Journal of Korea Multimedia Society Vol. 25, No. 1, January 2022(pp. 52-60)
https://doi.org/10.9717/kmms.2022.25.1.052

Intra-class Local Descriptor-based Prototypical Network for Few-Shot Learning

Xi-Lang Huang[†], Seon Han Choi[††,†††]

ABSTRACT

Few-shot learning is a sub-area of machine learning problems, which aims to classify target images that only contain a few labeled samples for training. As a representative few-shot learning method, the Prototypical network has been received much attention due to its simplicity and promising results. However, the Prototypical network uses the sample mean of samples from the same class as the prototypes of that class, which easily results in learning uncharacteristic features in the low-data scenery. In this study, we propose to use local descriptors (i.e., patches along the channel within feature maps) from the same class to explicitly obtain more representative prototypes for Prototypical Network so that significant intra-class feature information can be maintained and thus improving the classification performance on few-shot learning tasks. Experimental results on various benchmark datasets including mini-ImageNet, CUB-200-2011, and tiered-ImageNet show that the proposed method can learn more discriminative intra-class features by the local descriptors and obtain more generic prototype representations under the few-shot setting.

Key words: Few-Shot Learning, Image Classification, Local Descriptors

1. INTRODUCTION

In recent years, deep learning methods have brought impressive improvement on various computer vision tasks such as object detection [1,2], image classification [3,4], and semantic segmentation [5]. The successes of the deep learning methods can generally be attributed to a large number of available training images and sophisticated network structures. Although the network structures can be easily reconstructed and adjusted according to the target task, it is usually difficult to collect the required number of training images in real-life applications such as eye diseases [6], skin diseases [7], to fully train the network. On the other hand, using the limited labeled images to directly train a supervised model often leads to overfitting issues and unsatisfactory performance.

In contrast to the networks that require such an amount of training images to realize decent performance, humans are capable of learning from the prior knowledge and generalizing to the novel concepts through a few examples. Motivated by this phenomenon, few-shot learning [8-17] has been proposed to imitate the learning behavior of humans under limited training examples. In few-shot learning, a dataset is generally split into a meta-training set (base classes), a meta-test set (novel classes), and a meta-validation set. The learning objective is to help the networks efficiently learn

※ Corresponding Author : Seon Han Choi, Address: (48513) 45 Yongso-ro, Nam-gu, Busan, Korea, TEL : +82-51-629-6240, FAX : +82-51-629-6230, E-mail : shchoi@pknu.ac.kr
Receipt date : Dec. 23, 2021, Revision date : Jan. 7, 2022
Approval date : Jan. 12, 2022
[†] Dept. of Artificial Intelligence Convergence, Pukyong National University (E-mail : huangxl901@pukyong.ac.kr)
[††] Dept. of Artificial Intelligence Convergence, Pukyong National University
[†††] Dept. of Electronic and Electrical Eng., Ewha Womans University
※ This work was supported by the National Research Foundation of Korea (NRF) funded by the Korea Government (Ministry of Science and ICT) under Grant 2019R1G1A1098951.

significant features from the meta-training set and generalize the learned network to the meta-test set for prediction given a few labeled examples each time. To this end, few-shot learning exploits the episodic learning mechanism to mimic the low-data scenery in the meta-test procedure. Specifically, in order to consistence with the few-shot setting in the meta-test procedure, the meta-training/meta-validation set is divided into multiple small tasks, each task is composed of a support set that consists of N classes with K (usually 1 or 5) labeled samples per class and a query set that consists of some unlabeled samples from the same N classes. This setting is usually abbreviated as the N-way K-shot task. In each meta-training episode, the network learns to extract useful information from the support set and perform classification on the query set, and the classification losses are used to help update the network parameters. The meta-validation set is used for selecting the best network parameters when the defined episode number is exhausted. Finally, the meta-test procedure evaluates the learned network on the meta-test set under the N-way K-shot setting.

To effectively achieve the learning process of learning useful features given only a few labeled examples, researchers have been proposed a variety of methods in solving few-shot learning. The few-shot learning methods can be typically categorized into metric-based methods, model-based methods, and optimization-based methods. The metric-based methods [8, 10] aim to utilize a suitable distance metric (e.g., Euclidean distance and cosine distance) to perform nearest neighbor classification in the feature spaces. The model-based methods [11,12] focus on designing the model structure to quickly update the parameters to the corresponding task with a few samples. Optimization-based methods [14,15] exploit the meta-training set to learn a good initialization model and generalize the model to the meta-test set with a few optimization steps.

Among the above-mentioned methods, the metric-based methods have been received tremendous attention due to their simplicity and powerful generalization ability. Unlike the optimization-based methods that often require second-order gradients and the model-based methods that use an external memory buffer to store the prior knowledge, metric-based methods simply consist of a feature embedding network and a classifier that reflects the distance metric. In metric-based methods, the representative works are Prototypical network [8], and Relation network [9], Matching network [10], etc. The Prototypical network provides a straightforward idea on solving the few-shot learning problem. It aims to learn a generalized model to make the query samples as close as possible to their corresponding class prototypes, which are obtained by averaging the support samples from the same class. Despite the promising performance that has been achieved by the Prototypical network, the averaging strategy treats the samples from the same class as equally important while some samples may not be important as others, which easily results in learning uncharacteristic features in such a low-data scenery. To demonstrate this, we exemplify the procedure of getting the prototype in Fig. 1. Given 5 sample images from the same class, the Prototypical network flattens the feature maps of each sample obtained by the embedding network as the feature vector (i.e., embedding feature). The prototype is then can be obtained by taking the sample mean of feature vectors (each multiply by 0.2). However, each sample may contain a different level of valuable information. For samples with more stars, they have more discriminative features about a bird, while the samples with fewer stars may also contain some less important features for bird classification such as the branches, leaves, etc. Thus, simply using the mean vectors as prototypes may be prone to learning redundant features.

To alleviate the above issue of Prototypical net-

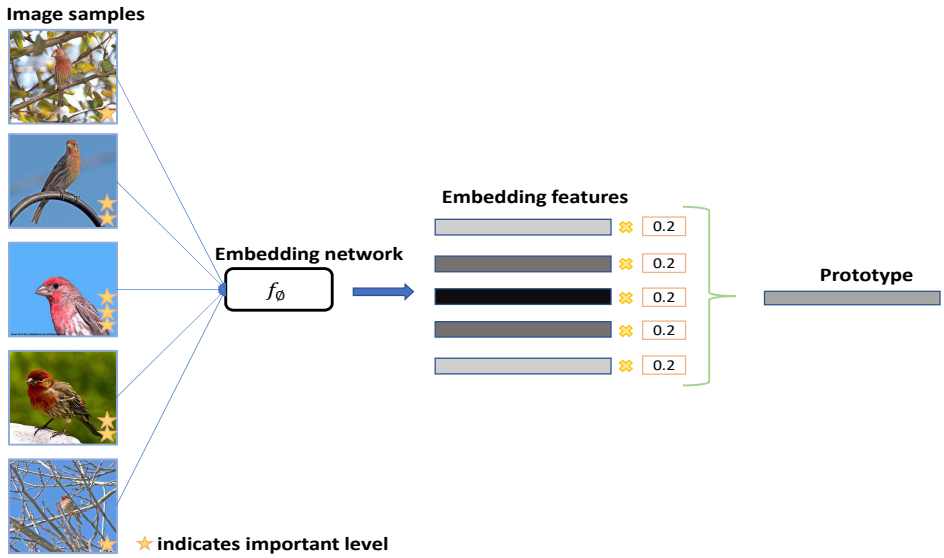

Fig. 1. The illustration of Prototypical network.

work, we suggest using the local descriptors of samples from the same class to emphasize the important features while neglecting the uncharacteristic features such that the prototypes can maintain more valuable information about the class. Concretely, we use the local descriptors of feature maps obtained by the embedding network, each local descriptor is a feature vector along the channel dimension of the feature map. For each sample from the same class, we use Euclidean distance to compute the similarity score between each of its local descriptors and the local descriptors of other samples on the corresponding position. The softmax function is subsequently applied to the sum of the similarity scores of the local descriptors at the corresponding positions between the samples to obtain the feature scoring maps, which aim to emphasize the important spatial features of each sample. The feature scoring maps are later multiplied with the corresponding feature maps along the channel dimension in order to distribute the similarity scores to each local descriptor, and the obtained feature maps are finally flattened as vectors and aggregated as class prototypes through summation. The proposed method is trained in an end-to-end manner and evaluated on various few-shot learning benchmark datasets: mini-ImageNet, tiered-ImageNet, and CUB-200-2011 to evaluate the proposed method. Our method outperforms the Prototypical network in the standard 5-way 1-shot and 5-way 5-shot few-shot learning tasks, showing that the prototype representations obtained by the proposed method contain more generic features of the corresponding class.

The rest of this study is organized as follows: Section 2 introduces the related work of few-shot learning, and Section 3 defines the problem of few-shot learning and related notations. Section 4 describes the proposed method and the experimental results are given in Section 5. The final section is the conclusion.

2. RELATED WORK

Few-shot learning can be broadly categorized into optimization-based, model-based, and metric-based methods. Each group of methods aims to address the few-shot learning from different perspectives. We summarize the main idea of each method and introduce their representative works.

2.1 Metric-based methods

Metric-based methods learn a suitable similarity metric to encourage the network to decrease the intra-class distance and increase the inter-class distance in the feature space. Snell et al. [8] assumed that the prototype representations of each class is the mean vector of samples from its class in the feature space, and utilized the Euclidean distance to perform classification by finding the nearest class prototype for each query sample. Sung et al. [9] parameterized the distance metric into a learnable network, and learned a non-linear distance metric using the relation scores between samples. Vinyals et al. [10] used the cosine distance between support samples and query samples to generate the attention score for classification.

2.2 Model-based method

Model-based methods focus on designing the model structure to quickly adapt the model parameters to the novel samples. Munkhdalai et al. [11] exploited a sophisticated weight update mechanism to achieve fast generalization ability, in which the gradient generated during the training process is used as the generation of fast weights. The model consists of a meta learner and a base learner. The meta learner is used to learn generalized information between training tasks and the memory mechanism is used to store the information. The base learner quickly adapts to the new task and interacts with the meta learner to produce predictive output. Santoro et al. [12] presented a Memory-Augmented Neural Networks (MANN) using an explicit memory buffer to store the class label information and combine it with implicit information from LSTM to speed up parameter updates.

2.3 Optimization-based methods

Optimization-based methods aim to learn a good initialization model from the meta-training set and quickly adapt to the novel meta-test set with a few gradient steps. Finn et al. [13] porposed a Model-Agnostic Meta-learning (MAML) to learn a good parameter initialization from the meta-training set by N-step gradient descent in the 'inner loop' procedure and fine-tunes the parameters in the 'outer loop' procedure to learn task-specific parameters. Nichol et al. [14] can be viewed as a 'simple' version of MAML, it does not require differentiating through the outer loop, making it more suitable and less computationally expensive where many optimization steps are required. Ravi et al. [15] interpreted the stochastic gradient descent as the update rule for the cell state in the LSTM, and proposed training an LSTM-based meta-learner to learn an update rule for training a classifier. Lee et al. [16] embraced the advantage of the support vector machine under low-data scenery to help feature embedding network to learn good feature representations that can generalize well to the novel samples.

3. BACKGROUND

3.1 Problem definition

In this section, we start by introducing some notations and terminologies used in few-shot learning. To train the network that generalizes well to the unseen samples, a dataset is generally split into three sets: a meta-training set D_{train}, a validation-test set D_{val}, and a test set D_{test}, which is respectively used to train the network in a supervised manner, select the best network parameters, and evaluate the generalization ability of the network. Each set contains the disjoint label spaces with each other. In order to mimic the circumstance that only a few labeled samples are available for training, few-shot learning adopts the episodic learning mechanism. Under the episodic learning regime, a support set that includes a few labeled samples and a query set that contains some unlabelled samples are generated in each episode. In

general, the support set is composed of N random classes with K labeled samples per class and the query set contains some unlabelled samples from the same N classes. This setting is called an N-way K-shot task. In the meta-training procedure, the model learns from the support set and generates the loss for parameter update by performing classification on the query set. The meta-validation is usually conducted to select the best network parameters when the total number of the episode is exhausted. In the meta-test procedure, the novel classes from the meta-test set are used to construct the support set and the query set for evaluating the selected network

3.1 Prototypical network

The main idea of the Prototypical network is to construct the prototypes using the sample mean of the support set. Given a support set of labeled samples $S=\{(x_1,y_1),...,(x_{NK},y_{NK})\}$, where $x_i \in R^D$ is an D-dimensional feature vector, and $y_i \in \{1,...,N\}$ is the corresponding label. We denote S_N as the subset of the support set with samples in the Nth class and f_Φ as the embedding network f with parameters Φ. Then, the prototypes for each class can be defined as:

$$C_N = \frac{1}{|S_N|} \sum_{(x_i,y_i) \in S_N} f_\Phi(x_i) \tag{1}$$

where C_N is the prototype (i.e., mean vector) of the corresponding samples that belong to the Nth class.

For classification, the Prototypical network employs the Euclidean distance d on each query sample and class prototypes in the embedding space to produce a probability distribution of classes based on a softmax function over the distance:

$$p_\Phi(y=n|x) = \frac{\exp(-d(f_\Phi(x), C_n))}{\sum_N \exp(-d(f_\Phi(x), C_N))} \tag{2}$$

The training process is to minimize the negative log-probability loss $J(\Phi) = -\log p_\Phi(y=n|x)$ of the ground truth label n via the stochastic gradient descent.

The Prototypical network provides a simple yet competitive method in solving the few-shot learning. By repeatedly training the network with the similarity scores between the support set and query set, the network learns to decrease the intra-class distance and cluster the samples from the same class.

4. METHODOLOGY

Due to the low-data scenery in the training and test phase, the mean vectors of samples in the support set are prone to learning redundant features without considering the priority of valuable features and resulting in degrading the semantic representation of prototypes. Here, we argue that different samples in the same class may differ greatly in their feature representations. To this end, we proposed a intra-class local descriptor-based prototypical network to emphasize the valuable features while neglecting the uncharacteristic features. Our method utilizes the similarity scores between local descriptors of intra-class samples on the same spatial position, which aims to emphasize the common feature on that position. The softmax function is later applied to these similarity scores to generate attention values along the channel dimension to determine how much information should be remain. The final prototype of the proposed method is subsequently obtained by summing the vectors flattened from the feature maps. In essence, the Prototypical network can be regarded as a special situation of our proposed method. In other words, the Prototype network can be regarded as using the mean operation to generate equal attention values for each local descriptor, regardless of the priority of each local descriptor.

The use of local descriptors is to explore the intra-class feature, we exemplify the proposed method with 3 labeled samples from the same class, which is shown in Fig. 1. The learnable em-

bedding network first extracts the feature map for each given support sample, whose size is $C \times H \times W$, in which C refers to the number of channels, H and W refer to its height and width. Using the feature map, we calculate the sum of the similarity between the local descriptor of each sample and other local descriptors in the same location. As shown in Fig. 1, the similarity sum of the local descriptor L_1 is computed from the similarity of L_1 itself and its similarity to L_2 and L_3. The softmax function is used to get the weight of each local descriptor on the same position, where the sum of the total weights is 1. After calculating the weights of local descriptors for each sample, we multiply the descriptors by their corresponding weights along the channel dimension to determine the amount of information that each descriptor should retain. Subsequently, we flatten the weighted feature maps into vectors and perform the element-wise sum on the vectors from the same class to get the prototype. The final classification is then performed by calculating the similarity between prototypes and query samples via the Euclidean distance.

5. EXPERIMENTS

5.1 Dataset

we evaluate the proposed method on three widely used benchmark datasets: miniI-mageNet [11], tiered-ImageNet [17], and CUB-200-2011 [18]. The mini-ImageNet is a subset of ImageNet for few-shot learning tasks. The dataset contains 60,000 images from 100 classes. We follow the class split setting used by Ravi and Larochelle [15], in which 64 classes are used for meta-training, 16 classes for meta-validation, and 20 classes for meta-test. Each class includes 600 images with a size of 84×84. The tiered-ImageNet dataset is another subset derived from ImageNet but contains a much larger number of classes compare to the mini-ImageNet. It includes 608 classes in total, in which 351 classes are used as meta-training, 97 classes for meta-validation, and 160 classes for meta-test. Each class contains approximately 1,300 images and the images are resized to 84×84. The CUB-200-2011 is the abbreviation of Caltech-UCSD birds-200-2011, which contains 11,788 images from 200 fine-grained bird classes. We follow the previous work [19] to randomly select 100 classes for meta-training, 50 classes for meta-validation, and

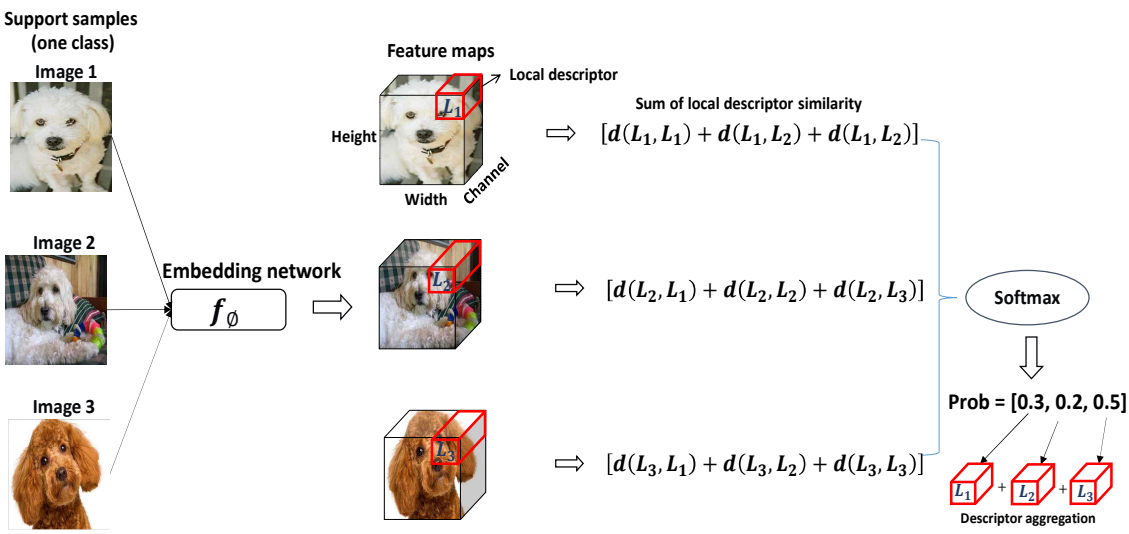

Fig. 2. Demonstration of local descriptor-based Prototypical network.

50 classes for meta-test. During meta-training, we follow the previous works [8,9,16] to apply the standard data augmented setting to samples from the support set including random crop, random horizontal flip.

5.2 Implementation details

We conduct experiments on the widely used 4-layer ConvNet (Conv-4). The Conv-4 consists of 4 consecutive 64-channel convolution blocks with each block including a convolutional layer, a BatchNorm layer, a leakyReLu layer, and a max-pooling layer. Then, given the input images of size 84×84, the output shape of feature maps for Conv-4 is thus $64\times5\times5$.

We perform the standard 5-way 5-shot (5W5S) and 5-way 1-shot (5W1S) classification task for the proposed method. The network is meta-trained for 200 epochs, each epoch contains 1,000 episodes. We use Adam as the optimizer with the default initial learning rate (i.e., 0.001), and the learning rate decay to its half at every 50 epochs. During the meta-training, the number of query samples is set to 15 per class for loss generation. We reported the testing results with a 95% confidence interval computed over 600 test episodes.

5.2 Main results

From Table 1, the proposed method only achieves a little improvement or similar performance under the 5-way 1-shot task on each dataset compare to the Prototypical network. The underlying reason for this phenomenon is that the proposed method is identical to the Prototypical network under the 5-way 1-shot setting. The proposed method cannot compute the similarity of local descriptors when there is only 1 sample for each class, while the Prototypical network directly uses that single sample as the class prototype. On the other hand, the Relation network shows strong generalization ability in the 5-way 1-shot setting, which can be attributed to the learnable distance metric effectively learning the non-linear metric when there is only one sample per class. In the 5-way 5-shot setting, the proposed method outperforms the Relation network, Matching network, and Prototypical network on three benchmark datasets, which demonstrates that the interaction between local descriptors from the same class can bring more valuable information to the prototypes, and thus making the prototypes more generalize to its class and discriminative to other classes.

6. CONCLUSION

In this study, we proposed the local-descriptor-based prototypical network that uses the intra-class local descriptors from the same spatial position to improve the semantic representation of prototypes. The existing Prototypical network simply employs mean vectors as the prototypes under the few-shot setting, which easily results in learning uncharacteristic features and degrades the

Table 1. Few-shot classification accuracies (%) on mini-ImageNet, tiered-ImageNet, and CUB-200-2011 using Conv-4 backbone.

Dataset	Method							
	Relation Network		Matching Network		Prototypical Network		Proposed	
	5W1S	5W5S	5W1S	5W5S	5W1S	5W5S	5W1S	5W5S
mini-ImageNet[ξ]	50.44±0.82	65.32±0.70	43.56±0.84	55.31±0.73	49.42±0.78	68.20±0.66	49.68±0.89	69.37±0.64
tiered-ImageNet	54.48±0.93	71.32±0.78	49.56±0.92	68.76±0.71	53.31±0.89	72.69±0.74	52.89±0.91	73.26±0.68
CUB-200-2011[ξ]	62.34±0.94	77.84±0.68	60.52±0.88	75.29±0.75	50.46±0.88	76.39±0.64	52.87±0.95	78.26±0.61

[ξ] indicates the results are reported by [19]. The best results of 5W1S and 5W5S are marked in bold. All the results are mean accuracies over 600 test episodes with 95% confidence intervals.

generalization of prototypes. To alleviate this issue, the proposed method utilizes the relationship between local descriptors to emphasize the valuable information while neglecting the redundant ones. To evaluate the effectiveness of the proposed method, we conducted experiments on the widely used few-shot learning benchmark datasets. The experimental results on the 5-way 5-shot task show that the proposed method can learn more discriminative prototypes compare to the Prototypical network. However, since each class has only 1 sample, the proposed method is not feasible enough to solve the 5-way 1-shot problem. Future work will focus on considering inter-class information to solve this problem.

REFERENCE

[1] N. Zhang, Y. Feng, and E.J. Lee, "Activity Object Detection Based on Improved Faster R-CNN," *Journal of Korea Multimedia Society*, Vol. 24, No. 3, pp. 416-422, 2021.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *European Conference on Computer Vision*, pp. 213-229, 2020.

[3] X. Yang, Y. Ye, X. Li, R.Y. Lau, X. Zhang, and X. Huang, "Hyperspectral Image Classification with Deep Learning Models," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 56, No. 9, pp. 5408-5423, 2018.

[4] S.W. Park and D.Y. Kim, "Comparison of Image Classification Performance in Convolutional Neural Network According to Transfer Learning," *Journal of Korea Multimedia Society*, Vol. 21 No. 12, pp. 1387-1395, 2018.

[5] F. Jiang, A. Grigorev, S. Rho, Z. Tian, Y. Fu, W. Jifara, et al., "Medical Image Semantic Segmentation Based on Deep Learning," *Neural Computing and Applications*, Vol. 29, No. 5, pp. 1257-1265, 2018.

[6] X.L. Huang, C.Z. Kim, and S.H. Choi, "An Automatic Strabismus Screening Method with Corneal Light Reflex Based on Image Processing," *Journal of Korea Multimedia Society*, Vol. 24, No. 5, pp. 642-650, 2021.

[7] K. Mahajan, M. Sharma, and L. Vig, "Meta-DermDiagnosis: Few-Shot Skin Disease Identification Using Meta-Learning," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 730-731. 2020.

[8] J. Snell, K. Swersky, and R.S. Zemel, "Proto-Typical Networks for Few-Shot Learning," *arXiv Preprint,* arXiv:1703.05175, 2017.

[9] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, and T.M. Hospedales, "Learning to Compare: Relation Network for Few-Shot Learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199-1208, 2018.

[10] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching Networks for One Shot Learning," *Advances in Neural Information Processing Systems*, pp. 3630-3638, 2016.

[11] T. Munkhdalai and H. Yu, "Meta Networks," *International Conference on Machine Learning,* pp. 2554-2563. 2017.

[12] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-Learning with Memory-Augmented Neural Networks," *International Conference on Machine Learning*, pp. 1842-1850, 2016.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *International Conference on Machine Learning,* Vol. 70, pp. 1126-1135, 2017.

[14] A. Nichol, and J. Schulman, "Reptile: A Scalable Metalearning Algorithm," a*rXiv Preprint,* arXiv:1803.02999, 2018.

[15] S. Ravi and H. Larochelle, "Optimization as a Model for Few-Shot Learning," *International*

Conference on Learning Representations, 2017.

[16] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-Learning with Differentiable Convex Optimization," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10657-10665, 2019.

[17] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J.B. Tenenbaum, et al., "Meta-Learning for Semi-Supervised Few-Shot Classification," *arXiv Preprint,* arXiv:1803.00676, 2018.

[18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, *The Caltech-UCSD Birds-200-2011 Dataset*, *Technical Report*, 2011.

[19] W.Y. Chen, Y.C. Liu, Z. Kira, Y.C.F. Wang, and J.B. Huang, "A Closer Look at Few-Shot Classification," *arXiv Preprint,* arXiv:1904.04232, 2019.

Xi-Lang Huang

He received the M.S. degree in electrical engineering from the Pusan National University, Busan, South Korea, in 2018. He is currently pursuing the Ph.D. degree in electrical engineering with the Pukyong National University, Busan, South Korea. His current research interests include modeling and simulation of discrete-event systems, efficient simulation optimization, and computer vision.

Seon Han Choi

He received the B.S., M.S., and Ph.D. degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012, 2014, and 2018, respectively. In 2018, he was a Post-Doctoral Researcher with the Information and Electronics Research Institute, KAIST. From 2018 to 2019, he was a Senior Researcher with the Korea Institute of Industrial Technology. In 2019, he joined the Department of IT Convergence and Application Engineering, Pukyong National University, Busan, South Korea, as an Assistant Professor. His current research interests include the modeling and simulation of discrete-event systems, efficient simulation optimization under stochastic noise, evolutionary computing, and machine learning.