Adaptive Multi-prototype Relation Network

Xiaoxu Li*, Tao Tian*, Yuxin Liu§, Hong Yu†, Jie Cao* and Zhanyu Ma‡
* Lanzhou University of Technology, Lanzhou, China
† Ludong University, Yantai, China
§ University of Melbourne, Melbourne, Australia
‡ Beijing University of Posts and Telecommunications, Beijing, China
E-mail:{hy@ldu.edu.cn, mazhanyu@bupt.edu.cn}

***Abstract*—Based on Relation Network, we propose a new network structure that can adaptively adjust the number of prototypes according to data distribution. Our method, called the Adaptive Multi-prototype Relation Network(AMRN), aims at extracting more reasonable prototype representation for different data distribution in few-shot learning case. Instead of representing each class as a single prototype in the relational network, we represent each class with one or more prototypes, and solve the problem of embedding network with the relational network connection, which can improving the classification accuracy in few-shot learning. Besides, our method can easily extend to other network structures, which is also a useful reference for other metric learning approaches.**

I. Introduction

Since the invention of deep neural networks [1], it has become a trending topic and achieved great success in many fields [2] [3] [4] [5]. However, the traditional deep learning method still has notable limitations, especially in visual recognition tasks [6] [7] [8], the deep neural network requires a considerable amount of training samples and time [9] [10] [11] [12] to build an effective model, and there will even be an additional cost if annotating the training samples manually. Moreover, it is challenging to collect enough data for training in some fields, with insufficient samples, the trained deep neural networks are troubled with the problem of overfitting or underfitting [13] [14] [15] [16].

Non-parametric methods are suitable for small sample tasks, and the most popular one is the nearest neighbor methods. Nearest neighbor methods are high-capacity models that represent a class by storing all of its examples, hence can capture complex distributions. In recent years, metric-based prototype parametric methods [17] have achieved significant progress. One of the prototypical methods is Gaussian mixture models, it belongs to low-capacity models which can fit simple distributions robustly by representing a class by its examples' means and variances. However, models using methods are only suitable for specific data distributions, and they can not adjust model size flexibly in terms of the data distribution. To solve this problem, Kelsey R.Allen and Evan Shelhamer proposed Infinite Mixture Prototypes (IMP) [18] which represent a class as a set of clusters and the number of clusters is data determined. IMP can learn a deep embedding network, which can adjust the model capacity based on the given data.

The metric-based learning method has gained some achievements in small samples classification. However, it mainly uses specific Euclidean distance [17] or Cosine distance as the measurement method, which restricts the adjustability of the metric method when encountering various data sets. Base on the Prototype network, Sung proposed Relation Network [19] for few-shot image recognition task [20], it aims to learn the similarity between query images and labelled sample images from training. After learning how to measure the similarity between samples through the neural network, the performance of classification with small data has improved significantly.

IMP provides a more reasonable scheme for the structure of the classification model, whereas Relational Network provides a more efficient way as measurement function. To retain the advantages of IMP and relation network, we introduce Adaptive Multi-prototype Relation Network, given the support data, it can generate one or more prototype for each class.

The main contribution of the paper is that it provides a new network structure that can adjust the network capacity according to the data distribution. And the classification accuracy based on four different datasets has improved.

II. Related Work

Metric Learning [21] Approaches are most relevant to our work, it builds a distance measurement method which allows samples in the same class close to each other and samples from different classes are further apart. Use appropriate distance measurement, non-parameter estimation methods can be applied to build small data classifiers. Using KNN [22] as an example, it does not require training, its performance mainly depends on the choice of distance measure. With the development of neural networks based machine learning methods, some end-to-end nearest neighbor classifiers build by different neural networks structure have been proposed. These models can integrate the advantages of both parameters and non-parameters models that can learn the presentation of unlabeled new samples quickly and suitable for small samples classification task.

Recently there are numerous few-shot learning [23] models, such as Matching networks [8], Siamese Neural Networks [6], Prototypical Networks [17] and Relation Network [19] . Matching networks [8] can generate labels for unseen classes without changing the network model. Its main innovations are in the modelling and training process, in the modelling process, memory and attention-based networks are selected

978-988-14768-8-3/20/$31.00 ©2020 APSIPA

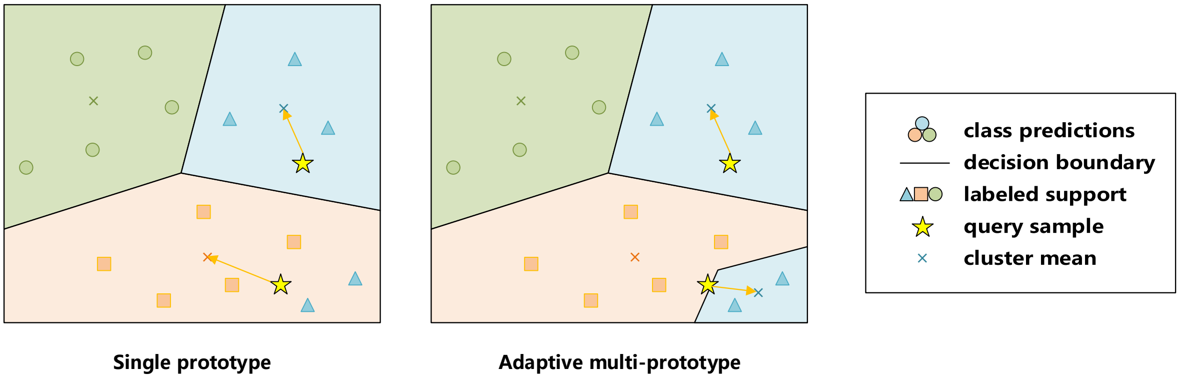

Figure 1. The prototype obtained by the Prototype Network is shown in the figure on the left. The prototype is obtained by averaging the support set samples according to classes, classified by comparing the query samples and the prototype. Our strategy is to cluster the supporting set samples and get one or more prototypes according to the data distribution, as shown in the figure, such a method is more conducive to making correct predictions.

to improve learning speed. In the training process, the proposed condition matching idea emphasizes that the training and testing should be carried out under the same condition, which requires during the training stage, the network should only see a small number of samples from each class which will be consistent with the testing process. Siamese Neural Networks [6] restrict the input structure and automatically discover the features that generalized from the new samples, this was trained through supervised metric learning based on twin networks which share the same weights and network parameters. During training, to learn the distance between pairs of input data samples, we send sample pairs into the twin networks and combine the two outputs. At the top layer, use the cross-entropy loss to determine whether inputs pairs belong to the same class. Prototypical Networks [17] is efficient and straightforward, it learns an embedding function to encode each input into a feature vector. By calculating the mean value of feature vectors in the same class, each class will generate a prototype representation. Based on the idea that feature vectors in the same class should gather around the prototype representation, the classification problem becomes to find the nearest neighbor of different prototype representations. A typical prototype network only gives one prototype representation to each class during the modelling process. It results in failing of fitting the distribution of complex data sets in the embedding space. To solve this problem, IMP is preferred as it provides multi-prototype representations to each class according to the distribution of feature vectors. As shown in figure 1, the IMP method can improve classification accuracy considerably. When measuring the distance between the feature vector of a test sample and multi-prototype representations in the evaluation step, instead of using Cosine distance or Euclidean distance selected by normal prototype network, we use a learnable network to find more appropriate distance measurement method.

The learnable of distance measurement is proposed in Relation Network, which is also a popular few-shot learning neural networks. It focuses on measuring the similarity between query and support samples, the class decision is judge by relation scores. Compared with Relation Network, the Prototype Network focuses on projecting samples into a more effective embedding space while the similarity of samples in the embedding space is measured by artificial method (such as Euclidean distance and Cosine distance). Whereas, the Relation Network intends to learn an effective distance measuring method for training samples to perform the few-shot classification task but without building a more nature prototype representation for class information. IMP offers more suitable prototype representations for each class and exhibits the capability of multi-prototype to improve classification performance.

III. Methodology

A. Problem Definition

In the task of Few-Shot Learning (FSL), its training and testing process are called meta-training and meta-testing.

A few-shot training set contains a number of classes, with multiple samples in each class. In the training stage, N classes will be randomly selected from the training set, with K samples in each class (total $N * K$ samples) to construct a meta-task, which will be input as the support set of the model. A batch sample is then extracted from the remaining data from these N classes as the query set of the model. The FSL task then requires the model to learn how to distinguish the N classes from the $N * K$ samples, which is called the N-way K-shot problem.

During the training process, different meta-tasks are randomly sampled from the training dataset in each epoch, in other words, each training meta includes a different combination of classes. This mechanism enables the model to learn more common features through different meta-tasks. These features will focus more on expressing the underlying nature of each class and features belong to the same class are more gathered. Since each meta-task has different training label, when

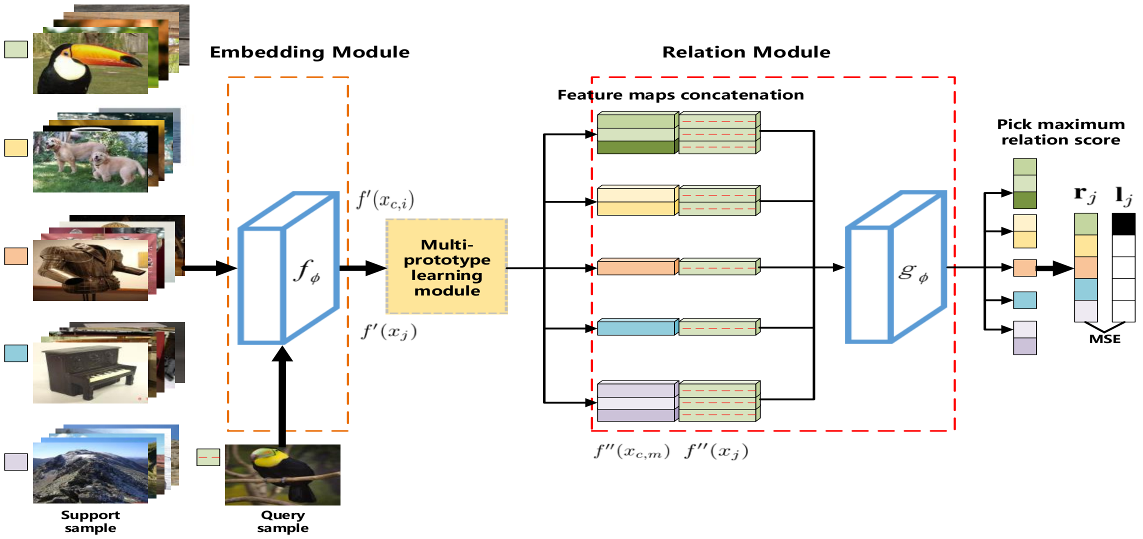

Figure 2. Adaptive Multi-prototype Relation Network for a 5-way 5-shot problem with one query example.

encountering a small data samples, learned features could get rid of the label information and the overfitting problem can be restrained. Models learned through this mechanism can also be suitable for new meta-tasks, in the meta testing phase, support and query set built on the new class can be classified accurately by using the trained meta-model.

B. *Network Architecture*

As shown in Figure 2, our model is mainly composed of embedding module and relation module. The embedding module extracts the image features, following the distribution of embedding features in different classes, each class is mapped as multi-prototype representations by multi-prototype learning module. Using DP-Means algorithm to implement the multi-prototype learning module, details of the module can be found in Section III C. After combining feature map of prototype representations and query set, relation pairs are feeding into relation module, we could get predicted labels by selecting maximum relation score. Compared with the Relation Network, the most significant improvement is the additional multi-prototype learning module, which enhances the ability to express the nature of the data distribution.

In the K-shot problem, the data processing flow is as follows: embedding module f_ϕ extracts features of support samples and query samples, then feed K features of support sample to multi-prototype learning module, instead of obtaining a single prototype by element-wise sum features in the same class, cluster these vectors to generate one or more prototype representations for each class. Finally, relation module makes predictions by comparing the similarity between the query sample and the prototypes. When dealing with the one-shot problem, our network degenerates into the normal Relational Network, so only the K-shot processing is described here.

For K-shot where $K > 1$, as illustrated in Figure 2. Samples x_j in the query set Q, and samples $x_{c,i}$ in the support set S are into through the embedding module f_ϕ, which produces feature maps $f'(x_{c,i})$ and $f'(x_j)$. K feature maps $f'(x_{c,i})$ in the same class are mapped into M prototype representations $f''(x_{c,m})$ by multi-prototype learning module and $f'(x_j)$ is mapped as $f''(x_j)$ too. $1 \leq m \leq K$, when $m = 1$, our network is equivalent to Relation Network, $m > 1$, the prototype representations $f''(x_{c,m})$ and $f''(x_j)$ are combined together and feed into relation module g_ϕ, which eventually produces a scalar in the range from 0 to 1 to represent the similarity of input pairs, which is called relation score, then taking the maximum relation score of the relation pairs for each class as predict label.

C. *Multi-prototype learning module*

The key of the multi-prototype learning module is the DP-Means algorithm, DP-means [24] is a deterministic, hard clustering algorithm derived via nonparametric Bayesian for the Dirichlet process. As illustrated in Algorithm 1, DP-means iterates over the data points, computing the minimum distance to all existing cluster means of each point. If this distance is greater than the threshold λ, a new cluster will be created with mean equal to the point. It optimizes a k-means-like objective for reconstruction error plus a penalty for making clusters.

D. *Loss function*

After the relational network calculation, predict the value of each class by taking the maximum value, and then calculate the Mean Square Error (MSE) loss to train model, $\mathbf{r}_j$ is relation score vector, $\mathbf{l}_j$ is one-hot vector of label.

Table I
COMPARISON OF THE CLASSIFICATION PERFORMANCE ON THE Omniglot, mini-Imagenet, Stanford-cars, AND Caltech101 DATASETS. THE MODULES ARE: Prototype network, Matching Nets, Relation Net AND Adaptive Multi-prototype Relation Network (Ours). ALL OURS ACCURACY RESULTS ARE AVERAGED OVER 600 TEST EPISODES AND ARE REPORTED WITH 95% CONFIDENCE INTERVALS.

Methods	Omniglot	mini-Imagenet	Stanford-cars	Caltech 101
Prototype network	99.7	63.15	62.14	-
Matching Nets	98.9	55.31	64.74	-
Relation Net	99.708	62.663	64.59	74.089
Ours	**99.718**	**64.719**	**65.086**	**76.195**

Table II
ABLATION STUDY OF Relation Net AND Adaptive Multi-prototype Relation Network (Ours) ON Omniglot, mini-Imagenet, Stanford-cars DATASETS. ALL OURS ACCURACY RESULTS ARE AVERAGED OVER 600 TEST EPISODES AND ARE REPORTED WITH 95% CONFIDENCE INTERVALS.

	Omniglot		mini-Imagenet		Stanford-cars	
Methods	5-way	10-way	5-way	10-way	5-way	10-way
Relation Net	99.708	99.64	62.663	46.743	64.59	44.761
Ours	**99.718**	**99.66**	**64.719**	**48.94**	**65.086**	**45.825**

Algorithm 1 Dp-means clustering algorithm

```
Input: $x_1, ..., x_n$: input data, $\lambda$: cluster penalty parameter
Output: Clustering $\ell_1, ..., \ell_2$ and number of cluster $k$
  Init.$k = 1, \ell_1 = \{x_1,...,x_n\}$,and $\mu_1$ the global mean.
  Init.cluster indicators $z_i$ for all $i = 1, ..., n$.
  Repeat until convergence
  for each point $x_i$ do
      Compute $d_{ic} = ||x_i - \mu_c||^2$ for $c = 1, ..., k$
      If $\min_c d_{ic} > \lambda$,set $k = k + 1, z_i = k$,and $\mu_k = x_i$.
      otherwise, set $z_i = \mathrm{argmin}_c d_{ic}$.
  Generate cluster $\ell_1$,...,$\ell_k$ based on $z_1$,...,$z_k$:$\ell_j = \{x_i|z_i = j\}$
  For each cluster $\ell_j$,compute $\mu_j = \frac{1}{|\ell_j|}\sum_{x\in\ell_j} x$
```

$$\varphi, \theta \leftarrow \arg\min_{\varphi,\theta} \sum_{j=1}^{N} |\mathbf{r}_j - \mathbf{l}_j|^2$$

We may consider a better solution to obtain single predict value without the maximizing procedure. Because maximizing the information along the way might cause other parts to lose useful gradient information. If this mechanism can be resolved, the overall performance of the network can be improved.

IV. Experiments

A. Datasets

We used the Omniglot, mini-Imagenet, Caltech-101 and fine-grained dataset Stanford-cars. By conducting the classification experiment of the above datasets, we proved the effectiveness of our network. All experiments are implemented based on PyTorch.

Omniglot [14] is a dataset of 1,623 handwritten characters from 50 alphabets. There are 20 examples of each character, where the images are resized to 28∗28 pixels and each image is rotated by multiples of 90°. This gives 6,492 classes in total, which are then split into 4,112 training classes, 688 validation classes, and 1,692 test classes.

Mini-ImageNet [8] is a reduced version of the ILSVRC'12 dataset [25], which contains 600 84∗84 images for 100 classes randomly selected from the full dataset. We use the split from Ravi & Larochelle [26] with 64/16/20 classes for train/val/test.

Stanford-cars Dataset contains 16,185 images of 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images, where each class has been split roughly in a 50-50 split. Classes are typically at the level of Make, Model, Year, e.g. 2012 Tesla Model S or 2012 BMW M3 coupe.

Caltech 101 The Caltech 101 dataset consists of a total of 9146 images, split between 101 different object categories, as well as an additional background/clutter category. Each object category contains between 40 and 800 images on average. Common and popular categories such as faces tend to have a larger number of images than less used categories. Each image is about 300∗200 pixels in dimension. Images of oriented objects such as airplanes and motorcycles were mirrored to be left-right aligned, and vertically oriented structures such as buildings were rotated to be off axis.

B. Implementation details

As most few-shot learning models utilise four convolutional blocks for embedding module [17] [8], we follow the same architecture setting for a fair comparison. More concretely, embedding module is composed of four convolution block, each convolutiona block contains a 64 channels, $3 * 3$ Convolution kernel convolution layer, a batch normalization and a ReLU non-linearity layer respectively. The first two blocks also contain a $2 * 2$ maxpooling layer while the latter two do not. The relation module consists of two convolutional blocks and two fully-connected layers, each of convolutional block contains a $3 * 3$ convolution kernel, 64 channels convolution layer, followed by batch normalization layer, ReLU activation layer and $2 * 2$ max-pooling layer. Take the sample of mini-ImageNet dataset as an example, in case of 5-way 5-shot, feeding $25 * 3 * 3 * 84$ tensors to embedding module, output is $25 * 64 * 19 * 19$ vector, these tensors are expanded by class and processed by multi-prototype learning module to produce $\sum_{i=1}^{5} m_i * 64 * 19 * 19$ prototypes, the prototypes and query features are concatenated into $\sum_{i=1}^{5} m_i * 128 * 19 * 19$ relation pairs as input of relation module, finally end up with a $5 * 1$ relationship tensors. In our experiment, the values of batch size, learning rate and episode are 10, 0.001 and $500,000$ respectively.

C. Results

In different datasets, shown in table I that the classification accuracy has been improved to a certain extent, especially in the coarse-grained datasets mini-Imagenet and Caltech-101, where the classification accuracy has increased by more than 2%. It also performs better than the original model on fine-grained datasets. Table I shows that for other typical classification models, our model also has advantages over several representative small sample datasets.

As can be seen from Table II, in the experiment of 10-way 5-shot, with the increase of sample class, our classification performance is still superior to the original model. Although our network structure has increased somewhat, the training time to reach the optimal model has not increased much.

Based on the above datasets, given a random seed, we have done groups of the control experiment. From the experimental results, the performance of our model is superior to the Relation Network. Especially for the datasets with significant variance, performance improvement is more prominent. Verifying for the same class, the assumption that one or more prototypes exist.

When it comes to datasets with insignificant intraclass distance, like Omniglot dataset, there is no more prototype within the class, by adjusting the clustering radius λ, the network can be reduced to the Relation Network. Experiments on coarse-grained data sets mini-Imagenet and Caltech 101 also prove that there may be one or more prototypes on datasets with relatively discrete data distribution. Hence our network structure performs better in these two datasets.

V. Conclusion

In this paper, we introduce Adaptive Multi-Prototype Relation Network, a Network structure that combines the characteristics of IMP and Relational Network. Through a series of experiments, there are a few conclusions in this work. Firstly, the data mapped to the high-dimensional space, especially in the coarse-grained datasets, still have different distributions among the same class, according to the distribution of data the choice of model is particularly essential. Secondly, Among various metric learning methods, the learnable measurement method is usually more appropriate than the traditional measurement method. We believe this idea is also worthy as a reference to many metric learning models.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2019YFF0303300 and Subject II No. 2019YFF0303302, in part by the National Natural Science Foundation of China (NSFC) under Grant 61906080, Grant 61763028, Grant 61773071, Grant 61922015, and Grant U19B2036, in part by the Beijing Natural Science Foundation Project No. Z200002, in part by the Beijing Academy of Artificial Intelligence (BAAI) under Grant BAAI2020ZJ0204, in part by the Beijing Nova Programme Interdisciplinary Cooperation Project under Grant Z191100001119140, in part by Key Research and Development Project of Shandong Province No.2019GGX101036, in part by the Hong-liu Outstanding Youth Talents Foundation of Lanzhou University of Technology.

References

[1] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455, 2009.

[2] Harrison Edwards and Amos Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.

[3] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016.

[4] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*, 2017.

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

[6] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[8] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[9] Alex Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25(2), 2012.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 06 2016.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Xiaoxu Li, Dongliang Chang, Zhanyu Ma, Zheng-Hua Tan, Jing-Hao Xue, Jie Cao, Jingyi Yu, and Jun Guo. Oslnet: Deep small-sample classification with an orthogonal softmax layer. *IEEE Transactions on Image Processing*, 2020.

[13] Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 464–471. IEEE, 2000.

[14] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[15] Xiaoxu Li, Liyun Yu, Xiaochen Yang, Zhanyu Ma, and Jun Guo. Remarnet: Conjoint relation and margin learning for small-sample image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2020.

[16] Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29:4683–4695, 2020.

[17] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[18] Kelsey R Allen, Evan Shelhamer, Hanul Shin, and Joshua B Tenenbaum. Infinite mixture prototypes for few-shot learning. *arXiv preprint arXiv:1902.04552*, 2019.

[19] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[20] Xiaoxu Li, Liyun Yu, Dongliang Chang, Zhanyu Ma, and Jie Cao. Dual cross-entropy loss for small-sample fine-grained vehicle classification. *IEEE Transactions on Vehicular Technology*, 68(5):4204–4212, 2019.

[21] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2006.

[22] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 986–996. Springer, 2003.

[23] Xiaoxu Li, Zhuo Sun, Jing-Hao Xue, and Zhanyu Ma. A concise review of recent few-shot meta-learning methods. *arXiv preprint arXiv:2005.10953*, 2020.

[24] Brian Kulis and Michael I Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*, 2011.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[26] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.