2018 IEEE 30th International Conference on Tools with Artificial Intelligence

Boosting Few-shot Image Recognition via Domain Alignment Prototypical Networks

Jiang Lu[†,‡], Zhong Cao[†], Kailun Wu[†], Gang Zhang[‡], Changshui Zhang[†]
[†]*Institute for Artificial Intelligence, Tsinghua University (THUAI)*
Beijing National Research Center for Information Science and Technologies (BNRist)
State Key Lab of Intelligence Technologies and Systems, Department of Automation, Tsinghua University, Beijing, P.R.China
[‡]*China Marine Development and Research Center (CMDRC), Beijing, China*
E-mail: {lu-j13, caozhong14, wukl14}@mails.tsinghua.edu.cn; gangzhang2008@163.com; zcs@mail.tsinghua.edu.cn

***Abstract*—Human has the ability of drawing inferences about other things from only one instance. Few-shot learning is aimed at imitating this generalized learning behavior of human beings, where the learning machine is expected to recognize novel categories not seen in the training set, given only a few training data for each novel category. In this paper, we enhance the Prototypical Network for few-shot learning tasks by introducing a domain alignment module, which takes into account the domain shifts existing between different categories. Compared to original Prototypical Network (PN), the most excellent model for few-shot learning at present, our proposed Domain Alignment Prototypical Network (DA-PN) is able to abate the distribution differences among the data of training and test classes, further optimizing the embedding space of prototype feature for each category and then boosting few-shot recognition. Comprehensive empirical evidence demonstrates that the proposed DA-PN can yield state-of-the-art few-shot recognition performance on the public benchmark dataset mini-ImageNet as well as a novel proposed few-shot dataset MNIST&CIFAR10.**

***Index Terms*—Few-shot Learning, Image Recognition, Prototypical Networks, Domain Shifts, Distribution Alignment**

I. Introduction

The recent success in deep learning seriously hinges on the large-scale training data and time-consuming human labeling efforts [1]. Despite the impressive results achieved by deep neural networks (DNN) on many machine visual tasks, such as image classification [2]–[4], object detection [5]–[7], and image caption [8], [9], [11], one setting presenting an intractable challenge is that of few-shot learning, as shown in Fig. 1, where for new target categories only a few labeled examples are available. As we all known, human beings are able to draw inferences about novel instances from a few scraps of information, *e.g.* inferring the full range of appearance for "panda", given only one or two images about "panda". In this context, few-shot learning has been developed to imitate the generalized learning ability of human beings [12]–[15].

Previous works for few-shot learning can be broadly divided into two major categories. Some works advised to tackle this problem from the perspective of meta-learning [16]–[18], which involves a meta-learner that tries to quickly learn a classifier on novel classes given a few training data for this new task [19]–[21]. In a different line of work, other efforts for few-shot learning can be concluded into metric-learning based methods, which attempt to learn a good feature representation

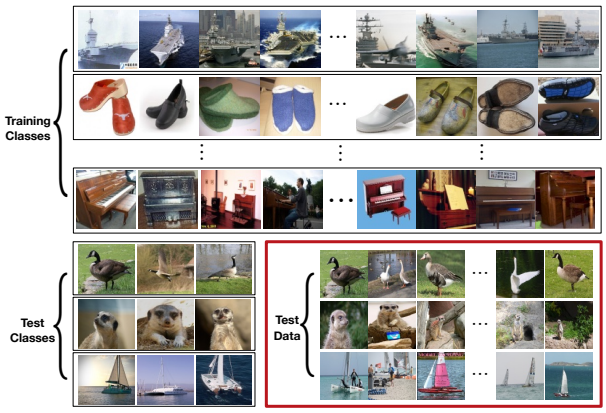

Fig. 1. Example of few-shot image recognition setup. The top represents the training classes each of which has abundant training samples. The bottom represents 3 test classes not seen in training set, each of which only has 3 training samples. Inside the red box are test samples to be classified into one test class. In this illustration, the above few-shot recognition task is also known as 3-way 3-shot classification task.

space clustering the features of same objects but separating the features of different objects [15], [22]–[24]. Among these metric-learning approaches, Prototypical Networks (PN) [23] maintains the state-of-the-art performance on few-shot image recognition task, which is based on the idea of learning to classify test samples by measuring their distances to the prototype feature of the novel classes. Specially, the prototype feature of one novel class can be computed as the cluster center of feature vectors of the few available training samples for that class embedded by a learnable DNN model. Despite the superiority and simplicity, PN ignores an intuitional but unavoidable fact that there exist domain shifts between training and test classes since the test data to be classified belong to the novel test classes never appeared in training classes. Consequently, we believe that the PN, merely learned on the domain of training classes and then adapted to accommodate new classes without any adjustment, may be insufficient to fit for the domain of test classes.

This paper proposes a Domain Alignment Prototypical Networks (DA-PN) to cope with few-shot recognition tasks, which takes better account of the domain shifts between training and test classes. Despite the homogeneous image modality

2375-0197/18/$31.00 ©2018 IEEE
DOI 10.1109/ICTAI.2018.00048

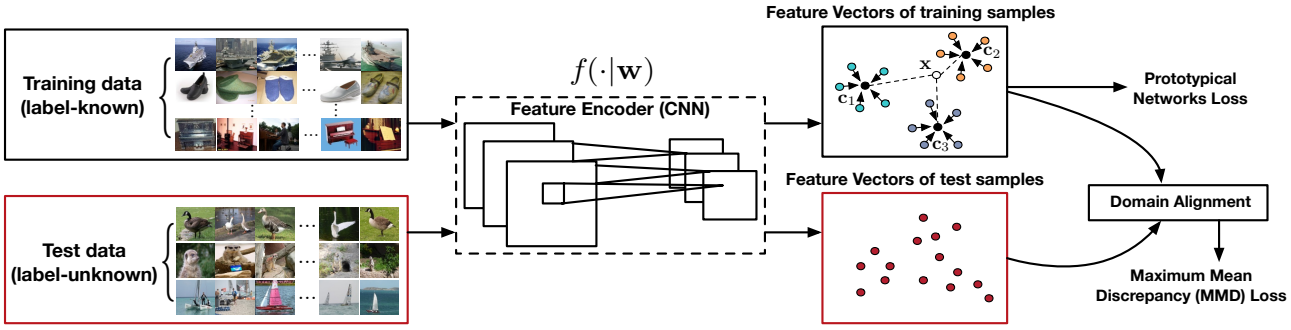

Fig. 2. Domain Alignment Prototypical Networks (DA-PN) for few-shot learning. Except for the original operations of Prototypical Networks (PN), in each "training episode", DA-PN embeds some test data, randomly sampled from test set, into feature space by current learned feature encoder $f(\cdot|\mathbf{w})$, and then combines the MMD loss calculated by domain alignment module and the classification loss computed by PN so as to achieve end-to-end training for $f(\cdot|\mathbf{w})$.

from which the training data and test data are both, there is an apparent dataset shift because of the different categories. In view of this, we design a domain alignment module to bridge the distribution differences existing among the data in training and test classes, which involves the minimization of Maximum Mean Discrepancy (MMD) [25] between training dataset and test dataset in feature space. Similar to PN [23] or other related methods [20]–[22], [24], we also leverage the *episode* training strategy to optimize our DA-PN model (Note that the episode is designed to mimic the few-shot task by subsampling classes as well as data points from training data), where the main difference is that we devise an extra MMD loss in each "training episode", which is calculated by domain alignment module upon the extracted features of training data and test data randomly sampled from test set, as shown in Fig. 2. Extensive experiments demonstrate that DA-PN can yield state-of-the-art few-shot classification performance on the public benchmark dataset mini-ImageNet as well as a novel few-shot dataset MNIST&CIFAR10 proposed by us.

II. Related Work

A. Few-shot learning

Recently, few-shot learning is a hot topic so that vast approaches have been developed to deal with this problem. Due to the space limitations, here we briefly introduce the most relevant metric-learning based approaches to our work. Siamese Neural Network [15] learned how to evaluate the pairwise similarity between test data and the few labeled data and formulated the few-shot recognition as a similarity-based image matching task. Matching Network [22] leveraged an attention mechanism over the learned features of the few labeled data to construct a differentiable nearest neighbor classifier. Based on the idea that there exists a prototype feature for each class under an ideal embedding, around which other feature points of the same class will cluster, PN [23] proposed to learn the prototype feature for one novel class as the cluster center of feature vectors of the few labeled data from that class, whose capability has also been proved to be state-of-the-art by the experimental comparison on benchmark dataset mini-ImageNet. However, its generalization ability hinges on the degree of domain difference existing between training classes and novel (or test) classes. That means it can be beneficial to bridge training domain and test domain.

B. Domain Alignment

Domain shifts is a stubborn trouble existing widely in information retrieval and computer vision tasks, which is highlighted when test samples are drawn from a different data distribution with training data. To address this problem, Domain Adaptation or Domain Alignment, which focuses on learning classification models on target data by exploiting extra knowledge distilled from a related source task, has recently seen a surge of activity [26]–[30], among which the regularization term of minimizing the Maximum Mean Discrepancy (MMD) [25], [26], [29] or maximizing a domain confusion [27], [28] has been added into the objective function so as to align source and target feature representations. Instructively, we exploit a domain alignment module in our DA-PN model under few-shot setup by treating the training classes and novel classes as different domains.

III. Method

A. Notation

If the goal is to separate n novel classes for many test samples from these n classes using only k labeled support samples for each class, we usually call it an n-way k-shot classification task. We denote by D_{trn} the training set including N training classes, D_{tst} the test set. The N_s classes, a subset of that N classes, is randomly selected to form a "training episode" to mimic the ultima few-shot task. In each "training episode", a support set $S = \{(\mathbf{x}_1^s, y_1^s), \ldots, (\mathbf{x}_n^s, y_n^s)\}$ as well as a query set $Q = \{(\mathbf{x}_{n+1}^s, y_{n+1}^s), \ldots, (\mathbf{x}_{n+m}^s, y_{n+m}^s)\}$ are constructed by randomly sampling data from that N_s classes, where $\mathbf{x}_i^s$ are the image inputs and y_i^s are corresponding labels. We denote by S_c the support data set whose elements all belong to class c, where $c \in \{1, \ldots, N_s\}$. Let K_s and K_q denote the numbers of support data and query data per class respectively, then we have $n = N_s K_s$ and $m = N_s K_q$. In addition, the $f(\cdot|\mathbf{w})$ parameterized by deep weights $\mathbf{w}$ is used to denote the embedding function implemented by a learnable CNN taking images $\mathbf{x}_i$ as inputs and outputting its feature vector.

Algorithm 1 Episode loss computation of DA-PN.

Input: D_{trn} including N training classes, D_{tst}, K_s, K_q, N_s (described in Sec. III-A), l, hyperparameters α and β.
$\{1,\ldots,N_s\} \leftarrow \mathrm{RandomSample}(\{1,\ldots,N\})$. ▷ *Randomly sample N_s support classes from all N training classes.*
Build $S = \{(\mathbf{x}_1^s, y_1^s), \ldots, (\mathbf{x}_n^s, y_n^s)\}$. ▷ *Randomly sampling n support data from D_{trn} about the N_s classes.*
Build $Q = \{(\mathbf{x}_{n+1}^s, y_{n+1}^s), \ldots, (\mathbf{x}_{n+m}^s, y_{n+m}^s)\}$. ▷ *Randomly sampling m query data from D_{trn} about the N_s classes.*
Build $T = \{(\mathbf{x}_1^t, y_1^t), \ldots, (\mathbf{x}_l^t, y_l^t)\}$. ▷ *Randomly sampling l test data from D_{tst}.*
$\mathbf{p}_c \leftarrow \frac{1}{|S_c|} \sum_{(\mathbf{x}_i^s, y_i^s) \in S_c} f(\mathbf{x}_i^s|\mathbf{w})$. ▷ *Calculate the prototype feature for each support class $c \in 1,\ldots,N_s$.*
$\mathcal{L} = \frac{1}{|Q|} \sum_{(\mathbf{x}_j^s, y_j^s) \in Q} -\log p(y^s = y_j^s|\mathbf{x}_j^s, \mathbf{w})$. ▷ *Calculate the negative log-probability loss $\mathcal{L}$.*
Compute $\mathcal{D}$ following Eq. 5 where $\mathbf{x}_i^s \in S \cup Q$, $\mathbf{x}_i^t \in T$. ▷ *Calculate the MMD distance $\mathcal{D}$.*
Output: Unified episode loss $\mathcal{J} = \mathcal{L} + \alpha\mathcal{D}$.

B. Prototypical Networks

Prototypical Networks propose that there exists an ideal non-linear embedding function $f(\cdot|\mathbf{w})$ which clusters the feature vectors of data in same class close to its corresponding prototype representation in feature space. The prototype feature vector $\mathbf{p}_c$ of class c can be derived as follow:

$$\mathbf{p}_c = \frac{1}{|S_c|} \sum_{(\mathbf{x}_i^s, y_i^s) \in S_c} f(\mathbf{x}_i^s|\mathbf{w}). \tag{1}$$

For a new query sample $\mathbf{x}^s$, it probability distribution over all N_s support classes can be formulated by a softmax function over negative distances to these N_s prototype features:

$$p(y^s = c|\mathbf{x}^s, \mathbf{w}) = \frac{\exp(-d(f(\mathbf{x}^s|\mathbf{w}), \mathbf{p}_c))}{\sum_{c' \in \{1,\ldots,N_s\}} \exp(-d(f(\mathbf{x}^s|\mathbf{w}), \mathbf{p}_{c'}))}, \tag{2}$$

where $d(\cdot,\cdot)$ is actualized by the metric of squared Euclidean distance. Furthermore, the minimization for the negative log-probability on whole query set Q is conducted as the network objective loss function of PN:

$$\mathcal{L} = \frac{1}{|Q|} \sum_{(\mathbf{x}_j^s, y_j^s) \in Q} -\log p(y^s = y_j^s|\mathbf{x}_j^s, \mathbf{w}). \tag{3}$$

The above loss function can be optimized via an end-to-end gradient-based back-propagation algorithm.

C. Domain Alignment Module

Since the novel classes under few-shot recognition setup have never appeared in training set, there are intuitional distribution shifts between training data and test data. Therefore, we exploit a domain alignment module to reduce the domain shifts by minimizing the Maximum Mean Discrepancy (MMD) [10] between the training classes and novel classes in the learned feature space. In fact, MMD is a nonparametric distance metric to evaluate the similarity between two different distributions in Reproducing Kernel Hilbert Spaces $\mathcal{H}$ (RKHS) with feature mapping function $f(\cdot|\mathbf{w})$ and kernel function $k(\cdot,\cdot)$ (detailed deduction can be found in [10]), which is formalized as:

$$\mathcal{D} = \left\| \mathrm{E}_{\mathbf{x}^s \sim P_s}(f(\mathbf{x}^s|\mathbf{w})) - \mathrm{E}_{\mathbf{x}^t \sim P_t}(f(\mathbf{x}^t|\mathbf{w})) \right\|_{\mathcal{H}}^2, \tag{4}$$

where P_s and P_t denote the distributions of source domain (*i.e.* data from training classes) and target domain (*i.e.* data from test classes), respectively. An empirical estimate of the MMD in Eq. 4 can be obtained by replacing the population expectations with empirical expectations as follows:

$$\begin{aligned} \mathcal{D} = &\sum_{i=1}^{n+m} \sum_{j=1}^{n+m} \frac{k(f(\mathbf{x}_i^s|\mathbf{w}), f(\mathbf{x}_j^s|\mathbf{w}))}{(n+m)^2} \\ &+ \sum_{i=1}^{l} \sum_{j=1}^{l} \frac{k(f(\mathbf{x}_i^t|\mathbf{w}), f(\mathbf{x}_j^t|\mathbf{w}))}{l^2} \\ &- 2 \sum_{i=1}^{n+m} \sum_{j=1}^{l} \frac{k(f(\mathbf{x}_i^s|\mathbf{w}), f(\mathbf{x}_j^t|\mathbf{w}))}{l(m+n)}, \end{aligned} \tag{5}$$

where $n+m$ and l are the number of training data and test data randomly sampled in each episode (m and n have been defined in Sec. III-A), and $k(\mathbf{x},\mathbf{y}) = \exp(-\beta\|\mathbf{x}-\mathbf{y}\|^2)$ is the Gaussian Kernel.

D. Learning and Inference

To balance the classification performance and the transitive ability of PN, we integrate the network objective loss of original PN and the MMD distance of domain alignment module into an unified optimization problem:

$$\min_{\mathbf{w}} \mathcal{J} = \mathcal{L} + \alpha\mathcal{D}, \tag{6}$$

where α is a trade-off hyperparameter between the negative log-probability loss $\mathcal{L}$ and MMD distance $\mathcal{D}$. The detailed episode loss computation is provided in Algorithm. 1, which can be optimized with gradient-based back-propagation algorithm. As for inference, we classify the test sample $\mathbf{x}^t$ into the class $\arg\min_c d(f(\mathbf{x}^t|\mathbf{w}^*), \mathbf{p}_c)$, where $\mathbf{w}^*$ denotes the well-trained encoder weights and $\mathbf{p}_c$ the prototype feature of test class c computed by the provided few samples for class c.

IV. Experiments

The DA-PN is evaluated on the challenging few-shot benchmark dataset mini-ImageNet [22]. Additionally, in order to better verify the transitive capacity of DA-PN to domain shifts in few-shot tasks, we construct one new few-shot recognition dataset named MNIST&CIFAR10 by combining some images of the two off-the-shelf image datasets MNIST [33] and CIFAR10. Experimental results on above two task datasets show the proposed method can yield performance improvements over recent competitive approaches.

TABLE I
FEW-SHOT RECOGNITION RESULTS ON MINI-IMAGENET.

Methods	5-way Recognition	
	1-shot	5-shot
Baseline Nearest Neighbors*	$41.08 \pm 0.70\%$	$51.04 \pm 0.65\%$
Matching Networks* [22]	$43.40 \pm 0.78\%$	$51.09 \pm 0.71\%$
Matching Networks FCE* [22]	$43.56 \pm 0.84\%$	$55.31 \pm 0.73\%$
Meta-Learner LSTM* [20]	$43.44 \pm 0.77\%$	$60.60 \pm 0.71\%$
MAML [21]	$48.70 \pm 1.84\%$	$63.11 \pm 0.92\%$
mAP-SSVM [24]	$50.32 \pm 0.80\%$	$63.94 \pm 0.72\%$
mAP-DLM [24]	$50.28 \pm 0.80\%$	$63.70 \pm 0.70\%$
PN [23]	$49.42 \pm 0.78\%$	$68.20 \pm 0.66\%$
DA-PN (ours)	$\mathbf{50.56} \pm 0.85\%$	$\mathbf{69.62} \pm 0.76\%$

*Results reported by [20].

TABLE II
FEW-SHOT RECOGNITION RESULTS ON TASK MNIST-CIFAR10.

Methods	5-way Recognition	
	1-shot	5-shot
Baseline Nearest Neighbors	$21.03 \pm 0.93\%$	$25.64 \pm 0.86\%$
PN [23]	$27.92 \pm 0.64\%$	$37.75 \pm 0.89\%$
DA-PN (ours)	$\mathbf{30.26} \pm 0.61\%$	$\mathbf{39.08} \pm 0.73\%$
Methods	**10-way Recognition**	
	1-shot	**5-shot**
Baseline Nearest Neighbors	$12.33 \pm 1.02\%$	$15.68 \pm 0.96\%$
PN [23]	$16.64 \pm 0.63\%$	$23.46 \pm 0.69\%$
DA-PN (ours)	$\mathbf{21.53} \pm 0.87\%$	$\mathbf{28.02} \pm 0.98\%$

TABLE III
FEW-SHOT RECOGNITION RESULTS ON TASK CIFAR10-MNIST.

Methods	5-way Recognition	
	1-shot	5-shot
Baseline Nearest Neighbors	$22.48 \pm 1.29\%$	$23.78 \pm 0.94\%$
PN [23]	$51.57 \pm 0.76\%$	$61.21 \pm 0.62\%$
DA-PN (ours)	$\mathbf{57.97} \pm 0.83\%$	$\mathbf{77.87} \pm 0.66\%$
Methods	**10-way Recognition**	
	1-shot	**5-shot**
Baseline Nearest Neighbors	$10.13 \pm 1.94\%$	$11.92 \pm 1.05\%$
PN [23]	$35.73 \pm 0.68\%$	$65.42 \pm 0.61\%$
DA-PN (ours)	$\mathbf{45.03} \pm 0.88\%$	$\mathbf{72.71} \pm 0.81\%$

mini-ImageNet. mini-ImageNet dataset is a subset of the large scale visual image dataset ILSCRC-12 [31] and originally built in [22]. It consists of 100 different classes with 600 RGB images per class, each of size 84×84. In our experiments we follow the splits proposed by [20] for fair comparison with other few-shot learning algorithms, which includes 64 classes for training, 16 classes for validation and 20 classes for testing. The evaluation setup on mini-ImageNet is to learn a few-shot recognition model on training classes by monitoring its generalization performance on validation classes and then using the learned model to play classification for test classes based on the provided few data for each test class.

MNIST&CIFAR10. MNIST&CIFAR10 dataset is formed by randomly selecting 600 samples per class of MNIST and CIFAR10 so that there are 20 classes and 12,000 images in all for this new dataset where 10 classes are digits "0" to "9" of MNIST and another 10 classes contains "airplane" to "truck" of CIFAR10. For this dataset we design two few-shot tasks, MNIST-CIFAR10 (*i.e.* training on MNIST classes and testing on CIFAR10 classes) and the opposite CIFAR10-MNIST. To ensure the same size of the images from these two different datasets, we have preprocessed the gray images of MNIST into RBG forms of size 32×32.

Evaluation Metric. An n-way k-shot recognition task, also called a test episode, is actualized by randomly sampling n classes from all test classes and k samples per class, and then evaluating on another t test samples per class from that n classes. For a fair comparison with previous relevant works, we also evaluate our approach about 5-way 1-shot/5-shot recognition tasks for mini-ImageNet and 5-way/10-way 1-shot/5-shot for MNIST&CIFAR10 by an averaged accuracy over 600 random test episodes with 95% confidence intervals and each of test episode contains $t = 15$ test samples per class.

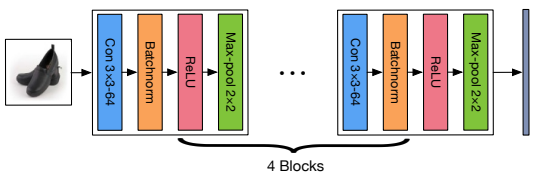

Fig. 3. Feature encoder architecture $f(\cdot|\mathbf{w})$ for the proposed DA-PN.

Experimental Setting. Similar to the PN, we also leverage a 4 blocks based ConvNet model each of which comprises 64-channel 3×3 convolution, batch normalization, ReLU nonlinearity and 2×2 max-pooling as feature encoder, as depicted in Fig. 3. The encoder leads to a 1,600-dimensional feature vector for one input image of mini-ImageNet and 256-dimensional feature vector for one input image of MNIST&CIFAR10. For mini-ImageNet, we use 30-way training episodes (*i.e.* $N_s = 30$) for 1-shot recognition and 20-way training episodes (*i.e.* $N_s = 20$) for 5-shot recognition. For MNIST&CIFAR10, we use 10-way training episodes (*i.e.* $N_s = 10$) for all few-shot settings. Meanwhile, we match $K_s = k$, $K_q = 15$ and $l = n * K_q$ for n-way k-shot tasks on both datasets. All models are implemented with Pytorch and run on an NVIDIA Tesla P40 GPU by Adam [34] using an initial learning rate of 0.001. We cut the learning rate by half every 2,000 episodes and stop training when the validation accuracy does not increase.

Compared Approaches. For mini-ImageNet, we compared our DA-PN with the recent competitive approaches, such as Matching Networks [22], Meta-Learner LSTM [20], Model-Agnostic Meta-Learning (MAML) [21], mAP-SSVM/mAP-DLM [24], PN [23] and Baseline Nearest Neighbors [20]. For MNIST&CIFAR10, the first compared baseline method is the Nearest Neighbors, which involves a simple nearest neighbor classifier on the 256-dimentional features of test data extracted by the feature encoder learned on the 10 training classes (*i.e.* digits "0" to "9" for task MNIST-CIFAR10, objects "airplane" to "truck" for task CIFAR10-MNIST). In addition, we reimplemented the PN under the same settings with our DA-PN on MNIST&CIFAR10 for comparing our method with the state-of-the-art few-shot learning algorithm.

Results on mini-ImageNet. We show our few-shot recognition results for mini-ImageNet in Table. I. Compared with the recent competitive few-shot learning approaches including PN, the proposed DA-PN leads to some persistent improvements under both 1-shot and 5-shot cases.

Results on MNIST&CIFAR10. Table. II and III show the experimental results on task MNIST-CIFAR10 and CIFAR10-MNIST of dataset MNIST&CIFAR10, respectively. Note that there exists a heavy domain gap between training classes and test classes for all designed few-shot recognition cases on MNIST&CIFAR10. The fact that the recognition results of Baseline Nearest Neighbors are close to random guess indicates the feature encoder learned on digits data of MNIST is not fit for objects data of CIFAR10 and vice versa. Despite the overall improvements from PN, the proposed DA-PN can further result in a better few-shot recognition performance by a large margin on all settings for MNIST&CIFAR10.

V. Conclusion

In this paper, we present Domain Alignment Prototypical Networks to tackle the few-shot image recognition tasks. It devises a domain alignment model under the framework of Prototypical Networks to abate the domain shifts existing between the data of different categories and then optimize the embedding space of prototype feature for each category. Furthermore, we design a modified episode training strategy to incorporate the domain alignment loss into the classification objective loss of Prototypical Networks. Experimental results show that our proposed model can boost few-shot image recognition compared with the recent competitive approaches.

Acknowledgment

This work is funded by the NSFC (Grant No.61473167), Beijing Natural Science Foundation (Grant No.L172037) and the German Research Foundation (DFG) in Project Crossmodal Learning, NSFC 61621136008 / DFG TRR-169.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553):436.

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Proceedings of the Advances in Neural Information Processing Systems, 2012. 1097-1105.

[3] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1-9.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770-778.

[5] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. 580-587.

[6] Girshick R. Fast r-cnn. Proceedings of the IEEE International Conference on Computer Vision, 2015. 1440-1448.

[7] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Proceedings of the Advances in Neural Information Processing Systems, 2015. 91-99.

[8] Fu K, Jin J, Cui R, et al. Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12):2321-2334.

[9] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional net- works for visual recognition and description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 2625-2634.

[10] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3156-3164.

[11] Wu Q, Shen C, Liu L, et al. What value do explicit high level concepts have in vision to language problems? Proceedings of the International Conference on Machine Learning, 2016. 203-212.

[12] Lake B, Salakhutdinov R, Gross J, et al. One shot learning of simple visual concepts. Proceedings of the Annual Meeting of the Cognitive Science Society, volume 33, 2011.

[13] Maas A, Kemp C. One-shot learning with bayesian networks. Cognitive Science Society, 2009.

[14] Lake B M, Salakhutdinov R R, Tenenbaum J. One-shot learning by inverting a compositional causal process. Proceedings of the Advances in Neural Information Processing Systems, 2013. 2526-2534.

[15] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. ICML Deep Learning Workshop, volume 2, 2015.

[16] Thrun S. Lifelong learning algorithms. Learning to learn. Springer, 1998: 181-209.

[17] Brazdil P B, Soares C, Da Costa J P. Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. Machine Learning, 2003, 50(3):251-277.

[18] Schweighofer N, Doya K. Meta-learning in reinforcement learning. Neural Networks, 2003, 16(1):5-9.

[19] Santoro A, Bartunov S, Botvinick M, et al. Meta-learning with memory-augmented neural networks. Proceedings of the International Conference on Machine Learning, 2016. 1842-1850.

[20] Ravi S, Larochelle H. Optimization as a model for few-shot learning. International Conference on Learning Representations, 2017.

[21] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the International Conference on Machine Learning, 2017. 1126-1135.

[22] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning. Proceedings of the Advances in Neural Information Processing Systems, 2016. 3630-3638.

[23] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. Proceedings of the Advances in Neural Information Processing Systems, 2017. 4080-4090.

[24] Triantafillou E, Zemel R, Urtasun R. Few-shot learning through an information retrieval lens. Proceedings of the Advances in Neural Information Processing Systems, 2017. 2252-2262.

[25] Gretton A, Borgwardt K M, Rasch M J, et al. A kernel two-sample test[J]. Journal of Machine Learning Research, 2012, 13(Mar): 723-773.

[26] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks. Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37. JMLR. org, 2015, 97-105.

[27] Tzeng E, Hoffman J, Darrell T, et al. Simultaneous deep transfer across domains and tasks. Proceedings of the IEEE International Conference on Computer Vision, 2015, 4068-4076.

[28] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. Proceedings of the International Conference on Machine Learning, 2015, 1180-1189.

[29] Long M, Zhu H, Wang J, et al. Unsupervised domain adaptation with residual transfer networks. Proceedings of the Advances in Neural Information Processing Systems, 2016, 136-144.

[30] Carlucci F M, Porzi L, Caputo B, et al. Just DIAL: Domain alignment layers for unsupervised domain adaptation. International Conference on Image Analysis and Processing. Springer, Cham, 2017:357-369.

[31] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3):211252.

[32] A. Krizhevsky. Learning multiple layers of features from tiny images. Masters thesis, Department of Computer Science, University of Toronto, 2009.

[33] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[34] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.