# A Survey on Meta-learning Based Few-Shot Classification

Weizhi Huang[1], Ming He[2(✉)], and Yongle Wang[2]

[1] School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China
[2] College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
heming@hrbeu.edu.cn

**Abstract.** Data-intensive applications have achieved great success in the field of machine learning. How to ensure that the machine can still learn correctly in the absence of labeled samples is the next challenging problem to be solved. This paper first introduces the problem definition of few-shot learning. Secondly, the existing small few-shot learning methods based on meta-learning are comprehensively summarized. Specifically, they are divided into three categories: metric-based learning methods, optimization-based learning methods and model-based learning methods. We conducted a series of comparisons among various methods in each category to show the advantages and disadvantages of each method. Finally, the limitations of existing methods are analyzed, and the future development direction of few-shot learning research is prospected.

**Keywords:** Few-shot learning · Deep learning · Meta-learning

## 1 Introduction

Inspired by the human process of learning new things, we hope that machines can also learn new knowledge from a few samples. In recent years, few-shot learning has attracted more and more attention. At the same time, meta-learning methods [23] are also developing rapidly. The purpose of meta-learning is to let the machine learn to learn. In practice, it is found that the meta-learning method fits well with the few-shot learning problem. The main purpose of this paper is to conduct a comprehensive study of meta-learning methods based on solving few-shot problems, focusing on the analysis of typical strategies. In this paper, the existing few-shot learning methods are divided into three categories, namely, metric-based learning methods, optimization-based learning methods and model-based learning methods. The latest research on these three categories will be discussed separately. It must be noted that there is no clear boundaries between these three categories. This paper will analyze various methods of each category in detail.


X. Jiang (Ed.): MLICOM 2021, LNICST 438, pp. 243–253, 2022.
https://doi.org/10.1007/978-3-031-04409-0_23

The rest of the article is structured as follows. Section 2 is an overview of few-shot learning, including definition, problem description and common data sets. Section 3 introduces the past learning problems related to few-shot learning and makes comparisons to clarify the application scenarios of small sample learning. Section 4 introduces metric-based learning methods for few-shot learning. Section 5 introduces the few-shot learning methods based on optimal learning strategies. Section 6 introduces the few-shot learning approach based on optimal model architecture. Finally, Sect. 7 makes a conclusion about the limitations of existing methods and future research.

## 2 Overview

### 2.1 Definition and Problem Description

Few-shot learning is a kind of machine learning problem that how to learn the model in the case of insufficient effective samples. The dataset contains three categories: Training set, Support set, and Query set. There are large-scale labeled data in the training set, containing numerous different classes, and many samples in each class, which is used to train the model. The support set usually contains N classes. Generally, there are no classes intersection between the support set and the training set. Each class of the support set has K samples. According to the size of the support set, we call the few-shot learning problem based on the support set as the N-way K-shot problem. The query set is used for the final model test. The goal of few-shot learning is to use the model trained by the training set to identify the labels of the samples in the query set. The class of the query set is not included in the training set but included in the support set, that is, the class of the query set is new for the model.

### 2.2 Major Application

Most of the existing few-shot learning is applied in the field of computer vision, such as handwritten character recognition [17] and image classification [15], because it's easy to obtain the visual information. And this method has been well-tested on previous machine learning problems. At present, there are two benchmark datasets miniImageNet [29] and Omniglot [16] in image classification and character recognition, and they have achieved high accuracy on two datasets. Therefore, more computer vision applications can be explored, such as image segmentation, neural style transfer, image reconstruction and image generation. In addition to computer vision applications, other fields have been gradually using the idea of few-shot learning such as few-shot translation and few-shot language modeling in natural language processing.

## 3 Relevant Learning Problems

In the field of machine learning, there are many cross-domain learning problems with few-shot learning, including Weakly Supervised Learning [31], Transfer Learning [20] and Multitask Learning [3]. This section will clarify the relevance and differences between these problems and few-shot learning, so as to determining its applicable scenarios.

Weakly supervised learning: including Semi-supervised learning [32] and Active learning [25]. Semi-supervised learning refers to learning optimal assumptions in mixed data with and without labels. Active learning refers to reducing the cost of labeling by some technical means or mathematical methods. Few-shot learning is different from this, it can be supervised learning, semi-supervised learning and reinforcement learning, which depends on what data are available in addition to limited supervised information.

Transfer Learning: Use the experience of the source task to improve the learning of the target task. The knowledge learned from the source domain and source task of a large amount of training data is transferred to the target domain and target task with limited training data. The inner thinking is based on human inferences, whiles improve the utilization of data. Transfer learning method is widely used in few-shot learning. When the given supervision information is limited to direct learning, few-shot learning needs to transfer prior knowledge from the source task to the current few-shot learning task.

Multi-task Learning: Multi-task learning is a derivation transfer learning method. The main task uses the domain-specific information possessed by the training signal of the related task as an inductive bias to improve the generalization effect of the main task. Multi-task learning involves parallel learning of multiple related tasks at the same time, the gradient is backpropagated at the same time, and multiple tasks help each other learn through the underlying shared representation to improve the generalization effect. To put it simply, multi-task learning puts multiple related tasks together to learn. In the learning process, a shallow shared representation is used to share and complement each other with information related to the learned field, promote each other's learning, and enhance generalization effect.

## 4 Metric-Based Learning Approach

In the metric-based learning method, we often use the metric criteria we designed to judge the distance of the sample in the feature space. Such learning methods generally include a feature encoder $E$ and a metric function $M$. The feature encoder $E$ is used to extract the input feature and convert it into a feature vector on the new feature space. The distance between the vectors is judged by the metric function $M$. Commonly used measurement criteria are Euclidean Distance, Minkowski Distance, Cosine Similarity and so on. In the problem of few-shot learning, the model trained on the training set by this method has the ability to judge the similarity between samples.

### 4.1 Fixed Distance Metric

Vinyals et al. used the model architecture of Matching Networks [29] to deal with the problem of few-shot learning. It is mainly embedding [12] the support set and the query set, and then uses the query set sample to calculate the attention of each support set sample, and the label of each category is linearly weighted according to the attention score to determine the category of the sample. Snell et al. also proposed a Prototypical Networks [26] model similar to Matching Networks, whose main idea is to take the average obtained by embedding the samples of each category as the characterization of the category, and determine its category by the distance between the query sample and the average characteristics of each category in the feature space.

Siamese Network [14] is a two-way neural network. Koch et al. used Siamese Network to solve few-shot image classification. The core idea of the method is to use the training set to train a neural network, so that the neural network has the ability to identify whether the characteristics of two samples are similar, and to classify the test samples in this way. There are two main methods for constructing training samples. One is by constructing positive and negative samples, sampling two pictures at a time, and if the same type is set as a positive sample, otherwise it is a negative sample. Use positive and negative samples to train a neural network to measure the similarity between two pictures. The second method is the idea of using Triplet Loss [24] proposed by Schroff et al., which is to construct triples. With the anchor as the center, the same kind is denoted as positive, and different types are denoted as negative. Use the metric function to calculate the distance between the anchor and the positive and negative in the feature space, denoted as $d^+$ and $d^-$. The model should make $d^+$ as small as possible, and $d^-$ as large as possible, and $d^-$ must be much larger than $d^+$, otherwise the model cannot distinguish between the two categories.

Similar to Siamese Network, Relation Network [27] first uses an embedding module to map the supported images and query images into feature vectors. But unlike Siamese Network, Relation Network does not directly calculate their distance after obtaining these vectors, but first connects their feature maps, and then passes through a relation module e to get their relation score, and finally according to This relationship score is classified.

In the case of a small number of samples, the models trained by conventional neural networks often have limited accuracy, and the pre-training method can solve this problem well. Pre-training refers to training a neural network on other large-scale labeled similar data sets to obtain a set of model parameters, initialize the model with the learned model parameters, and then perform fine-tune on the model in a small-sample task. The training method has been proven to have good results in few-shot learning.

Chen et al. first proposed the concept of Classifier-Baseline [4], which is to pre-train a classifier on the base class to learn visual representations, then delete the last fully connected layer that depends on the class, and use the cosine distance in the feature space, Use the nearest centroid to classify the query sample. This process can also be seen as estimating the weight of the fully

connected layer of the new class, but there is no need to train parameters for the new class. The article verifies that the Classifier-Baseline method is superior to many more advanced algorithms. Then the author proposed to use meta-learning to improve Classifier-Baseline, and proposed Meta-Baseline [4]. In Meta-Baseline, use the pre-trained Classifier-Baseline to initialize the model, and use the cosine similarity measure to perform meta-learning, that is, use the support set to fine-tunning the initialized model. The model is shown in Fig. 1.

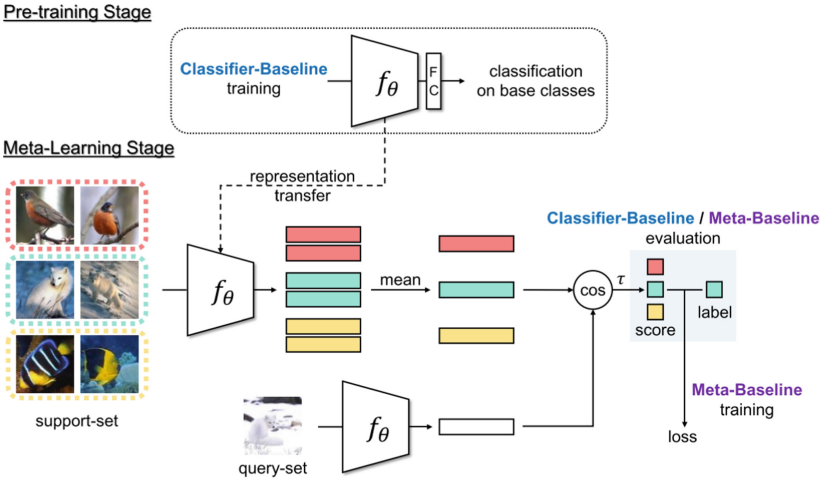


**Fig. 1.** Classifier-baseline & meta-baseline.

Dhillon et al. made improvements in the fine-tune stage of the pre-training method, and mainly proposed transductive fine-tuning [5], which is to fine-tune the deep network trained with standard cross-entropy loss. The performance of this method in the standard data set is better than the latest technology with the same hyperparameters.

Table 1 summarizes the performance of the metric-based approach on mini-Imagenet.

**Table 1.** Few-shot classification results trained with the mini-ImageNet dataset.

| Model | Backbone | 5-way 1-shot | 5-way 5-shot |
|---|---|---|---|
| Matching networks | ConvNet-4 | 46.60 ± 0.84% | 60.00 ± 0.73% |
| Prototypical networks | ConvNet-4 | 49.42 ± 0.78% | 68.20± 0.66% |
| Relation networks | ConvNet-4 | 50.44 ± 0.82% | 65.32 ± 0.70% |
| Classifier-baseline | ResNet-12 | 58.91 ± 0.23% | 77.76 ± 0.17% |
| Meta-baseline | ResNet-12 | 63.17 ± 0.23% | 79.26 ± 0.17% |

### 4.2 Metric-Based Cross-domain Learning

From the perspective of the generalization performance of small-sample classification, many existing metric-based methods have problems. That is, there are significant differences in the distribution of image features extracted from tasks in different domains. Therefore, as shown in Fig. 2, in the training phase, the metric function may overfit the feature distribution encoded only from the known domain, resulting in the inability to generalize to other domains.

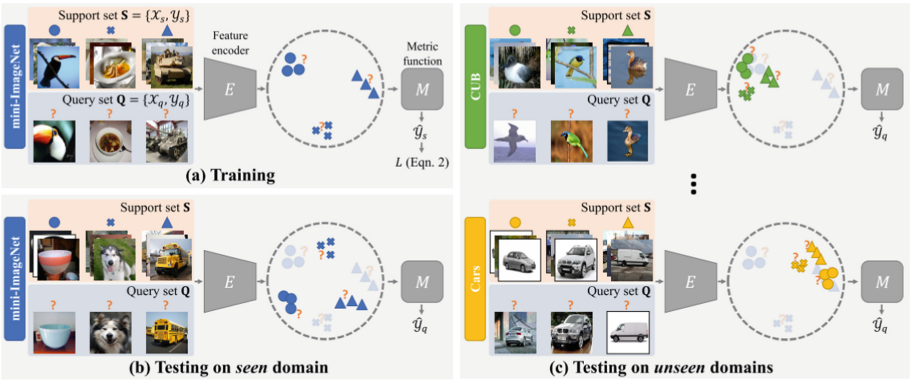


**Fig. 2.** Cross-domain problem formulation and motivation.

To solve the cross-domain problem, Tseng et al. proposed Feature-wise Transformation Layer [28], which was inserted after the BN layer [10] of Feature Encoder. The core idea is to use a feature-based transformation layer to enhance image features through affine transformation in the training phase to simulate various feature distributions in different domains, thereby improving the generalization ability of the measurement function in the test phase. In addition, there are two hyper-parameters in Feature-wise Transformation Layer that require careful manual adjustment, because it is difficult to model complex changes in image feature distribution in different fields. Based on this, the article developed a learning-to-learn algorithm to optimize the proposed feature conversion layer, that is, let the model learn hyperparameters by itself.

## 5 Optimization-Based Learning Approach

In traditional learning methods, many of the entire training steps need to be designed by humans, including the network architecture, the initialization parameters, and the way to update the parameters, etc. When we choose a different design, we get a different learning method. When the training steps are adjusted artificially, it is difficult to achieve the desired efficiency and accuracy. The optimization-based learning method considers whether the machine can learn to part of the training steps by itself.

### 5.1 Parameter Optimization Method

Finn et al. proposed the model-independent meta-learning method MAML [7], which pioneered optimization-based methods. MAML optimizes by letting the machine learn the initialization parameters itself, and the method expects the learned initialization parameters to achieve optimal results with a few updates. MAML can be applied not only to few-shot classification problems, but also to reinforcement learning and regression problems with better results. Nichol et al. proposed Reptile [19] with some improvements on MAML, firstly, by simplifying the parameter update operation and ignoring the second-order differentiation operation to improve the speed of the operation while maintaining its performance. Secondly, the rules of parameter update are relaxed and the constraints of parameter update are reduced.

In addition, Ravi et al. summarize the reasons why deep learning-based optimization algorithms are not applicable in less sample learning. Ones are that gradient-based optimization algorithms such as Adam [13], AdaDelta [30], and Adagrad [6] are not suitable for few-sample situations with a limit on the number of parameter updates. Secondly, for multiple separated tasks, random initialization parameters will affect the task's ability to complete optimization after a few updates. The article found a great similarity between the LSTM [9] internal update and the gradient descent process, based on which the LSTM-based gradient descent method [21] was proposed to allow the network to learn the LSTM network parameters as well as the initialization parameters by itself, enabling it to learn different tasks quickly. Unlike Ravi et al. Andrychowicz et al. proposed that most of the standard optimization procedures consider only first-order differentiation without considering second-order differentiation, which obviously results in a loss of performance. To solve such problems, an LSTM-based optimizer [1] was designed to overcome the drawback of considering only first-order derivatives by using the memory function possessed by recurrent neural networks.

Table 2 shows the performance of the parameter optimization method on mini-Imagenet

**Table 2.** Results of the parameter optimization method on mini-ImageNet.

| Model | Backbone | 5-way 1-shot | 5-way 5-shot |
|---|---|---|---|
| Meta-learn LSTM | ConvNet-4 | 43.44 ± 0.77% | 60.60 ± 0.71% |
| MAML | ConvNet-4 | 48.70 ± 1.84% | 63.11 ± 0.92% |
| Reptile | ConvNet-4 | 49.97 ± 0.32% | 65.99 ± 0.58% |

### 5.2 Cross-task Learning Approach Based on Parameter Optimization

Although meta-learning has achieved good results on many deep learning problems, including image classification and augmentation learning tasks, classical meta-learning approaches ignore an important issue of learning the optimal initial model on multiple tasks, i.e., how to ensure that the initial model obtained from learning is unbiased for all tasks.

Jamal et al. proposed a task-agnostic meta-learning method TAML [11] (Task Agnostic Meta-Learning), which enables the initial model to be treated equally for different tasks by adding a regularization term to it. The article designs two types of regularization, TAML based on entropy reduction maximization and TAML based on inequality minimization, to meta-train the initial model so that the model is less different when facing different tasks.

## 6 Model-Based Learning Approach

The model-based learning approach aims at finding the optimal architecture, where the model can update parameters quickly mainly thanks to the internal structure of the model or controlled by other meta-learning models. The major difference between this approach and the metric-based approach is that it does not make assumptions about the form of conditional probabilities, but relies on a model that can learn quickly.

In few-shot learning, where the goal of learning is to combine information learned in the past and learn new knowledge quickly, it is not surprising to use models with memory functions. Santoro et al. proposed a MANN [22] (Memory-Augmented Neural Networks) model based on a neural Turing machine [8] (Fig. 3). Although recurrent neural networks such as LSTM also have memory functions, they are mostly internal, whereas MANN uses memory to assist memory. MANN modifies the internal retrieval mechanism as well as the training settings and proposes a new addressing mechanism for assigning attention weights to memory vectors. The model has two major advantages of storing stable information and the size of the storage space is not limited by the size of the model parameters.

Furthermore, Munkhdalai et al. proposed the Meta-Networks [18] model, which consists of two main learners, namely the base learner and the meta-learner, in addition to also being equipped with external memory. The base learner is used to generate parameters Slow Weights based on common optimization methods (e.g., SGD), and the meta-learner is used to generate parameters Fast Weights for another neural network, which are mainly used for cross-task generalization. The combination of Slow Weights and Fast Weights is used throughout the neural network for prediction.

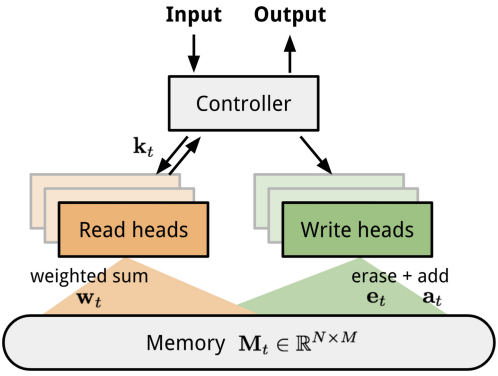


**Fig. 3.** Neural turing machine.

## 7 Conclusion

In this paper, we summarize and analyze the classical methods of few-shot learning from different perspectives by comparing the latest research. The metric-based learning method is limited because it is prone to overfitting when the number of samples is too small, and the method is relatively picky about the dataset, which can appear to perform well on some tasks but poorly on others. And this method has low robustness. Optimization-based learning methods or parametric methods usually require multiple update steps to reach a better point when updating weights using gradient descent because of the limitations in optimizer selection and learning rate settings. It makes the learning process so slowly when the model process on a new task. Model-based approaches are very good at handling learning with few samples. But they are usually poor at generalization because it is ambiguous whether the model can successfully embed a large training set into a base model.

Research on few-shot learning with deep learning has grown rapidly over the past few years, and as a result, applying few-shot learning models to practical applications will receive more attention. In this case, how to ensure the accuracy and computation efficiency of the model at the same time is one of the most challenging problems. Existing few-shot learning methods usually use previous knowledge from a single modality, while prior knowledge from multiple modalities [2] can provide prior knowledge for complementary views, but different modalities may contain different structures and need to be handled carefully. The use of multi-modal information in the design of few-shot learning methods is a direction for future research.

## References


1. Andrychowicz, M., et al.: Learning to learn by gradient descent by gradient descent. arXiv:1606.04474 [cs], November 2016
2. Baltrušaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2019). https://doi.org/10.1109/TPAMI.2018.2798607
3. Caruana, R.: Multitask Learning. Mach. Learn. **28**(1), 41–75 (1997). https://doi.org/10.1023/A:1007379606734
4. Chen, Y., Wang, X., Liu, Z., Xu, H., Darrell, T.: A new meta-baseline for few-shot learning. arXiv:2003.04390 [cs], April 2020
5. Dhillon, G.S., Chaudhari, P., Ravichandran, A., Soatto, S.: A baseline for few-shot image classification. arXiv:1909.02729 [cs, stat], October 2020
6. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135. PMLR, July 2017
8. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv:1410.5401 [cs], December 2014
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
10. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. PMLR, June 2015
11. Jamal, M.A., Qi, G.J.: Task agnostic meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11719–11727 (2019)
12. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, MM 2014, pp. 675–678. Association for Computing Machinery, New York, November 2014. https://doi.org/10.1145/2647868.2654889
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 [cs], January 2017
14. Koch, G.: Siamese neural networks for one-shot image recognition, p. 30
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017). https://doi.org/10.1145/3065386
16. Lake, B.M., Salakhutdinov, R., Gross, J., Tenenbaum, J.B.: One shot learning of simple visual concepts, p. 7
17. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015). https://doi.org/10.1126/science.aab3050
18. Munkhdalai, T., Yu, H.: Meta networks. In: International Conference on Machine Learning, pp. 2554–2563. PMLR, July 2017
19. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. arXiv:1803.02999 [cs], October 2018
20. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191
21. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning, p. 11 (2017)

22. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: Meta-learning with memory-augmented neural networks. In: International Conference on Machine Learning, pp. 1842–1850. PMLR, June 2016
23. Schaul, T., Schmidhuber, J.: Metalearning. Scholarpedia **5**(6), 4650 (2010). https://doi.org/10.4249/scholarpedia.4650
24. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
25. Settles, B.: Active learning. In: Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 6, no. 1, pp. 1–114, June 2012. https://doi.org/10.2200/S00429ED1V01Y201207AIM018
26. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. arXiv:1703.05175 [cs, stat], June 2017
27. 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 the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1199–1208 (2018)
28. Tseng, H.Y., Lee, H.Y., Huang, J.B., Yang, M.H.: Cross-domain few-shot classification via learned feature-wise transformation. arXiv:2001.08735 [cs], March 2020
29. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems, vol. 29, pp. 3630–3638 (2016)
30. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. arXiv:1212.5701 [cs], December 2012
31. Zhou, Z.H.: A brief introduction to weakly supervised learning. Natl. Sci. Rev. **5**(1), 44–53 (2018). https://doi.org/10.1093/nsr/nwx106
32. Zhu, X., Goldberg, A.B.: Introduction to semi-supervised learning. In: Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3, no. 1, pp. 1–130, January 2009. https://doi.org/10.2200/S00196ED1V01Y200906AIM006