HiLL PUBLISHING

Advances in Computer and Communication, 2024, 5(4), 210-217
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Graph-based Embedding Propagation for Transductive Few-shot Classification

Kangwen Niu

Department of Artificial Intelligence, Hefei University of Economics, Hefei 230012, Anhui, China.

How to cite this paper: Kangwen Niu. (2024) Graph-based Embedding Propagation for Transductive Few-shot Classification. *Advances in Computer and Communication*, 5(4), 210-217.
DOI: 10.26855/acc.2024.10.002

Received: July 30, 2024
Accepted: August 27, 2024
Published: September 25, 2024

***Corresponding author:** Kangwen Niu, Department of Artificial Intelligence, Hefei University of Economics, Hefei 230012, Anhui, China.

Abstract

The purpose of few-shot learning is to identify underlying patterns in data using only a limited number of training samples. Few-shot classification aims to train a classifier to differentiate between previously unseen classes based on a small number of labeled instances. Few-shot classification is highly difficult due to unseen classes and data-poor issues. Extraction of features that are typical of each category is, in our opinion, important for classification tasks. However, many existing approaches extract features directly from a small number of samples inside each class without taking into account the relationship information between them, leaving the classes with insufficient discriminative features. In this paper, I propose novel extensions of the Transductive Propagation Network (TPN), called Co-Attention TPN (CA-TPN) which is a method to extract representative feature information by considering the relational information among input data channels and class samples. Additionally, the study conducts semi-supervised few-shot classification through graph-based embedding propagation. First, this study focuses on examining the correlation between input data channels and developing a metric score formula to assess the significance of each channel feature. This approach involves incorporating a channel attention mechanism that enables us to weigh the relevance of each channel feature accurately. Second, I explore the relationship between the samples in each class and propose a novel prototype model that utilizes relation attention. This prototype with relation attention serves as a replacement for the traditional embedding feature of a single sample and enhances subsequent embedding propagation and downstream classification tasks. The approach I employ is validated on various benchmark datasets, demonstrating superior performance compared to existing few-shot classification methods and attaining the current state-of-the-art outcomes.

Keywords

Few-shot learning; Classification; Graph-based; Transductive Propagation; Co-Attention TPN (CA-TPN)

1. Introduction

Humans have a remarkable ability to learn from a small number of examples, a capability that is exemplified by children's effortless formation of new concepts. For example, children have no problem forming the concept of “giraffe” by only taking a glance at a picture in a book or hearing its description as looking like a deer with a long neck [1]. However, despite impressive advances in recognition systems [2-5], the most successful models still require vast amounts of labeled training data, which can be challenging to obtain for rare or specialized categories and expensive for those that require expert annotation. The reliance on labeled data significantly limits the scalability of recognition systems to learn and generalize across the long-tail categories in the real world. These categories are often more

diverse and nuanced than the head categories, and learning from a small number of examples is critical to enable recognition systems to overcome the challenges of limited labeled data and recognize new categories. Additionally, the ability to learn from a small number of examples is essential for enabling recognition systems to adapt to new and evolving categories in the real world, making them more adaptable and flexible to changes. Finding ways to reduce the reliance on labeled data and enable recognition systems to learn from a small number of examples is critical for the future development of these systems, as the demand for accurate and flexible recognition systems continues to grow.

Few-shot learning is a challenging task in machine learning due to the limited availability of labeled data, with only one or a few examples per category. One of the major obstacles in this regime is the difficulty in estimating the data distribution without introducing an inductive bias. Furthermore, obtaining labeled data is often expensive and time-consuming, while unlabeled instances are more readily available in real-world applications. To mitigate the data scarcity problem, leveraging unlabeled instances to aid in the learning process of few-shot models is a promising solution. This gives rise to two approaches: 1) semi-supervised few-shot learning, also known as inductive semi-supervised few-shot learning, which leverages unlabeled data to augment the labeled data, and 2) transductive semi-supervised learning, which utilizes all available test data, both labeled and unlabeled, to learn the model. The lack of labeled data in few-shot learning makes it challenging to train effective models that can generalize well to unseen examples. To address this problem, researchers have explored various techniques such as data augmentation, transfer learning, and meta-learning. However, these approaches still require some level of labeled data to train the model. In contrast, semi-supervised few-shot learning and transductive semi-supervised learning leverage the abundance of unlabeled data to enhance the learning process. In semi-supervised few-shot learning, the model is trained on both labeled and unlabeled data, allowing it to learn a more robust representation of the data. This can be achieved by combining unsupervised learning techniques such as clustering, auto-encoders, and generative models with supervised learning. By leveraging the unlabeled data, the model can learn to generalize better to unseen examples and improve its accuracy. On the other hand, transductive semi-supervised learning utilizes all available data, both labeled and unlabeled, during the test phase to refine the model's predictions. This can be done by incorporating the unlabeled examples into the optimization objective, thereby leveraging the additional information to improve the model's performance. This approach assumes that the test set is representative of the true data distribution and can provide valuable information about the underlying structure of the data. Overall, both semi-supervised few-shot learning and transductive semi-supervised learning show promising results in improving the performance of few-shot models. However, they also come with their own challenges, such as the risk of overfitting the unlabeled data or the assumption that the test set is representative of the true data distribution. Therefore, careful consideration should be given to the specific requirements and constraints of the problem at hand when choosing between these approaches.

In this work, we propose a new method Co-Attention Transductive Propagation Network (CA-TPN) based on TPN [6]. The primary objective of the CA-TPN is to improve the accuracy and efficiency of few-shot learning by incorporating co-attention with the prototype self-attention mechanism and novel channel attention mechanism. In our approach, a new inter-channel attention mechanism is proposed. We design a score to evaluate the importance of each channel feature and weight the feature embedding with channel information based on this. Because for image classification data, the feature information with each channel is also different and has a different relationship with that category feature information. Therefore, it is crucial for us to explore the correlation of each channel with the category features in order to extract the feature embedding that best represents the category information. In addition, a novel prototype generation approach is proposed to generate more representative prototype nodes of category features by introducing a self-noticing mechanism that considers the relationship information between samples in the same category instead of the traditional averaging to generate prototypes. Instead of the original taking each sample as a manifold structure node. The purpose of doing so is to construct more category-specific nodes to build undirected graphs for subsequent label propagation.

2. Main Approach

2.1 Problem Definition

Our approach adheres to the episodic paradigm [7], a widely used technique in the literature [8-11] for training a meta-learner to perform few-shot classification tasks. In this scenario, the aim is to develop classifiers for a new set of classes, denoted as C_{test} with limited labelled examples, using a relatively extensive labelled dataset featuring a

range of classes, denoted as C_{train}. More specifically, in every episode, a limited of N classes are randomly selected from the training set C_{train} to create a support set and a query set. In the N-way K-shot setting, the support set is comprised of K examples from each of the N classes, represented as $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{N \times K}, y_{N \times K})\}$. Conversely, the query set $Q = \{(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_T^*, y_T^*)\}$ consists of distinct examples from the same N classes. In each episode, the labelled training set, denoted as support set S, is utilised to train the model to minimize the loss of its predictions for the query set Q. The present methodology emulates the process of training classifiers for C_{test}, proceeding through each episode iteratively until reaching a state of convergence.

2.2 Our Framework

The framework we have proposed is based on the Transductive Propagation Network (TPN) [6]. As Figure 1 shows, it comprises four key components, namely: feature embedding that incorporates new channel attention, graph construction that utilises prototype with relation attention, label propagation that disseminates labels from the support set S to the query set Q, and a loss generation step that calculates a cross-entropy loss between the propagated labels and the ground-truths on Q. This approach enables the joint training of almost all parameters in the framework. The feature embedding component is designed to incorporate new channel attention. This mechanism is added after the first CNN layer and proposes a metric to weigh the importance of each channel's information. The aim is to assign different weights to each channel to generate feature embeddings with channel importance information. The graph construction component utilizes a prototype with relation attention to construct a graph that captures the relationships between samples. This is done by creating prototype nodes that capture inter-sample information within a category. The label propagation component disseminates labels from the support set S to the query set Q. This is done by propagating the labels through the constructed graph, utilizing the inter-sample information captured in the prototype nodes. The loss generation step calculates a cross-entropy loss between the propagated labels and the ground-truths on Q. This loss is used to update the parameters of the network during training. The two proposed methods are indicated by the two red blocks in Figure 1. The first proposed method is a new channel attention mechanism added after the first CNN layer. This mechanism proposes a metric to weigh the importance of each channel's information and assigns different weights to each channel to generate feature embeddings with channel importance information. This means that channels with more relevant category information are assigned higher weights. The second proposed method introduces a self-attention mechanism after the second CNN layer, inspired by a transformer [12], to construct prototype nodes with inter-sample information within a category to do label propagation subsequently instead of extracting features directly from a small number of samples or simply averaging them as prototype nodes.

Our framework differs significantly from TPN in that it incorporates a relation attention mechanism during the initial stages of feature embedding and graph construction. This mechanism enables the extraction of the most representative features for effective embedding propagation and downstream classification tasks. Further details of the two proposed methods are discussed in Section 2.2.1 and Section 2.2.2.

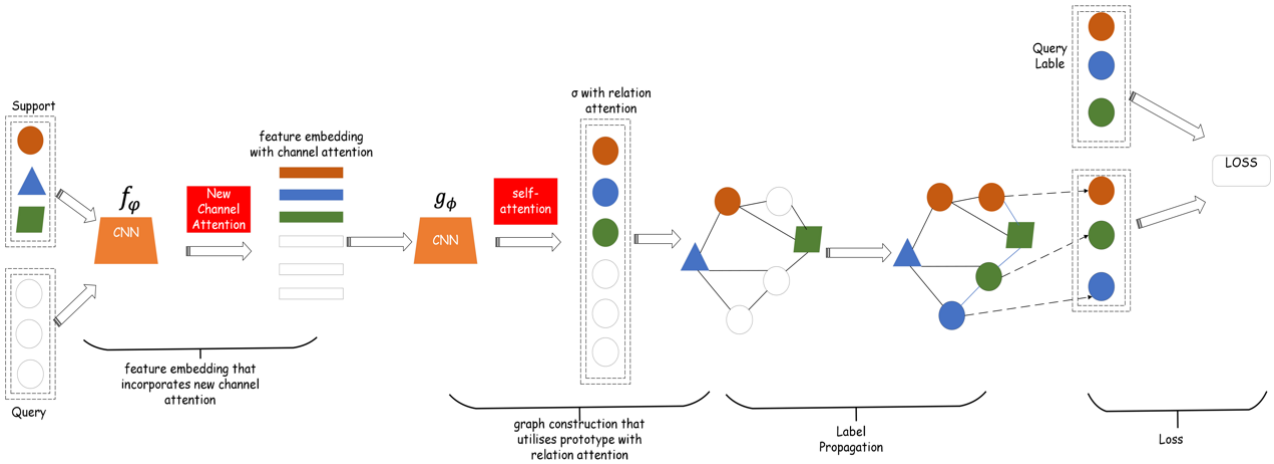

Figure 1. The overall framework of our improved TPN. The red block is our proposed approach.

2.2.1 New Channel Attention in Feature Embedding

We use a convolutional neural network f_φ to extract the features of an input x_i, where $f(x_i;\ \varphi)$ is the feature map and is a network parameter. Notwithstanding its generality, we have chosen to employ the identical architecture that has been utilised in multiple recent works [7, 9, 10]. Through this methodology, it is possible to offer equitable comparisons with experiments, thereby emphasising the impact of our attention method. The neural network consists of four convolutional blocks, with each block commencing with a 2D convolutional layer featuring a 3×3 kernel and a filter size of 64. Subsequent to each convolutional layer, a batch-normalization layer, a ReLU nonlinearity, and a 2×2 max-pooling layer are implemented. The embedding function f_φ is utilized for both the support set S and the query set Q.

In many existing methods, the importance of information among different channels from input data is not considered in feature extraction. However, in classification tasks, if we can determine which channel information is important and assign a relatively large weight to it, then when feature embedding is performed, it will carry feature information that is more representative of the image's category, which is advantageous for subsequent classification tasks.

To determine the feature information importance of different channels from input data, we propose a metric to measure this. Then to measure the variance for assigning weight to different channels, we define the benchmark feature M_i as the average of all channel features of a singer instance:

$$M_i = \frac{1}{C}\sum_{j=1}^{C} f_{i,j} \tag{1}$$

where $f_{i,j}$ denotes the feature embedding of j-th channel of i-th sample. C is the number of channels. Moreover, we first proposed a new but simple scoring formula on the channel information, called $\mathrm{R^{intra}}$ which is defined as follows:

$$R_{i,j}^{intra} = \frac{||f_{i,j} - M_i||^2}{H \times W}. \tag{2}$$

Here, H/W is the height and width of the input sample. $R_{i,j}^{intra}$ denotes the score of j-th channel, in i-th sample. This score R^{intra} indicates the difference between each channel and the baseline feature, and a smaller score means that the information of that channel is more important. Based on this score, it is then converted into weights assigned to each channel by linear transformation. The dimension of $\mathrm{R^{intra}}$ is the same as the number of channels, and after that, this weight is multiplied by each channel to obtain the feature embeddings with the channel's attention. The simple workflow is shown in Figure 2.

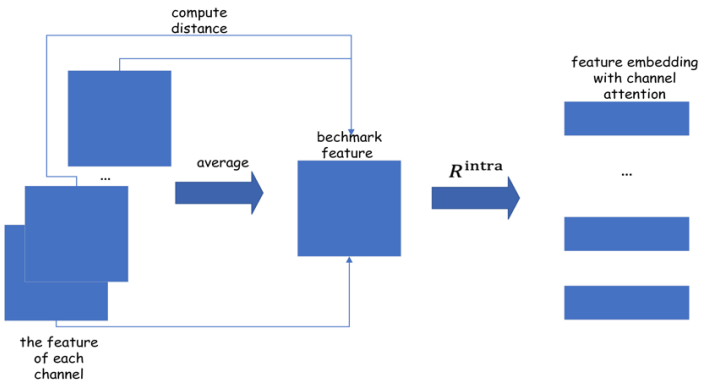

Figure 2. Our proposed New Channel Attention.

2.2.2 Prototype with Relation Attention in graph construction

The current graph nodes are constructed directly from the extracted embedding features of a single sample in each class. However, in few-sample settings, these features may not accurately represent the class, so a prototype approach is employed, wherein the average feature of each class is used as the node of the graph. While the previous prototype

method only averaged the feature embeddings, there exists a relationship between the data within each class in the task. By incorporating the relation attention which consider the information between samples into the prototype, it becomes more robust and representative of the class.

Inspired by transformer's major breakthrough in Natural Language Processing (NLP) field [12], and the self-attention mechanism is its core idea. We try to introduce the self-attention into the construction of protype node in graph construction to get the representative class feature embedding.

We follow the self-attention formula in transformer as proposed by Let Vaswani et.al. [12], First, the feature embedding which is extracted by g_ϕ of samples in each class is linearly transformed, using the transform matrices W^Q, W^K, W^V:

$$Q = x_i \times W^Q, K = x_i \times W^K, V = x_i \times W^V. \tag{3}$$

Second, subsequent to applying the outcomes of Eq (3), the conventional self-attention formula is employed to evaluate the attention weight of each sample. The obtained weight is subsequently multiplied with V (the input after linear transformation) to derive the feature embedding, which signifies the information score pertaining to the inter-sample relationships within the same class. The function defined as follows:

$$P_1 = Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{(d_k)}}\right)V, \tag{4}$$

where d_k is the number of columns of the matrix Q, K, i.e., the vector dimension. The formula Eq (4) computes the inner product between each row vector of matrices Q and K. To maintain gradient stability during the training process, the inner product is divided by the square root of d_k, thereby preventing it from becoming overly large. Then the original average feature embedding is calculated:

$$P_2 = \frac{1}{K}\sum_{K} g_\phi(x_i), \tag{5}$$

$$P = \lambda\overline{P_1} + (1-\lambda)P_2, \tag{6}$$

where K is number of samples in x_i class. Ultimately, the summation of the outputs from the aforementioned equations (Eq (4), Eq (5)) yields the definitive feature embedding for each sample in same category, reflective of the relational attention they possess. $\overline{P_1}$ means the average of P_1, that is the average of samples with self-attention in each category. λ denotes weighting parameter. A simple schematic about our proposed method that generate prototype node with self-attention is shown in Figure 3.

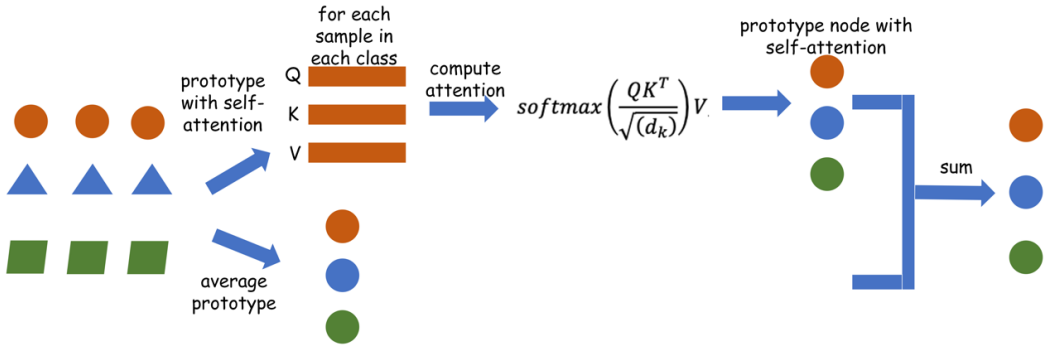

Figure 3. Our proposed prototype node with self-attention.

3. Experiments and Results

We evaluate and compare our method with state-of-the-art approaches on four datasets, namely miniImageNet [2], tieredImageNet [13], CIFARFS-FS [14], and CUB [15]. The former is the most prominent few-shot learning benchmark, while the latter is a much larger few-shot learning dataset that was recently released. The details of the datasets are as follows.

3.1 Experimental Setup

In accordance with Snell et al.'s [9] methodology, we have implemented the episodic training approach. Specifically, we have generated a collection of *N*-way *K*-shot training tasks in order to simulate the *N*-way *K*-shot test scenarios. Snell et al. introduced a training strategy called the "Higher Way", which involved incorporating a greater number of training classes in each episode compared to test cases. Nevertheless, it has been observed that training with a larger number of examples than during the testing phase yields advantageous results. The term used to refer to this phenomenon in our experimental study is "Higher Shot". In order to address the 1-shot and 5-shot test problem, we have implemented a training approach that involves utilizing 5-shot training for the former and 10-shot training for the latter. The query number is uniformly set to 15 across all settings, and the resulting performance metrics are computed as the average over 600 randomly generated episodes drawn from the test set.

To ensure equitable comparison with alternative approaches, we have elected to utilize a commonly employed Convolutional Neural Network (CNN) as the feature embedding function denoted f_{φ}. The value of the hyper-parameter k in the k-nearest neighbour graph, has been established as 20. Additionally, the value of α for label propagation has been determined to be 0.99, in accordance with the recommendations put forth by Zhou et al. [16].

We trained all models utilized in our work using the Adam optimization algorithm, as proposed by Kingma et al. [17], with an initial learning rate of 0.001. In the case of miniImageNet, the learning rate was decreased by half every 10,000 iterations, while for tieredImageNet, the learning rate was reduced by half every 25,000 iterations. The rationale behind the employment of the larger decay step pertains to the fact that tieredImageNet encompasses a greater number of classes, each of which contains a larger number of examples, thereby necessitating more training iterations. The training procedure is repeated until the verification error rate reaches a plateau.

3.2 Main Results

We compared our method with several state-of-the-art methods in various settings. The experimental results are shown in Table 1. The transductive methods tend to perform better than the pure inductive and supervised methods. Our proposed CA-TPN with a backbone containing a 4-layer convolutional network exhibits better performance than other state-of-the-art methods on various benchmark datasets. Importantly, when comparing the original TPN, we achieved significant improvements of 1.9% and 1.8% in both the 1-shot setting and the 5-shot setting for each dataset. It can be observed that on both datasets, miniImageNet, and tieredImageNet, our proposed method performs significantly better than the original TPN for both 1-shot and 5-shot settings. These results strongly demonstrate the effectiveness of our proposed method in accurately extracting representative features rich in attentional information. The reason behind this may be because, for classification tasks, it is particularly important to extract representative feature information for each category, while the original TPN algorithm does not specifically consider feature extraction regarding category information, yet for few-shot classification, the number of samples for each class is small, and if only extracted directly, it does not represent important feature information for that category. Our algorithm ameliorates this shortcoming by considering important information about the representative class at the feature extraction stage, thus enabling better generalization to invisible new classes and making correct predictions when doing the final classification task.

Table 1. Few-shot classification accuracies compared with other state-of-art methods

Methods	Backbone	miniImageNet		tieredImageNet		CIFRAS-FS		CUB	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet [7]	4CONV	43.56	55.31	-	-	-	-	-	-
MAML [8]	4CONV	48.70	63.11	51.67	70.30	58.90	71.50	54.73	75.75
ProtoNet [9]	4CONV	49.42	68.20	53.31	72.69	55.50	72.00	50.46	76.39
Transductive methods									
TPN [6]	4CONV	55.51	67.46	59.91	71.01	-	-	-	-
CA-TPN (ours)	**4CONV**	**56.09**	**68.79**	**61.57**	**72.15**	**63.64**	**77.83**	**71.09**	**76.57**

Compared with other methods except TPN, our proposed CA-TPN method largely outperforms other methods, achieving accuracy improvements of 20.6% and 6.9% for the 1-shot and 5-shot conditions, respectively. This illustrates the superiority of the transductive method in the few-shot classification task, relative to previous popular algorithms. Because the transductive method here constructs an undirected graph with all unlabeled data and labeled data together, the labels of all unlabeled data are obtained by means of label propagation. It also solves the problem that the unknown categories have only a very small amount of data (1 or 5 training samples per category), which leads to unreliable learning of the classifier. Thus, it is clearly superior to other advanced algorithms.

4. Conclusion and Future Work

In this paper, we have proposed a simple method for few-shot classification to consider the sensitive information between channels from input image data, as well as, incorporate a self-attention mechanism to build prototype nodes that can better represent the features of a class in graph construction which is used to do label propagation. The described method has the capability to extract the most relevant and significant features from a given dataset, specifically those that are associated with each class, to facilitate accurate few-shot classification tasks. The feature embedding stage is a key step that produces the representative feature of each class, specifically for few-shot classification from a few samples. Therefore, our approach is dedicated to extracting the features that are most representative of the class information. Extensive experiments on four datasets demonstrate the effectiveness of our method and yield state-of-the-art results. Also, compared with other existing popular methods, our algorithm yet achieves higher results.

The work that can be done in the future can focus on improving the structure of the graph and determining whether or not there is a method that is superior to label propagation. These are the kinds of topics that can be examined in great detail in the future. In addition, investigating whether learning techniques based on graph neural networks may be utilised in place of label propagation is another something that should be looked at.

References

[1] Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. Acm Computing Surveys (CSUR), 53(3), 1-34.

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[3] Chen, Z. M., Wei, X. S., Wang, P., & Guo, Y. (2019). Multi-label image recognition with graph convolutional networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5177-5186).

[4] Li, Y. (2022, January). Research and application of deep learning in image recognition. In 2022 IEEE 2nd international conference on power, electronics and computer applications (ICPECA) (pp. 994-999). IEEE.

[5] Li, G., Zhang, M., Li, J., Lv, F., & Tong, G. (2021). Efficient densely connected convolutional neural networks. Pattern Recognition, 109, 107610.

[6] Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S. J., & Yang, Y. (2018). Learning to propagate labels: Transductive propagation network for few-shot learning. arXiv preprint arXiv:1805.10002.

[7] Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. Advances in Neural Information Processing Systems, 29.

[8] Finn, C., Abbeel, P., & Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks. In International conference on machine learning (pp. 1126-1135). PMLR.

[9] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems, 30.

[10] Ding, Y., Tian, X., Yin, L., Chen, X., Liu, S., Yang, B., & Zheng, W. (2019, September). Multi-scale relation network for few-shot learning based on meta-learning. In International Conference on Computer Vision Systems (pp. 343-352). Cham: Springer International Publishing.

[11] Fallah, A., Mokhtari, A., & Ozdaglar, A. (2020, June). On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In International Conference on Artificial Intelligence and Statistics (pp. 1082-1092). PMLR.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[13] Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., ... & Zemel, R. S. (2018). Meta-learning for semi-

supervised few-shot classification. arXiv preprint arXiv:1803.00676.

[14] Oreshkin, B., Rodríguez López, P., & Lacoste, A. (2018). Tadam: Task dependent adaptive metric for improved few-shot learning. Advances in Neural Information Processing Systems, 31.

[15] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

[16] ZhuЃ, X., & GhahramaniЃн, Z. (2002). Learning from labeled and unlabeled data with label propagation.

[17] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.