ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

A Review of Few-shot Learning

Shangwan Yu[1, a]

[1]School of Computer and Information Technology, Liaoning Normal University, Liaoning, China

[a]E-mail: 2847790665@qq.com

Abstract

Few-shot learning is an emerging research field in recent years, aiming at solving machine learning tasks with limited samples.Since a large number of samples cannot be obtained for many scenarios in the real world, such as lung cancer in the medical field, the method of few-shot learning is popular in many fields. However, due to the limitation of sample size, the problem of model overfitting is often caused. This paper first describes the definition of few-shot learning, systematically sorts out the current work related to few-shot learning, and specifically introduces the research progress of three types of few-shot learning models based on data enhancement, transfer learning and meta-learning. Then the siamese network, the model-agnostic meta-learning and the prototype network are studied in three small samples The classical method is described in detail. Finally, the future development direction of few-shot learning is prospected.

Keywords

Few-shot learning; Siamese network; Model-agnostic meta-learning; Prototype network.

1. Introduction

With the rapid development of artificial intelligence and machine learning, the research of machine intelligence has attracted wide attention. In recent years, deep learning algorithms based on a large number of annotated samples have made outstanding achievements in many fields such as computer vision and natural language processing. However, in the real world, a large amount of data cannot be obtained in many scenarios or the cost of annotating a large amount of data is too high. Therefore, research based on a small number of samples has gradually been favored by people. Children have the ability to quickly learn to recognize animals from a small number of samples. Analogous to human intelligence, few-shot learning is designed to be quickly learn a model to solve a problem when machines have a small number of labeled samples.

Few-shot learning first emerged in the field of computer vision, and has achieved advanced results in image classification, image recognition and other tasks. Face recognition has become the hottest technology in this wave. In recent years, few-shot learning has also taken its place in the fields of natural language processing and sound signal processing. Emotion analysis, the voice in navigation software and Siri have been deeply embedded in our daily life. In addition to its applications in the fields of images, text and sound, few-shot learning is increasingly being applied in medicine, drug development and disease diagnosis [1] and other problems have achieved remarkable results. Few-shot learning can also be applied in the field of robotics, which trains robots to complete specific tasks.

Compared with traditional weakly supervised learning, Not only few-shot learning can deal with the problems of classification and regression, but also solve the problem of reinforcement learning. At present, the methods to solve few-shot learning are mainly divided into three categories [2], the first method is to solve the problem of insufficient sample number through

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

the strategy of enhancing the data set, the second method is to constrain the original sample space into a smaller sample space according to the prior knowledge, and the third method is the search strategy of the optimization algorithm, through good initialize parameters and fine tuning.

Starting from the definition of few-shot learning, this paper introduces the related research work of few-shot learning, and elaborates in detail the model structure of siamese network, the calculation methods of two kinds of loss functions, the prototype network and the model structure and implementation method of the model-agnostic meta-learning method.

2. Problem Definition

Machine learning is that a computer program learns from experience related to a particular task and improves its performance. Few-shot learning is a type of machine learning in which a computer program learns from a task-related experience and gains performance improvements. Few-shot learning experiences contain only a small amount of supervisory information. The data set of few-shot learning is divided into training set and test set. The training set contains many categories, and each category has a number of samples, K categories are randomly selected, and N samples from each category constitute a k-way n-shot support set. The training set is defined asD_N, the training set consists of a set of data pointsX_N,such as a picture.$Y_N$is labels that correspond to these

images Composition. Support set is defined as $D_{support}$that can be defined in the

following form:

$$D_{support}=\{(x_{ij},y_{ij})\}_{i=1,j=1}^{i=k,j=n}\sim D_N^n\ ,y_{ij}\in Y_N \tag{1}$$

From the remaining samples of k class, n 'samples of each class are randomly selected as the query set.

$$D_{quary}=\{(x_{ij},y_{ij})\}_{i=1,j=1}^{i=k,j=n'}\sim D_N^{n'} \tag{2}$$

In this way, the training set can be divided into several combinations of categories. The test set and the training set deal with samples in the same mode and divide them into support set and query set in the same way. Define the test set asD_T, the test set consists of a set of data pointsX_T,and a set of category labels $Y_T$corresponding to

these imags. The support set can be defined as:

$$D_{support}=\{(x_i,y_i)\}_{i=1}^{N_s}\subset X_T\times Y_T \tag{3}$$

Then select from the remaining samples of the test set$N_q$as a query set.

$$D_{quary}=\{(x_i)\}_{i=1}^{N_q}\sim D_T^{N_q} \tag{4}$$

Different from traditional machine learning, the category of test set in the few-shot learning method never appears in the training set. So the purpose of few-shot learning is to learn to distinguish the similarities and differences between things.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

3. Relevant Work

Due to the limitation of sample size, based on a small sample of fitting problems of machine learning, so one way to solve the few-shot learning is enhanced data, can through the conversion program to expand. A transformation program can be learned a set of geometric variations from similar classes, also can learn a set of automatic encoder from similar class. Benaim et al [3] proposed a variant for the single sample learning strategy. Autoencoder clone creates a copy of the method to expand the sample. Schwartz et al [4] proposed a method for learning to synthesize new samples of unseen categories based on an improved autoencoder is proposed. Training sets can also be enhanced by summarizing and adapting pairs of input and output from similar but larger datasets. Goodfellow et al [5] proposed generative adversarial network (GADN), which is a deep neural network architecture composed of two competing networks, is proposed to learn and simulate the distribution of data for the realization of numbers the enhancement of data.

Transfer learning is a common method to solve few-shot learning at the present stage, which can relieve the heavy task of traditional supervised learning to annotate a large number of data. Transfer learning is to apply knowledge learned in a certain field to different related fields. The siamese network introduced in this paper is a method of transfer learning, Koch et al [6] applied the Siamese network to few-shot learning, and the two sub-networks share weight parameters, and classification is performed by calculating similarity. Yang et al [7] migrated the Inception-V3 pre-training model and tested in a data set with 10 types of fruits, and high test accuracy was obtained. Long et al [8] proposed a method that the difference of joint distribution is measured by using JDD regular term, and the tagged data of source domain is fine-tuned to make the joint distribution of source domain and target domain similar.

Meta-learning is also a common method to solve few-shot learning. Meta-learning aims to make machines learn to learn. Meta-learning methods have been proposed and have been studied extensively by Ye Hanjia et al [9]. A priori method of model combination is proposed, which decomposes the model structure through the optimal conditions of objective functions, and estimates each component of the model respectively to obtain an effective classifier, so as to realize meta-learning. Al - Shedivat et al [10] developed a simple gradient - based meta-learning method for dynamic and adversarial scenarios. A new multi-competition environment was designed to test all aspects of samples continuous adaptation.

4. The Classic Approach

In this chapter, we mainly introduce three classical algorithms: Siamese network, model-agnostic meta-learning, and prototype network.

4.1. Siamese Networks

Siamese Network is a meta-learning method based on transfer learning to solve few-shot learning tasks. Implement small samples by migrating the method of differentiating sample similarity learned in other tasks fast learning of new tasks under conditions.

4.1.1. Structure of Siamese Network

The so-called siamese network is a pair of conjoined sub-neural networks, the two sub-neural networks share parameters and the same structure. As shown in Figure 1, for a pair of input samplesX_1andX_2, the feature vector f(X_1) and f (X_2) are extracted through convolutional neural network, and prediction the similarity of the samples by calculating the distance between the two sample features | | f (X_1)- f(X_2) | |

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

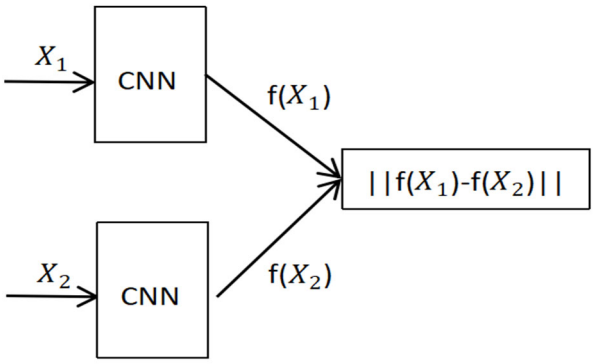

Figure 1. Structure of siamese network model

4.1.2. Two Methods for Calculating Losses

The first method of Cross Entropy Loss is to construct data sets into equal numbers of positive and negative samples. Positive samples can enable the neural network to learn what is the same type of information, while negative samples can enable the neural network to learn the differences between things. A picture is extracted from the training set, and another picture is extracted from the same class, and the 'Target' is set to 1 to get a positive sample. The positive sample set is composed of repeated positive sample sampling. The second sample in the negative sample is randomly sampled from a data set of a different category from the first sample, and the 'Target' is set to 0 to represent the two samples the similarity is 0.

The training process of positive samples is shown in Figure 2. Two samples of the same category are input, which are defined as X_1 and X_2.By convolution neural network to extract the most representative characteristics of f (1) and f (2), and the characteristics of two absolute value all the elements of subtraction of | f (1) - f(2) | to get the difference between the two characteristics. Through the connection layer and activation function, get a real number between 0 and 1 is used to measure the degree of similarity between two images. Since both images are in the same category the output should be close to 1.Loss function is defined as the difference between the label and the predicted output. The gradient is calculated through reverse relay, and the model parameters are updated with gradient descent to complete the training. The training process of the negative sample is the same as that of the positive sample. The input of the negative sample is two samples of different categories. The label is set to 0, hope the neural network's prediction is close to zero.

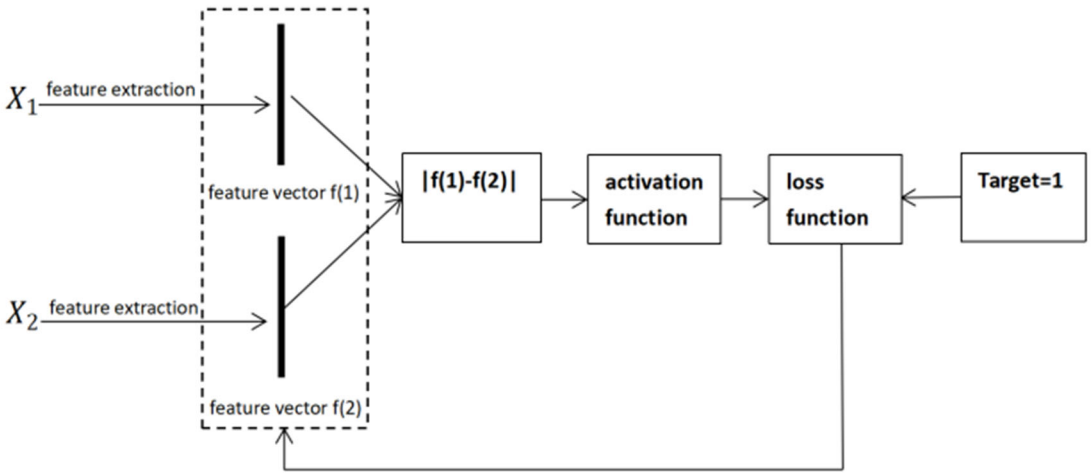

Figure 2. Training process of positive samples

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

The second loss calculation method (Triplet Loss) is shown in Figure 3. A random image was selected as the anchor point in the training set, which is expressed as X^a.The random sample in the same category as the anchor points is the positive sample and is denoted byX^+ ,the random sample in the different category as the anchor points is the negative sample and is denoted byX^-. In the Siamese network, the distance between the the anchor point and the positive sample eigenvector $d^+$and the distance between the anchor point and the negative sample eigenvectord^-. Since the positive samples are of the same category as the anchor points, the purpose of the training model is to make d^+small enough,d^-large enough.

The loss function is defined as follows, where α is a custom parameter. If $d^- \geqslant d^+ + \alpha$ then the loss is the loss function is zero, otherwise the loss function is $d^+ + \alpha - d^-$.

$$\text{Loss}(X^a, X^+, X^-) = \max\{0, d^+ + \alpha - d^-\} \tag{5}$$

The parameters of the convolutional neural network are updated by gradient descent algorithm to reduce the value of the loss function.

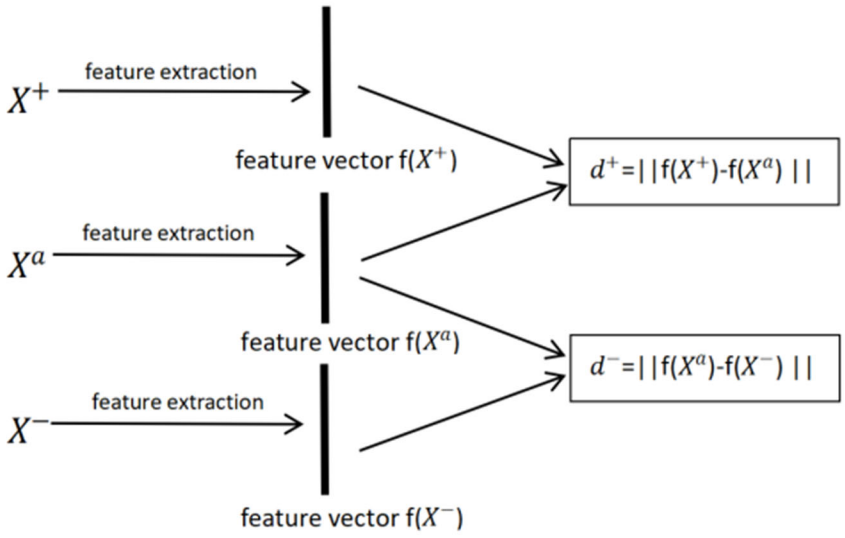

Figure 3. Method 2 training model

4.1.3. Two Test Methods for Calculating Loss

In both methods, the test sets are composed of a support set with known labels and a quary set with unknown labels, and the categories of samples in the test set have not appeared in the training set. Method 1: By comparing the similarity degree between the samples in each query set and each kind of samples in the support set to determine the category of samples. Method 2: Calculates the distance between each query set and each type of sample in the support set, the smallest distance is the category of the sample.

4.2. Model-independent Meta-learning

The basic processing unit of Meta Learning method is task. The aim is to train a model on a variety of different learning tasks so that the trained model combined with a small number of training samples can quickly solve new learning tasks. Different from traditional machine learning, meta learning solves the problem of learning a universal basic parameter rather than a mathematical model directly used for prediction.

In 2017, Finn [11] proposed the Model-Agnostic Meta-Learning (MAML) method. The core of this method is to initialize reasonable model parameters. If the model parameters are initialized

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

randomly, the ideal output results can be obtained after a large number of iterations. If the initial parameters of the model are reasonable, it does not need a large number parameters, through a small amount of training can get the final model.

4.2.1. Basic Idea of Algorithm

The figure shows the process of machine learning based on gradient descent method.$\theta^0$is the initialized parameter, the gradient is calculated according to the initial parameters and the training data, and is represented by g. The initial parameters are updated according to the gradient descent, and the final training parameters $\theta^*$are output after several training updates. In traditional machine learning, the network structure, initialization parameters, and method of updating parameters are all artificially designed in advance. The essence of meta-learning algorithm is to let the machine learn how are these designed, and the goal of model-agnostic meta-learning is a parameterθ^0 for learning initialization.

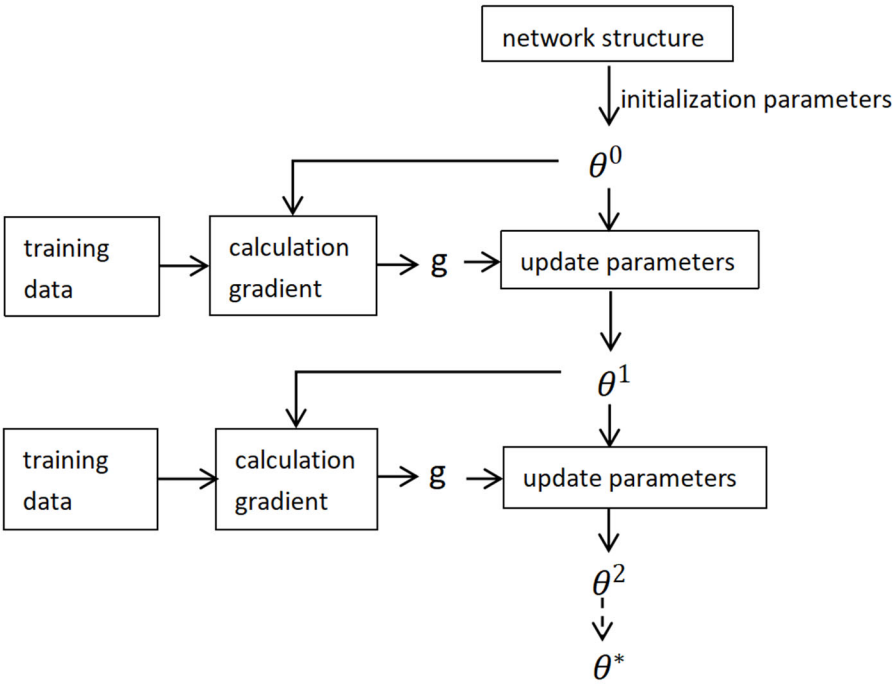

Figure 4. Machine learning process based on gradient descent method

4.2.2. Definition of Loss Function

Since the meta-learning method takes the task as the unit, the loss function L(F) is defined as the sum of the loss functions of N test tasks. The loss function for each test task is expressed as the difference between the output passing through the model and the original label.

$$L(F)=\sum_{n=1}^{N} l^n \tag{6}$$

Where: l^n—the loss function of the Nth task

In the meta-learning algorithm independent of the model, the initial parameters are obtained through learning. Let the initialization parameter of the meta-learning algorithm independent of the model be Φ. By doing training on different tasks, you end up learning different models. The loss function of model-agnostic meta-learning is defined as:

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

$$L(\Phi)=\sum_{n=1}^{N} l^{n}(\hat{\theta}^{n}) \tag{7}$$

Where: $\hat{\theta}^{n}$ —The model learned in the Nth task is related to the initialization parameter Φ.

$l^{n}(\hat{\theta}^{n})$ —the loss function of the Nth task

The loss function is minimized by gradient descent in training. In practice, due to small number of samples for sample learning is limited, and multiple gradient descent is likely to overfit, so the element is independent of the model.

The learning algorithm only does one gradient descent and outputs the final model training results θ^{*}It can be expressed as:

$$\theta^{*}=\Phi-\varepsilon\nabla_{\Phi} l(\Phi) \tag{8}$$

Where: ε—learning rate

$\nabla_{\Phi} l(\Phi)$—Φ gradient on a task.

4.2.3. MAML Model Training Ideas

Firstly, the task set P (T) and the super-parameter learning rate for training should be prepared. Among them, the combination of fewer training tasks but different tasks can form a large number of samples. The purpose of MAML is to learn from a large number of different tasks, so that the model has a strong enough generalization ability. The learning rate determines the weight updating speed of the model. Two learning rates are set in the algorithm, in which α is used to update the parameters of a single task and β is used to update the parameters of the meta-learner.

The algorithm starts by initializing the parameter Φ of the meta-learner, loops the input tasks, and calculates the current task support set loss gradient, and update the intermediate variable with gradient descentθ_i. At the end of the inner loop, the loss function of the query set for the current task is $L_{T_i}(f_{\theta_i})$, summing the loss function for all tasks $\sum_{T_i\sim p(T)} L_{T_i}(f_{\theta_i})$, calculate its gradient with respect to the parameter Φ, and update the parameter bar so that the loss function is maximum.

Table 1. Algorithm MAML-Training

Require1: Task distribution p(T)

Require2: α is the parameter of the base learner, β is the parameter of the meta learner

1: randomly initialize Φ

2: while not done do

3: Sample batch of tasks T_i~p(T)

4: for all T_i do

5: Evaluate $\nabla_{\Phi} L_{T_i}(f_{\Phi})$ with respect to K examples

6: Evaluate $\theta_i=\Phi-\alpha\nabla_{\Phi} L_{T_i}(f_{\Phi})$

7: end for

8: updat$\Phi\leftarrow\Phi-\beta\nabla_{\Phi}\sum_{T_i\sim p(T)} L_{T_i}(f_{\theta_i})$

9: end while

10: go back to step 3 and sample again

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

4.3. Prototype Network

The disadvantage of the siamese network is that it is necessary to compare the test samples with each sample in the support set to calculate the similarity, so as to determine the category of the test samples. Jake Snell et al [12] proposed the prototype network, and the test samples only need to be compared with the clustering center which supports each kind of samples in the test set, which reduces the time complexity degrees.

4.3.1. Basic Idea of the Algorithm

For the classification problem, the prototype network algorithm looks for the prototype center of each class in the sample space. Through the convolutional neural network, the input image x is transformed into a feature vector f(x), assuming that the data set has k categories, then the prototype center can be expressed as:

$$C_k = \frac{1}{S_k} \sum_{(x_i, y_i) \in S_k} f(x_i) \tag{9}$$

Where: $f(x_i)$—Support eigenvectors of set samples

y_i— Label of the sample

S_k—The number of samples a category supports in a set

The distance between each sample in the query set and each type of prototype is calculated by the softmax function. The highest probability by calculation is the category of the test sample.

$$p(y = k \mid x) = \frac{\exp(-d(f(x), c_k))}{\sum_{k'} \exp(-d(f(x), c_{k'}))} \tag{10}$$

Where: d (f (x),c_k) — Distance of the sample feature of the query set to f(x) and center c_k of the sample prototype k '—the true category corresponding to the sample

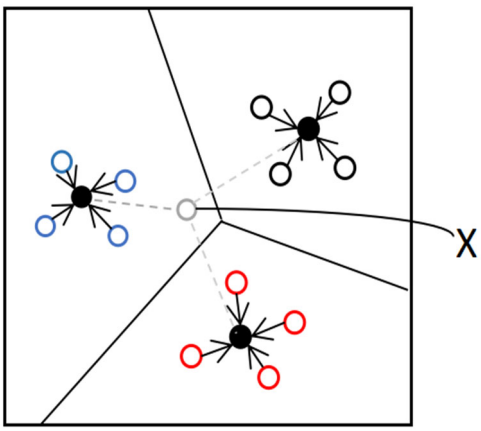

Figure 4. Basic idea of prototype network

The loss function is minimized through stochastic gradient descent, and the loss function is defined as:

$$L = -\log p(y = k' \mid x) \tag{11}$$

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

4.3.2. Algorithm Idea

The core idea of the training algorithm is shown in the table, which includes two cycles. The first cycle is used to select training support samples and query samples, and to calculate the sample center of each type of support samples. The initial loss function value is 0. The second loop updates the value of the loss function.

The training set is composed of N samples, which have a total of K categories and N_c samples are randomly seleceted constitute the support setS_k, random selection$N_Q$samples constitute the query setQ_k, the sample centerc_k of each category of support set sample is calculated.Set the initial loss function value L to 0, and update the loss function value through the test set samples.

Table 2. Prototypical Network-Training

Input:train set $D=\{(x_1, y_1),.....,(x_N, y_N)\}, y_i \in \{1,.....,K\}$

Output:the value of loss function L

1: V←Random Sample ({1,...,K},N_c)

2: for k in {1,.....N_c}do

3:S_k ←Random Sample ($D_{V_k}.N_s$)

4:Q_k←Sample ($D_{V_k} \backslash S_k.N_Q$)

5: $c_k \leftarrow \frac{1}{N_c} \sum_{(x_i,y_i)\in S_k} f(x_i)$

6: end the for

7: L← 0

8: for k in {1,.....N_c}do

9: for (x, y) in Q_k ($L \leftarrow L+\frac{1}{N_c N_Q}[d(f(x),c_k))+\log \sum_{K'} \exp(-d(f(x),c_k))]$

10: end for

11: end for

5. Conclusion

This paper first gives a detailed description of the definition of few-shot learning, and sorts out the related work of few-shot learning, and then explains in detail the siamese network, model-agnostic meta-learning method, prototype network three classic few-shot learning methods. Existing few-shot learning methods usually use a single form of information. In the future, few-shot learning can be used to deal with multi-form information, and more research achievements can be made in the fields of health care, endangered species detection and so on.

References

[1] Zhao Y. Convolutional neural network based carotid plaque recognition over small sample size ultrasound images [MS. Thesis]. Wuhan: Huazhong University of Science and Technology, 2018.

[2] Yaqing Wang, Quanming Yao,James T.Kwok.2020.Generalizing from a few examples:a survey on few-shot learning. arXiv preprint at arXiv.

[3] S. Benaim and L.Wolf.2018.One-shot unsupervised cross domain translation. In Advances in Neural Information Processing Systems. 2104-2114.

[4] E.Schwartz,L.Karlinsky,J.Shtok.2018.Delta-encoder:An effective sample synthesis method for few-shot object recognition.In Advances in Neural Information Processing Systems.2850-2860.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202109_4(9).0035

[5] I.Goodfellow,J.Pouget-Abadie,M.Mirza. 2014. Generative adversarial nets. In Advances in Neural Information Processing Systems. 2672-2680.

[6] Koch G,Zemel R,Salakhutdinov R.2015.Siamese neural networks for one-shot image recognition.In International Conference on Machine Learning.

[7] Yang G,Tang B,Cao S.2019.Research on small sample image recognition based on transfer learning. Journal of Physics: Conference Series.

[8] Long M S, Zhu H, Wang J M.2017.Deep transfer learning with joint adaption networks. Proceedings of the 34th International Conference on Machine Learning.

[9] Ye HJ, Zhan DC. 2018. Small sample learning based on model decomposition. Science in China: Information Science.

[10] Al-Shedivat M,Bansal T,Burda Y.2017.Continuous adaptation via meta-learning in nonstationary andcompetitive environments.arXiv preprint at arXiv.

[11] Chelsea Finn, Pieter Abbeel,Sergey Levine.2017.Model-Agnostic Meta-Learning for fast adaptation of deep networks.arXiv preprint at arXiv.