents the gradient, and β represents the gradient update pace of the meta-learner,

214   also known as the learning rate of the meta-learner.

215   Substituting Equation (3-1) into Equation (3-4), we can get:

216
$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i \sim p(T)} L_{T_i} f\left(\theta - \alpha \nabla_\theta L_{T_i} f(\theta)\right) \tag{5}$$

3. Algorithm design of multi-scale meta-relational network

218   In the multi-scale meta-relational network, we hope to find a set of characterization θ that
219   can make fine adjustments efficiently according to a small number of samples. Where, θ is

220   composed of feature extractor parameter φ and metric learner parameter ϕ. During the training

221   process, each task is composed of training set D_{train} and test set D_{test}. Where, D_{train} is used for the

222   inner optimization iteration, D_{test} serves as the target set of the outer optimization iteration, and the

223   support set adopts the support set D_S in D_{train}.

224   In the inner optimization iteration process, on the new task T_i with a small amount of data D_{train},

225   the learner converges to the optimal parameter θ_i after an iterative step of gradient descent, as

226   shown below:

227
$$\theta_i = \theta - \alpha \nabla_\theta L_{T_i,D_{train}} f(\theta) \tag{6}$$

228   Where α is a vector of the same size as $L'_{T_i,D_{train}}(\theta)$, representing the learning rate of the learner.

229   The direction of the vector represents the update direction, and the magnitude of the vector
230   represents the learning step.

231   For different training tasks, the optimal parameter θ_i suitable for specific tasks was obtained

232   through a step of gradient iteration, and then tested on D_{test} of different tasks, requiring

233    $L_{T_i,D_{test},D_S} f(\theta_i)$ to reach the minimum value. Therefore, the meta-optimization objectives of the

234   meta-training process are as follows:

235
$$min_\theta \sum_{T_i \sim p(T)} L_{T_i,D_S,D_{test}} f(\theta_i) \tag{7}$$

236   Substituting Equation (3-6) into Equation (3-7), we can find:

237
$$min_\theta \sum_{T_i:\ p(T)} L_{T_i,D_S,D_{test}} f\left(\theta - \alpha \nabla_\theta L_{T_i,D_{train}} f(\theta)\right) \tag{8}$$

238   For parameter θ, the gradient calculation formula of formula (3-8) is as follows:

239
$$g_\theta = \frac{\partial}{\partial \theta} L_{T_i,D_S,D_{test}} f\left(\theta - \alpha \nabla_\theta L_{T_i,D_{train}} f(\theta)\right)$$
(9)

240   Substituting Equation (3-6) into Equation (3-9) can be simplified as follows:

241
$$g_\theta = \frac{\partial L_{T_i,D_S,D_{test}} f(\theta_i)}{\partial f(\theta_i)} \times \frac{\partial f(\theta_i)}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \theta}$$

$$= \frac{\partial L_{T_i,D_S,D_{test}} f(\theta_i)}{\partial f(\theta_i)} \times \frac{\partial f(\theta_i)}{\partial \theta_i} \times (I - \alpha \frac{\partial \left(\nabla_\theta L_{T_i,D_{train}} f(\theta)\right)}{\partial \theta})$$
(10)

242   Where I is the unit vector.

243   It can be seen from Equation (3-10) that the update process of the multi-scale meta-relational
244   network involves the gradient calculation process, that is, when the gradient operator in the meta-
245   target is used to propagate the meta-gradient, additional reverse transfer is required to calculate the
246   Hessian vector product (Kemp & Maas, 2009).

247   In the multi-scale relational network, except for the sigMIod nonlinear function at the last full
248   connection layer of the metric learner, all other nonlinear functions are ReLU. The ReLU neural
249   network is almost linear in part, which indicates that the second derivative is close to zero in most
250   cases. Therefore, the multi-scale element relational network, like MAML, ignores the second
251   derivative in the calculation of the back propagation of the element gradient (Kemp & Maas, 2009),
252   as follows:

253
$$g_\theta = \frac{\partial L_{T_i,D_S,D_{test}} f(\theta_i)}{\partial f(\theta_i)} \times \frac{\partial f(\theta_i)}{\partial \theta_i}$$

$$= \frac{\partial L_{T_i,D_S,D_{test}} f(\theta_i)}{\partial \theta_i}$$

$$= \nabla_{\theta_i} L_{T_i,D_S,D_{test}} f(\theta_i)$$
(11)

254   Therefore, in the process of calculating the outer gradient, the multi-scale element relation

255   network stops the back propagation after calculating the gradient at θ_i and calculates the second

256   derivative at θ.

257   For parameter α, the gradient calculation formula of formula (3-8) is as follows:

258
$$g_\alpha = \frac{\partial}{\partial \alpha} L_{T_i,D_S,D_{test}} f\left(\theta - \alpha \nabla_\theta L_{T_i,D_{train}} f(\theta)\right)$$

$$= \frac{\partial L_{T_i,D_S,D_{test}} f(\theta_i)}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \alpha}$$

$$= -\nabla_\theta L_{T_i,D_{train}} f(\theta) \times \nabla_{\theta_i} L_{T_i,D_S,D_{test}} f(\theta_i)$$
(12)

259   In the outer optimization iteration process, stochastic gradient descent method is adopted, and
260   the updated iteration formula of model parameter θ is shown as follows:

$$\theta \leftarrow \theta - \beta \sum_{T_i \sim p(T)} \nabla_{\theta_i} L_{T_i, D_S, D_{test}} f(\theta_i) \tag{13}$$

Similarly, the update iteration formula of model parameter α is shown as follows:

$$\alpha \leftarrow \alpha + \beta \sum_{T_i \sim p(T)} \left(\nabla_{\theta_i} L_{T_i, D_S, D_{test}} f(\theta_i) \right) \left(\nabla_{\theta} L_{T_i, D_{train}} f(\theta) \right) \tag{14}$$

In each inner iteration, all tasks can be used to update the gradient, but when the number of tasks is very large, all tasks need to be calculated in each iteration step, so the training process will be slow and the memory capacity of the computer will be high. If each inner iteration uses one task to update the parameters, the training speed will be accelerated, but the experimental performance will be reduced, and it is not easy to implement in parallel. So, during each inner iteration, we take num_inner_task for the number of tasks. By selecting the number of tasks in the inner iteration process as a super parameter, memory can be reasonably utilized to improve the training speed and accuracy.

During the meta-training of multi-scale meta-relational network, after the end of an epoch, the accuracy on the meta-validation set was calculated, and the highest accuracy on the meta-validation set was recorded so far. If consecutively N epoches (or more, adjusted according to the specific experimental conditions) did not reach the optimal value, it could be considered that the accuracy was no longer improved, and the iteration could be stopped when the accuracy was no longer improved or declined gradually. Output the model with the highest accuracy, and then test it with this model.

Experiment and results

1. Multi-scale meta-relational network design

In the algorithm design of multi-scale meta-relational network:

1. Adopting the meta-SGD idea, the learning rate of the inner layer, that is, the learning rate of the learner, is taken as the learning vector and model parameter to learn together, so as to further improve the performance of the learner. 2. Considering that the small sample image classification method based on metric Learning can be adapted to the new task without fine-tuning, the multi-scale meta-relational network adopts MAML algorithm to learn and find the optimal parameters of the model in the meta-training process, and eliminates the inner gradient iteration in the meta-validation and meta-testing process.

2. Small sample learning benchmark dataset experiment

The model structure of multi-scale meta-relational network adopts multi-scale relational network, which is mainly composed of feature extractor and measure learner (Biggs, 1985). The second derivative is ignored in the back propagation process of the multi-scale element relational network. In practical implementation, the back propagation process can be truncated after the calculation of the outer gradient of the multi-scale element relational network is completed.

For experimental comparison, MAML and MetaSGD experiments are carried out simultaneously in this paper. According to Chelsea Finn et al., in the experiment of MAML and MetaSGD, a convolutional neural network composed of four layers of convolution and one layer

298   of full connection is adopted as a learning device. Each convolution module of the learner is
299   composed of a convolutional layer consisting of 64 filters of size 1, a batch standardization layer,
300   a modified linear element layer and a maximum pooling layer of size 2. All the convolutional
301   layers are filled with zero. The fourth convolution module is followed by the full connection layer.
302   The number of neurons in the full connection layer is 64, and the nonlinear function log_SOFTmax
303   is adopted. Different from the multi-scale relational network, the cross-entropy loss function is
304   used to optimize the learning object.

305   *a) Omniglot data set*

306   In the Omniglot data set experiment, all samples were processed into a size of 28×28 and a
307   random rotation was used to enhance the data. From 1623 classes, 1200, 211 and 212 classes were
308   selected as meta-training set, meta-validation set and meta-test set respectively. In this section,
309   single-sample image classification experiments of 5-ways 1-shot and 20-ways 1-shot and small-
310   sample image classification experiments of 5-ways 5-shot and 20-ways 5-shot are conducted.

311   In the experiment, 100 episodes of multi-scale meta-relational network were taken as one
312   epoch. The iteration times of the meta-training set, meta-validation set and meta-test set are
313   accordingly 70,000, 500 and 500. Other super-parameter selections on the Omniglot dataset are
314   shown in Table 1.

315   Table 1 Omniglot data set experiment super parameter value

316

317   1)Single sample image classification

318

319   Figure 2 Accuracy and loss iteration curves of 5-way 1-shot in multi-scale meta-relational network

320   It can be seen from Figure 2 that, when the number of iterations is 42400, the multi-scale
321   meta-relational network achieves the highest accuracy rate of 99.8667% in the 5-way 1-shot
322   experiment on the meta-validation set. The iteration time of the multi-scale meta-relational
323   network is less than that of the multi-scale relational network (method 1 in this paper). Therefore,
324   compared with the convergence of the multi-scale relational network when the iteration reaches
325   114,000, the learning speed of the multi-scale meta-relational network is faster than that of the
326   multi-scale relational network.

327   The trained model was tested on the meta-test set, and the test results were shown in Table 2.
328   According to Table 2:

329   1. The accuracy rate of the 5-way 1-shot experiment on the meta-test set of the multi-scale
330   meta-relational network is higher than that of MAML and Meta-SGD, and about 0.22% higher
331   than that of the multi-scale meta-relational network.

332   2. MAML and Meta-SGD based on optimized initial characterization need fine-tuning on
333   new tasks, while the multi-scale relational network based on metric Learning can achieve good
334   generalization performance on new tasks without fine-tuning. By comparing the two methods, the
335   accuracy rate of the 5-way 1-shot experiment on the meta-test set of the multi-scale relational
336   network is higher than MAML, but slightly lower than meta-SGD.

337

338   Figure 3 Accuracy and loss iteration curves of 20-way 1-shot in multi-scale meta-relational network

339   As can be seen from FIG. 3, when the number of iterations is 57,700, the multi-scale meta-
340   relational network (method 2 in this paper) achieves the highest accuracy rate of 98.86% in the 20-
341   way 1-shot experiment on the meta-validation set. Compared with the multi-scale relational
342   network (method 1 in this paper), the learning speed of multi-scale meta-relational network is
343   faster than that of multi-scale relational network when it iterates to 89,000.

344   The trained model was tested on the meta-test set, and the results were shown in Table 2
345   below. According to Table 2:

346   1. The accuracy rate of the 20-way 1-shot experiment of the multi-scale meta-relational
347   network on the meta-test set is about 0.47% higher than that of the multi-scale meta-relational
348   network, which is higher than MAML and Meta-SGD.

349   2. MAML and Meta-SGD based on optimal initialization representation were compared with
350   the multi-scale relational network based on metric Learning. The accuracy of the multi-scale
351   relational network in the 20-way 1-shot experiment on the meta-test set was higher than that of
352   MAML and meta-SGD.

353   Table 2 Omniglot data set single sample classification experimental results

354

355   2)Small sample image classification

356

357   Figure 4 Accuracy and loss iteration curves of 5-way 5-shot in multi-scale meta-relational network

358   As can be seen from FIG. 4, when the number of iterations is 29100, the multi-scale meta-
359   relational network (method 2 in this paper) achieves the highest accuracy of 99.89% in the 5-way
360   5-shot experiment on the meta-validation set. Compared with the multi-scale relational network
361   (method 1 in this paper), the learning speed of the multi-scale meta-relational network is faster
362   than that of the multi-scale relational network when it iterates to 293500.

363   The trained model is tested on the meta-test set, and the results are shown in Table 3.
364   According to Table 3:

365   1. The accuracy rate of the 5-way 5-shot experiment on the meta-test set of the multi-scale
366   meta-relational network is about 0.14% higher than that of the multi-scale meta-relational network,
367   but lower than MAML and Meta-SGD.

368   2. MAML and Meta-SGD based on optimal initial representation were compared with the
369   multi-scale relational network based on metric Learning. The accuracy of the 5-way 5-shot
370   experiment on the meta-test set of the multi-scale relational network was lower than that of MAML
371   and meta-SGD.

372

373   Figure 5 Accuracy and loss iteration curves of 20-way 5-shot in multi-scale meta-relational network

374   As can be seen from FIG. 5, when the number of iterations is 65800, the multi-scale meta-
375   relational network (method 2 in this paper) achieves the highest accuracy rate of 99.82% in the
376   20-way 5-shot experiment on the meta-validation set. Compared with the multi-scale relational
377   network (method 1 in this paper), the learning speed of multi-scale meta-relational network is
378   faster than that of multi-scale relational network when it iterates to 119500.

379   The trained model is tested on the meta-test set, and the results are shown in Table 3.

380 According to Table 3:

381   1. The accuracy rate of the 20-way 5-shot experiment of the multi-scale meta-relational

382 network on the meta-test set is about 0.19% higher than that of the multi-scale meta-relational

383 network, which is also higher than MAML and Meta-SGD.

384   2. MAML and Meta-SGD based on optimized initial characterization were compared with the

385 multi-scale relational network based on metric Learning. The accuracy of the multi-scale relational

386 network in the 20-way 5-shot experiment on the meta-test set was higher than that of MAML and

387 meta-SGD.

388   Table 3 Small sample classification experimental results in Omniglot data set

389

390   In summary, the experimental results on Omniglot data set are as follows:

391   1. The classification accuracy of the 5-way 5-shot experiment on Omniglot data set was

392 slightly lower than that of meta-SGD and MAML, and the other experiments were all higher than

393 that of the three methods of multi-scale relational network, MAML and Meta-SGD. And the

394 training speed of multi-scale meta-relational network is faster than that of multi-scale relational

395 network.

396   2. MAML and Meta-SGD based on optimal initial characterization were compared with the

397 multi-scale relational network based on metric Learning. Except that the classification accuracy of

398 5-way 1-shot experiment on Omniglot data set was slightly lower than that of META-SGD and

399 MAML, the classification accuracy of 5-way 5-shot experiment was lower than that of META-

400 SGD and MAML, the other two groups of experiments were both higher than that of META-SGD

401 and MAML.

402   ***b) Miniimagenet data set***

403   According to the ratio of 64:16:20, it is divided into meta-training set, meta-validation set and

404 meta-test set. In this chapter, the classification experiment of single-sample image of 5-ways 1-

405 shot and the classification experiment of small-sample image of 5-ways 5-shot is carried out.

406   In the experiment, the multi-scale meta-relational network took 500 episodes as one epoch.

407 The number of iterations of the meta-training set, meta-validation set and meta-test set can be

408 valued at 120,000, 600 and 600. Other super-parameter selections on the MiniImageNet dataset

409 are shown in Table 4.

410   Table 4 miniImageNet data set experiment super parameter value

411

412   1) Single sample image classification

413

414   Figure 6 Accuracy and loss iteration curves of 5-way 1-shot in multi-scale meta-relational network

415   As can be seen from Figure 6, with the increase of iteration, loss decreases to convergence,

416 and accuracy gradually increases to convergence. When the number of iterations is 87,000, the

417 multi-scale meta-relational network (method 2 in this paper) achieves the highest accuracy rate of

418 51.234% in the 5-way 1-shot experiment on the meta-validation set. The iteration time of multi-

419 scale meta-relational network is less than that of multi-scale relational network (method 1 in this

420 paper). Therefore, compared with the convergence of multi-scale relational network when iteration

421 reaches 155000, the learning speed of multi-scale meta-relational network is faster than that of
422 multi-scale relational network.
423   The trained model is tested on the meta-test set, and the results are shown in Table 5.
424 According to Table 5:
425   1. The accuracy rate of the 5-way 1-shot experiment on the meta-test set of the multi-scale
426 meta-relational network is about 0.35% higher than that of the multi-scale meta-relational network,
427 and higher than MAML and Meta-SGD.
428   2. MAML and Meta-SGD based on optimized initial characterization were compared with
429 the multi-scale relational network based on metric Learning. The accuracy of the 5-way 1-shot
430 experiment on the meta-test set of the multi-scale relational network was higher than MAML, but
431 lower than meta-SGD.
432   Table 5 The experimental results of 5-way 1-shot in Miniimagenetdata set
433
434   2)Small sample image classification
435
436   Figure 7 Accuracy and loss iteration curves of 5-way 5-shot in multi-scale meta-relational network
437   As can be seen from Figure 7, with the increase of iteration, loss decreases to convergence,
438 and accuracy gradually increases to convergence. When the number of iterations is 62000, the
439 multi-scale meta-relational network (method 2 in this paper) achieves the highest accuracy rate of
440 66.81% in the 5-way 5-shot experiment on the meta-validation set. Compared with the
441 convergence of multi-scale relational network (method 1 in this paper) when iterative to 140000,
442 the learning speed of multi-scale meta-relational network is faster than that of multi-scale relational
443 network.
444   The trained model is tested on the meta-test set, and the results are shown in Table 6.
445 According to Table 6:
446   1. The accuracy rate of the 5-way 5-shot experiment on the meta-test set of the multi-scale
447 meta-relational network is about 0.34% higher than that of the multi-scale meta-relational network,
448 which is higher than MAML and Meta-SGD.
449   2. MAML and Meta-SGD based on optimized initial characterization were compared with
450 the multi-scale relational network based on metric Learning. The accuracy of the multi-scale
451 relational network in the 5-way 5-shot experiment on the meta-test set was higher than that of
452 MAML and meta-SGD.
453   Table 6 The experimental results of 5-way 5-shot in Miniimagenet data set
454
455   In summary, the experimental results on Miniimagenet data set are as follows:
456   1. The classification accuracy of MNS on Miniimagenet data set is higher than that of MNS,
457 MAML and META-SGD, and the training speed is higher than that of MNS.
458   2. By comparing the MAML and Meta-SGD based on optimal initial representation and the
459 multi-scale relational network based on metric Learning, the classification accuracy of the 5-way
460 1-shot experiment on Miniimagenet data set was slightly lower than that of meta-SGD, and the
461 classification accuracy of the 5-way 5-shot experiment was higher than that of the META-SGD

462   and MAML.

Discussion

464   The multi-scale meta-relational network proposed in this paper learns specific classification
465   tasks by training the multi-scale relational network on a wide range of task Spaces. Then, MAML
466   algorithm is used to find a set of highly adaptive parameters, which can make good use of the
467   experiential knowledge learned in previous tasks and realize the ability of Learning.
468   1.By comparing the MAML and Meta-SGD based on optimized initial characterization with
469   the multi-scale relational network based on metric Learning, the overall experimental performance
470   of the two methods is almost the same in terms of accuracy, and it cannot be absolutely determined
471   that one method is better than the other.
472   2. MAML, Meta-SGD and multi-scale meta-relational networks based on optimal initial
473   characterization are compared separately, although Meta - SGD and MAML on 5 - way Omniglot
474   data set the classification accuracy of cyber-shot experiment was slightly higher than that of multi-
475   scale mata relation network, but Meta - SGD and MAML need to use a small amount of data from
476   a new task computing steps to update one or more of the gradient parameters, on a new task has
477   the largest generalization performance, and multi-scale meta network for not seen in the process
478   of training the new category of image, With the help of a small number of samples for each new
479   category, it has a good generalization ability. Therefore, the method combining metric learning
480   and learning optimization initial representation has higher performance than the method based on
481   optimization initial representation.
482   3. Compare the multi-scale relational network based on metric Learning with the multi-scale
483   meta-relational network separately. Multi-scale meta network will study MAML initialization
484   characterization method is introduced to measure the Learning, because the initialization
485   characterization is more suitable for MAML learning Yu Zaiyuan task distribution of training data
486   sets, thus reduced multi-scale mate network data sets the influence of the difference of distribution
487   in the small sample experiment results on benchmark data set higher than that of multi-scale
488   network, network learning speed and multi-scale relation in multi-scale network. Therefore, the
489   method combining metric learning and learning optimization initial representation has higher
490   performance than the method based on metric Learning.
491   Therefore, the method combining metric learning and optimal learning initial representation
492   has higher performance than the method based on metric or optimal initial representation. The
493   multi-scale meta-relational network enables the learned measurement method to have stronger
494   generalization ability, which not only improves the classification accuracy and training speed on
495   the benchmark set, but also avoids the situation that MAML needs fine-tuning.

Conclusion

497   Considering the difference of task set distribution, and in order to make the learned
498   measurement have stronger generalization ability, this paper designs a multi-scale meta-relational
499   network (MSNN), a classification method based on optimized initialization representation, for
500   image with few samples. First, the multi-scale meta-relational network adopts the meta-SGD idea

501 and takes the inner learning rate as the learning vector and model parameter to learn together.
502 Secondly, in the process of meta-training, the multi-scale meta-relational network adopts MAML
503 algorithm to learn and find the optimal parameters of the model, while in the process of meta-
504 validation and meta-test, the inner gradient iteration is eliminated and the test is carried out
505 directly. The experimental results show that the method combined with metric Learning and
506 optimized initial representation has higher performance than the method based on metric or
507 optimized initial representation. The multi-scale meta-relational network enables the learned
508 measurement method to have stronger generalization ability, which not only improves the
509 classification accuracy and training speed on the benchmark set, but also avoids the situation that
510 MAML needs fine-tuning.

511 Although the method in this paper improves the classification accuracy on the learning
512 benchmark set with few samples and the overfitting situation, it still needs to be improved in the
513 following aspects:

514 1) compared with miniImageNet data sets, Omniglot dataset is simpler, on the small sample
515 learning problems, Omniglot baseline is above 97% and the classification of the data set but on
516 miniImageNet data set classification effect is not too ideal, can find a better way of meta learning,
517 to reduce the influence of the difference of task set ascend miniImageNet data set classification
518 effect is worth exploring.

519 2) In the multi-scale element relational network, partial gradient information will be lost when
520 the second derivative is omitted. Whether a better way to simplify the second derivative can be
521 found is also the direction to be solved in future work.

Acknowledgements

523 The authors express their sincere appreciation and profound gratitude to research assistant
524 Xia Tian, WeiZheng Huang and LiHong Song, for their helping and supporting on collection and
525 sorting of the data.

Reference

527 Biggs, J. B. (1985). The role of metalearning in study processes. *British journal of educational psychology, 55*(3),
528 185-212.
529 Chen, X., Yin, L., Fan, Y., Song, L., Ji, T., Liu, Y., . . . Zheng, W. (2020). Temporal evolution characteristics of PM2.
530 5 concentration based on continuous wavelet transform. *Science of The Total Environment, 699*, 134244.
531 Ding, Y., Tian, X., Yin, L., Chen, X., Liu, S., Yang, B., & Zheng, W. (2019). *Multi-scale Relation Network for Few-*
532 *Shot Learning Based on Meta-learning.* Paper presented at the International Conference on Computer Vision
533 Systems.
534 Fe-Fei, L. (2003). *A Bayesian approach to unsupervised one-shot learning of object categories.* Paper presented at
535 the Proceedings Ninth IEEE International Conference on Computer Vision.
536 Finn, C., Abbeel, P., & Levine, S. (2017). *Model-agnostic meta-learning for fast adaptation of deep networks.* Paper
537 presented at the Proceedings of the 34th International Conference on Machine Learning-Volume 70.
538 George, D., Lehrach, W., Kansky, K., Lázaro-Gredilla, M., Laan, C., Marthi, B., . . . Wang, H. (2017). A generative
539 vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science, 358*(6368).
540 Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., . . . Agapiou, J. (2016).
541 Hybrid computing using a neural network with dynamic external memory. *nature, 538*(7626), 471-476.
542 Kemp, C., & Maas, A. L. (2009). *One-Shot Learning with Bayesian Networks.* Paper presented at the Proceedings of
543 the Annual Meeting of the Cognitive Science Society.

544 Koch, G., Zemel, R., & Salakhutdinov, R. (2015). *Siamese neural networks for one-shot image recognition.* Paper
545   presented at the ICML deep learning workshop.
546 Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). *One shot learning of simple visual concepts.* Paper
547   presented at the Proceedings of the annual meeting of the cognitive science society.
548 Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic
549   program induction. *Science, 350*(6266), 1332-1338.
550 Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like
551   people. *Behavioral and brain sciences, 40*.
552 Li, X., Yin, L., Yao, L., Yu, W., She, X., & Wei, W. (2020). Seismic spatiotemporal characteristics in the Alpide
553   Himalayan Seismic Belt.
554 Li, X., Zheng, W., Wang, D., Yin, L., & Wang, Y. (2015). Predicting seismicity trend in southwest of China based on
555   wavelet analysis. *International Journal of Wavelets, Multiresolution and Information Processing, 13*(02),
556   1550011.
557 Li, X., Zheng, W., Yin, L., Yin, Z., Song, L., & Tian, X. (2017). Influence of social-economic activities on air
558   pollutants in Beijing, China. *Open Geosciences, 9*(1), 314-321.
559 Liu, Q., Lu, J., Zhang, G., Shen, T., Zhang, Z., & Huang, H. (2020). Domain-specific meta-embedding with latent
560   semantic structures. *Information Sciences*.
561 Ma, Z., Zheng, W., Chen, X., & Yin, L. (2021). Joint embedding VQA model based on dynamic word vector. *PeerJ
562   Computer Science, 7*, e353.
563 Munkhdalai, T., & Yu, H. (2017). Meta networks. *Proceedings of machine learning research, 70*, 2554.
564 Ni, X., Yin, L., Chen, X., Liu, S., Yang, B., & Zheng, W. (2019). *Semantic representation for visual reasoning.* Paper
565   presented at the MATEC Web of Conferences.
566 Ravi, S., & Larochelle, H. (2016). Optimization as a model for few-shot learning.
567 Rezende, D., Danihelka, I., Gregor, K., & Wierstra, D. (2016). *One-Shot Generalization in Deep Generative Models.*
568   Paper presented at the International Conference on Machine Learning.
569 Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). *Meta-learning with memory-augmented
570   neural networks.* Paper presented at the International conference on machine learning.
571 Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). *Learning to compare: Relation
572   network for few-shot learning.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision
573   and Pattern Recognition.
574 Tang, Y., Liu, S., Deng, Y., Zhang, Y., Yin, L., & Zheng, W. (2020). Construction of force haptic reappearance system
575   based on Geomagic Touch haptic device. *Computer methods and programs in biomedicine, 190*, 105344.
576 Tang, Y., Liu, S., Deng, Y., Zhang, Y., Yin, L., & Zheng, W. (2021). An improved method for soft tissue modeling.
577   *Biomedical Signal Processing and Control, 65*, 102367.
578 Tang, Y., Liu, S., Li, X., Fan, Y., Deng, Y., Liu, Y., & Yin, L. (2020). Earthquakes spatio–temporal distribution and
579   fractal analysis in the Eurasian seismic belt. *Rendiconti Lincei. Scienze Fisiche e Naturali, 31*(1), 203-209.
580 Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. *Advances in
581   neural information processing systems, 29*, 3630-3638.
582 Yang, S., Liu, F., Dong, N., & Wu, J. (2020). Comparative Analysis on Classical Meta-Metric Models for Few-Shot
583   Learning. *IEEE Access, 8*, 127065-127073.
584 Yin, L., Li, X., Zheng, W., Yin, Z., Song, L., Ge, L., & Zeng, Q. (2019). Fractal dimension analysis for seismicity
585   spatial and temporal distribution in the circum-Pacific seismic belt. *Journal of Earth System Science, 128*(1), 22.
586 Zhang, R., Che, T., Ghahramani, Z., Bengio, Y., & Song, Y. (2018). Metagan: An adversarial approach to few-shot
587   learning. *Advances in neural information processing systems, 31*, 2365-2374.
588 Zheng, W., Li, X., Xie, J., Yin, L., & Wang, Y. (2015). Impact of human activities on haze in Beijing based on grey
589   relational analysis. *Rendiconti Lincei, 26*(2), 187-192.
590 Zheng, W., Li, X., Yin, L., & Wang, Y. (2016a). The retrieved urban LST in Beijing based on TM, HJ-1B and MODIS.
591   *Arabian Journal for Science and Engineering, 41*(6), 2325-2332.
592 Zheng, W., Li, X., Yin, L., & Wang, Y. (2016b). Spatiotemporal heterogeneity of urban air pollution in China based
593   on spatial analysis. *Rendiconti Lincei, 27*(2), 351-356.
594 Zheng, W., Li, X., Yin, L., Yin, Z., Yang, B., Liu, S., . . . Li, Y. (2017). Wavelet analysis of the temporal-spatial
595   distribution in the Eurasia seismic belt. *International Journal of Wavelets, Multiresolution and Information
596   Processing, 15*(03), 1750018.
597 Zheng, W., Liu, X., & Yin, L. (2021). Sentence Representation Method Based on Multi-Layer Semantic Network.

598   *Applied Sciences, 11*(3), 1316.

Figure 1

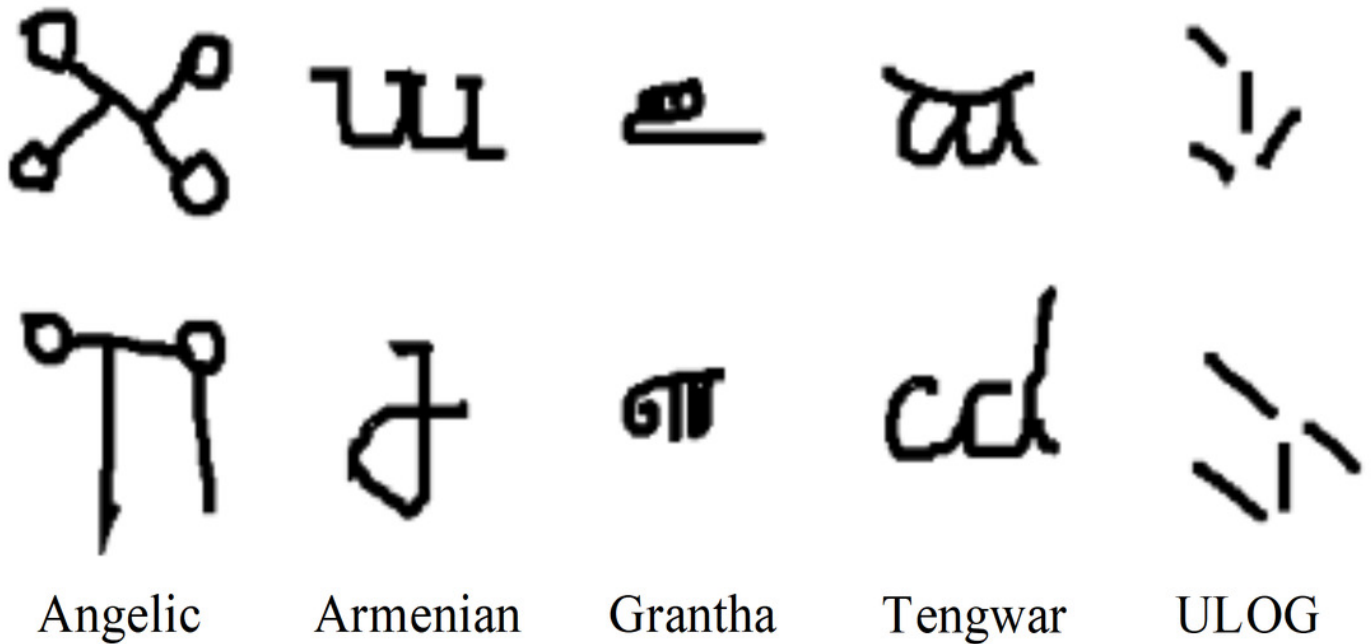
Omniglot dataset example

Angelic   Armenian   Grantha   Tengwar   ULOG

Figure 2

Accuracy and loss iteration curves of 5-way 1-shot in multi-scale meta-relational network

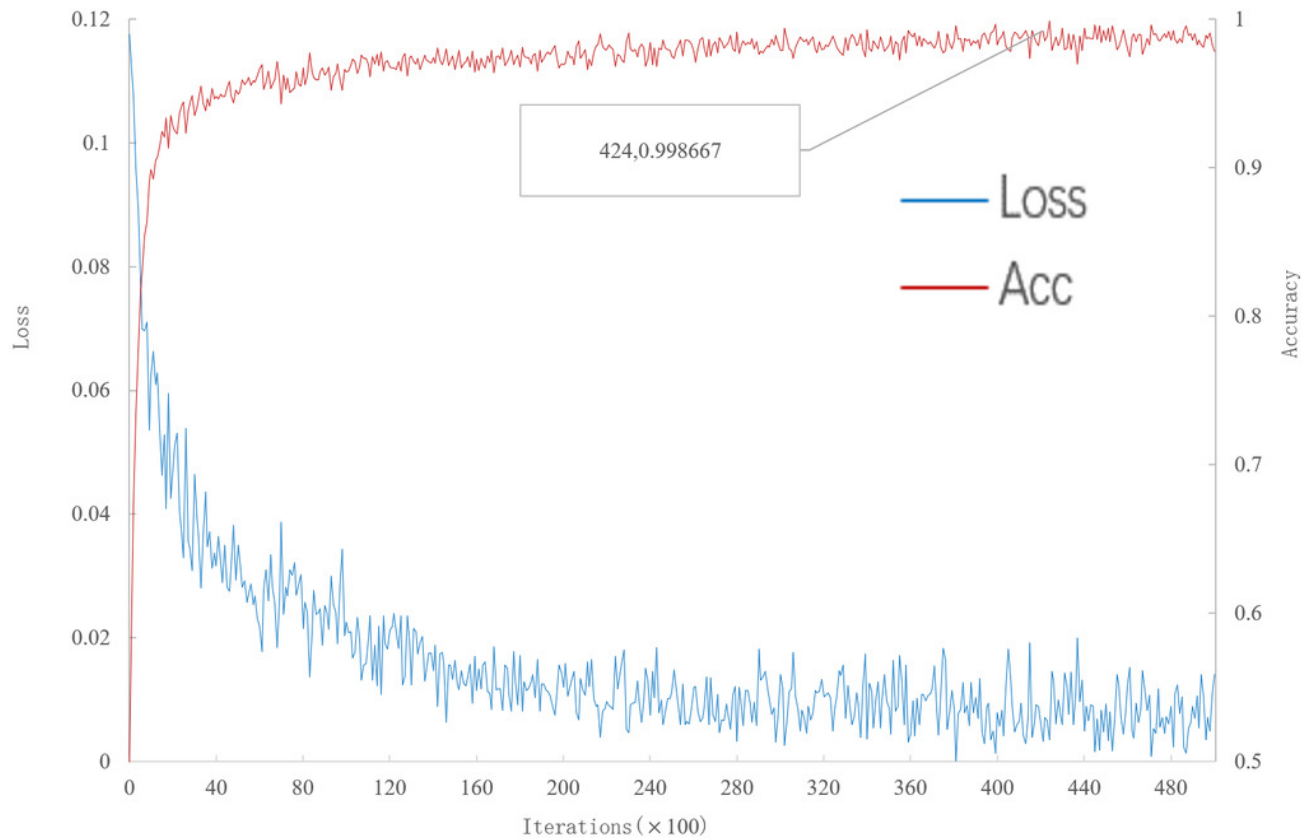



Figure 3

Accuracy and loss iteration curves of 20-way 1-shot in multi-scale meta-relational network

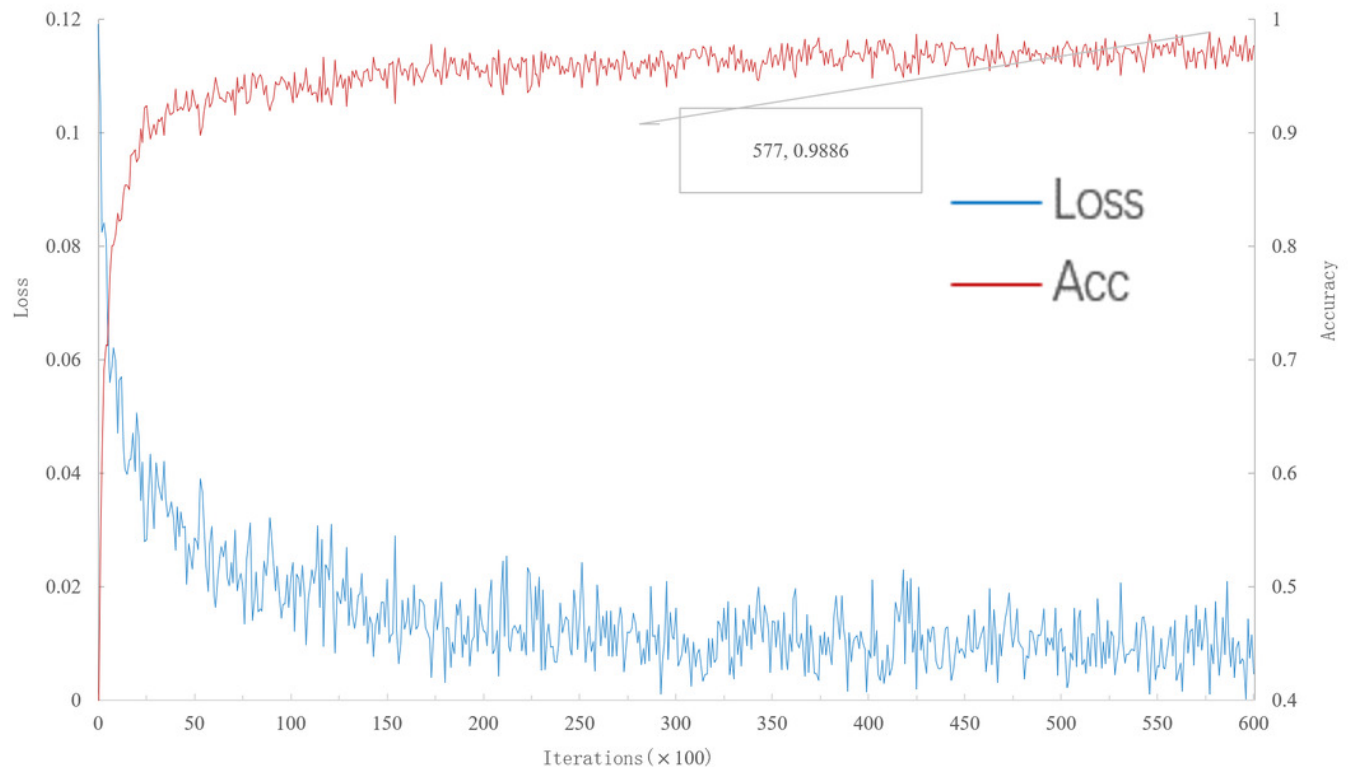



Figure 4

Accuracy and loss iteration curves of 5-way 5-shot in multi-scale meta-relational network

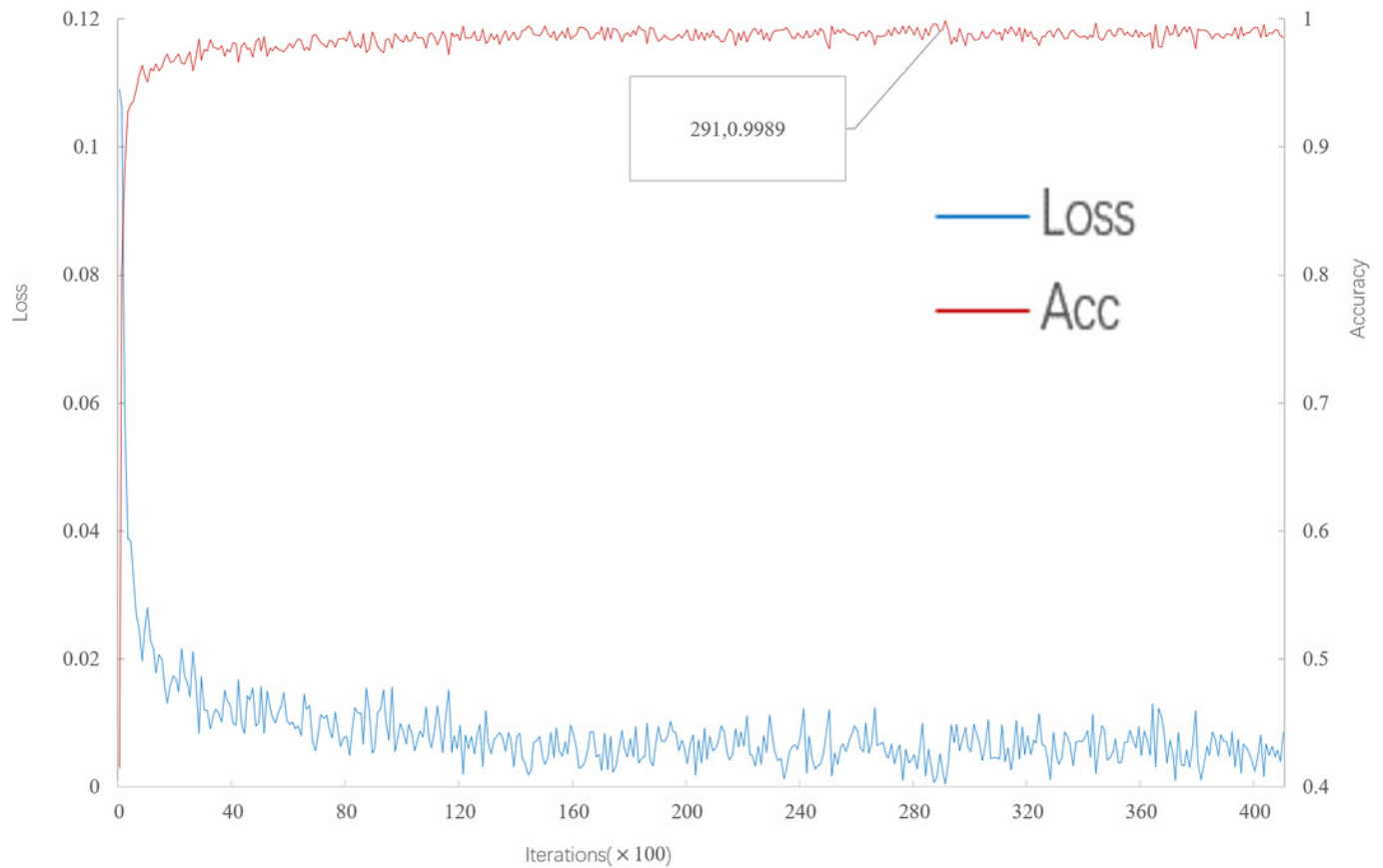



Figure 5

Accuracy and loss iteration curves of 20-way 5-shot in multi-scale meta-relational network

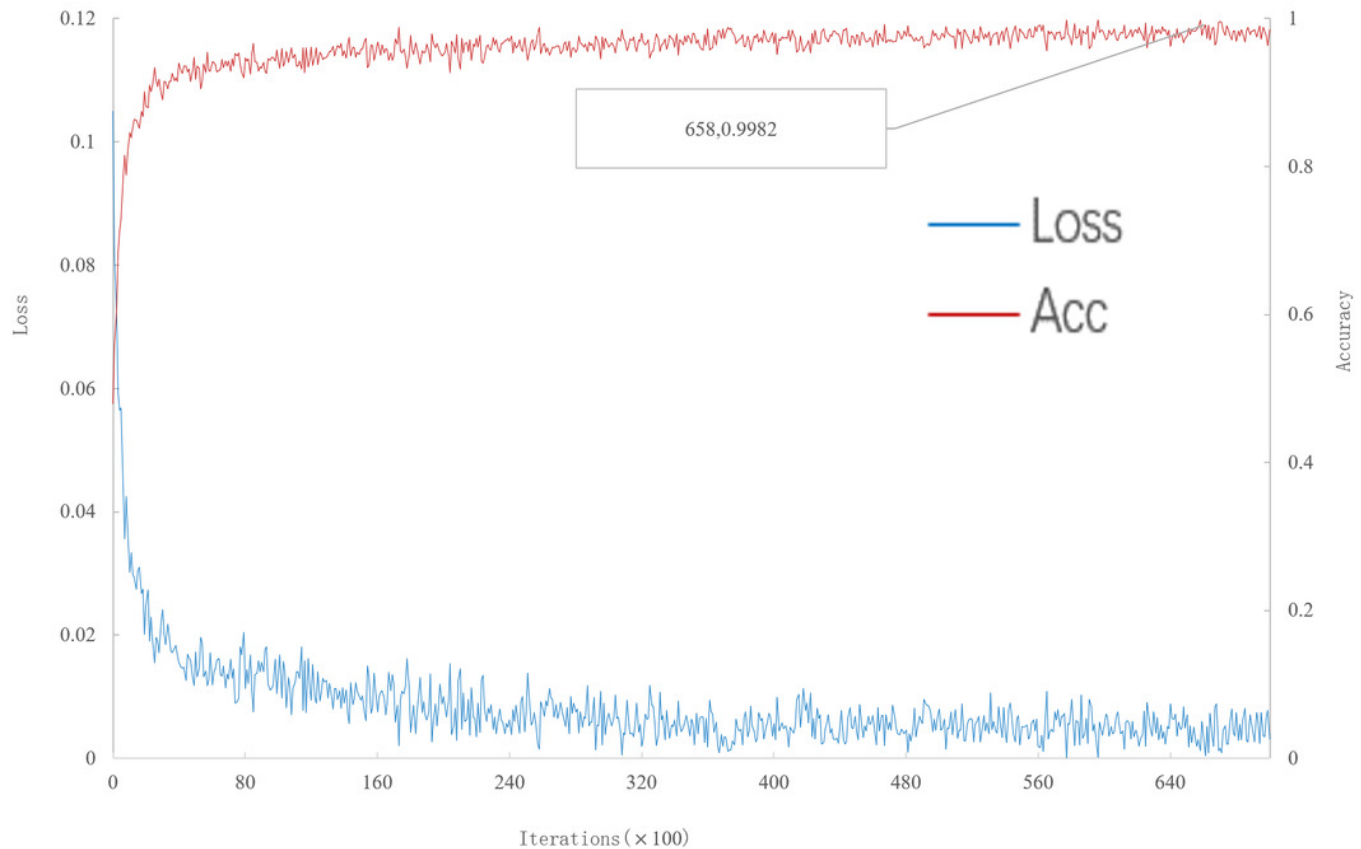



Figure 6

Accuracy and loss iteration curves of 5-way 1-shot in multi-scale meta-relational
network

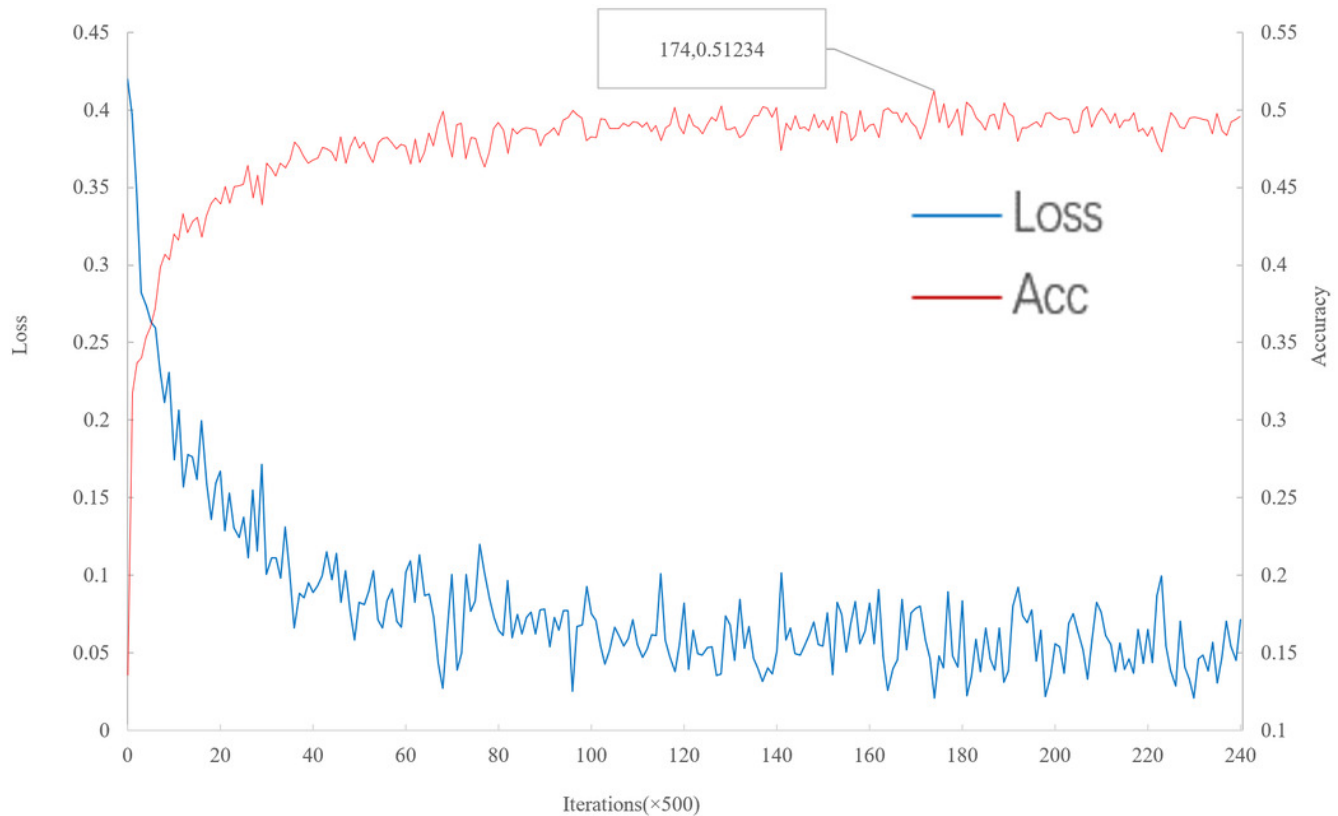
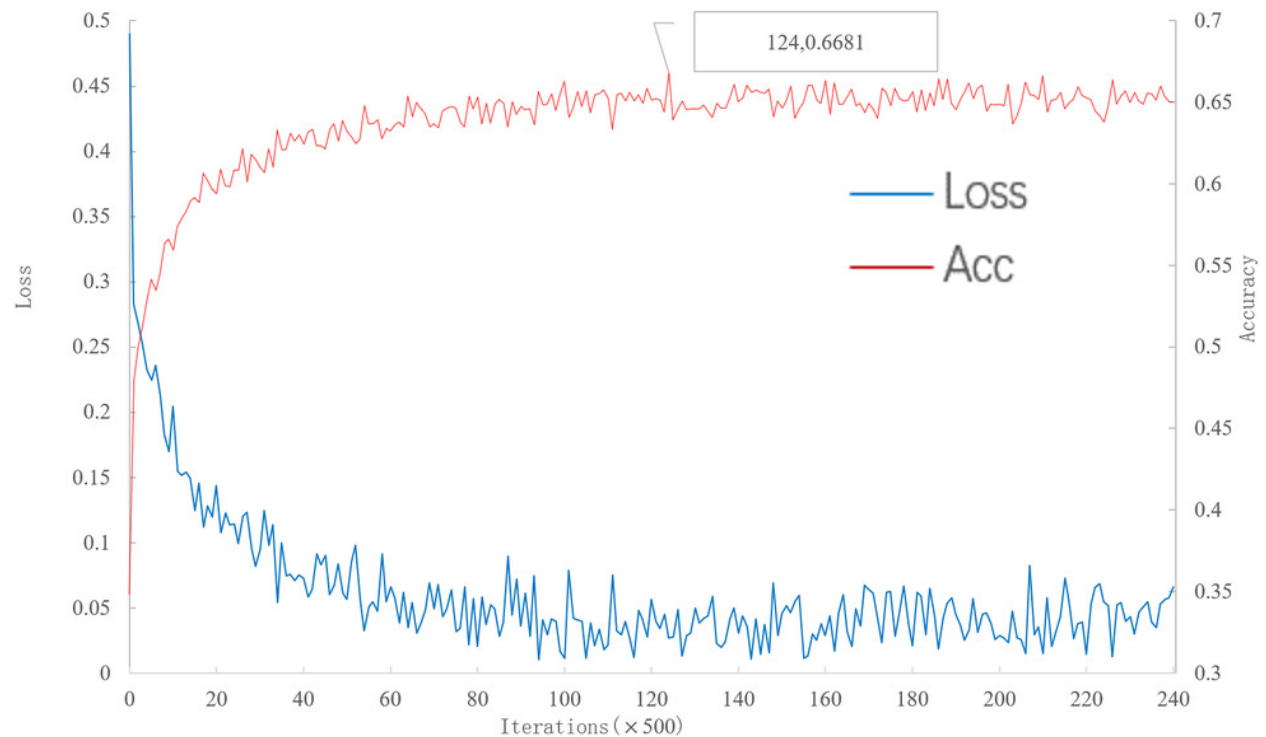



Figure 7

Accuracy and loss iteration curves of 5-way 5-shot in multi-scale meta-relational
network

Omniglot dataset experiment super parameter value

3 Table 1 Omniglot dataset experiment super parameter value

Meta-learning rate β	1-shot		5-shot		num_inner_task	
	b_1	b_2	b_1	b_2	5-way	20-way
0.003	8	8	5	5	32	16

Omniglot dataset single sample classification experimental results

2

Table 2 Omniglot dataset single sample classification experimental
results

Model	Fine-tuning	Accuracy	
		5-way 1-shot	20-way 1-shot
MAML	Y	97.80±0.32%	95.60±0.21%
Meta-SGD	Y	99.50±0.23%	95.83±0.36%
Multi-scale relational networks	N	99.35±0.25%	97.41±0.28%
Multi-scale meta-relational network	N	99.57±0.16%	97.88±0.20%

8
9
10
11

14
15

Small sample classification experimental results in Omniglot dataset

2
3

Table 3 Small sample classification experimental results in Omniglot dataset

Model	Fine-tuning	Accuracy	
		5-way 5-shot	20-way 5-shot
MAML	Y	99.89±0.10%	98.90±0.13%
Meta-SGD	Y	99.91±0.11%	98.92±0.22%
Multi-scale relational networks	N	99.70±0.08%	99.01±0.13%
Multi-scale meta-relational network	N	99.84±0.16%	99.20±0.12%

miniImageNet dataset experiment super parameter value

Table 4 miniImageNet dataset experiment super parameter value

Meta-learning rate β	1-shot		5-shot		num_inner_task	
	b_1	b_2	b_1	b_2	1-shot	5-shot
0.003	10	10	5	5	8	4

8
9
10
11
12

15
16
17
18

The experimental results of 5-way 1-shot in MiniImagenet dataset

2
3
4
5
6
7
8

10   Table 5 The experimental results of 5-way 1-shot in MiniImagenet

11   dataset

Model	Fine-tunning	Accuracy
MAML	Y	48.69±1.47%
Meta-SGD	Y	50.46±1.23%
Multi-scale relational networks	N	50.21±1.08%
Multi-scale meta-relational network	N	50.56±0.93%

13
14

17
18
19
20
21

The experimental results of 5-way 5-shot in Miniimagenet dataset

2

Table 6 The experimental results of 5-way 5-shot in Miniimagenet dataset

Model	Fine-tunning	Accuracy
MAML	Y	63.10±0.82%
Meta-SGD	Y	64.03±0.66%
Multi-scale relational networks	N	65.89±0.32%
Multi-scale meta-relational network	N	66.23±0.55%

6
7
8
9
10