Class-Discriminative Feature Embedding For Meta-Learning based Few-Shot Classification

Alireza Rahimpour[1,2] and Hairong Qi[1]

[1] Department of Electrical Engineering and Computer Science,
University of Tennessee, Knoxville, TN. {arahimpo, hqi}@utk.edu
[2] Ford Motor Company, Greenfield labs, Palo Alto, CA, 94304. arahimpo@ford.com

Abstract

Although deep learning-based approaches have been very effective in solving problems with plenty of labeled data, they suffer in tackling problems for which labeled data are scarce. In few-shot classification, the objective is to train a classifier from only a handful of labeled examples in a support set. In this paper, we propose a few-shot learning framework based on structured margin loss which takes into account the global structure of the support set in order to generate a highly discriminative feature space where the features from distinct classes are well separated in clusters. Moreover, in our meta-learning-based framework, we propose a context-aware query embedding encoder for incorporating support set context into query embedding and generating more discriminative and task-dependent query embeddings. The task-dependent features help the meta-learner to learn a distribution over tasks more effectively. Extensive experiments based on few-shot, zero-shot and semi-supervised learning on three benchmarks show the advantages of the proposed model compared to state-of-the-art.

1. Introduction

Deep learning has made major advances in many areas but still has limitations when it comes to problems with limited number of labeled data. Humans on the other hand are able to rapidly learn new classes. For example, a child can learn to recognize a new object by only seeing one picture of that object. Human can recognize objects even without seeing the examples of that object category and just by hearing the description of that object (similar to zero-shot learning). This significant gap between human and machine learning and the fact that many practical recognition systems should be able to recognize a new category from a handful of training images, provides fertile ground for few-shot learning developments.

Few-shot classification is a task in which a classifier must be able to generalize from few examples. Recently there has been a surge of interest in using meta-learning (learning-to-learn) for few-shot learning [33, 38, 29, 34]. These approaches use a meta-learning strategy which includes extracting some transferable knowledge from a set of tasks and transferring the knowledge to quickly adapt to new tasks without suffering from the overfitting that might happen when applying deep models to problems with small amount of data. Specifically, these meta-learning based models utilize sampled mini-batches called episodes during training, where each episode is designed to mimic the few-shot task by sub-sampling classes as well as data points. The use of episodes makes the training problem more faithful to the test environment and thereby improves generalization [38]. In fact, the meta-learner learns a strategy for generalizing to an unseen task from some similar task distributions. Here instead of learning the distribution of data samples (as in regular machine learning algorithms), the model learns the distributions of tasks.

Several successful directions have been explored recently for meta-learning-based few-shot learning, including learn to fine-tune [9, 28], sequence based methods [31], and metric learning models [38, 33, 34]. However, there are still challenges in solving the few-shot learning problem. For instance, even though reducing the intra-class variation is a very critical factor in the current few-shot classification problem setting, recent works seldom explicitly study it. In this work, we address this issue by defining a structured-based margin loss to explicitly decrease the intra-class distance between feature embedding of each class in the support set and create a structured support set embedding. The structured-based margin considers the relationship between all the support set samples in minimizing the loss and guides the model to learn a deep metric to cluster the support set embeddings and generates a highly discriminative feature space where all classes are well separated. We refer to the proposed Class-Discriminative Few-Shot learning framework in this paper as CDFS.

In episode-based few-shot learning frameworks a task is defined based on a support set and its relationship with the query in each episode. It has been shown in [24], that incorporating task information to the feature embedding can highly improve the performance of few-shot classification. The proposed context-aware query embedding in this paper incorporates task information into query embedding in each episode using attention mechanism and 1-D CNN.

Besides few-shot learning, we also show the applicability of our proposed model for zero-shot classification. In the zero-shot setting, each class comes with a category description (meta-data) giving a high-level description of the class rather than a number of labeled examples. The model therefore learns an embedding of the meta-data into a shared space to serve as the prototype for each class. Classification is performed, as in the one-shot scenario, by finding the nearest class prototype for an embedded query point.

The main contributions of this paper are summarized as follows:

- Regularizing the few-shot classification setting with a structured-based margin loss which takes into account the global structure of the support set feature space and learns to explicitly reduce the intra-class variation in order to map the data to a highly discriminative feature space where the few-shot classification is most effective.
- Proposing a context-aware query embedding module which takes into account the support set's context and generates task-dependent feature representations which would help the meta-learner to learn a distribution over tasks more effectively.
- Performing extensive experiments based on few-shot, one-shot, zero-shot and semi-supervised learning schemes to show the advantages of the proposed model compared to state-of-the-art.

2. Related Work

Recently there has been a resurgence of interest in few-shot learning based on meta-learning [9, 38, 33, 28, 31, 22, 34, 18, 10, 25]. The existing meta-learning models for few-shot classification can be divided into three types: the learning to fine-tune based, RNN based, and metric learning based. For instance, in [9] the MAML model aims to meta-learn an initial condition that is good for fine-tuning on few-shot problems. The model in [28] is an LSTM-based optimizer that is trained to be specifically effective for fine-tuning. In [31], a recurrent neural network iterates over examples of given problem and accumulates the knowledge required to solve that problem in its hidden activations. However, these recent works either require fine-tuning the target problem [9, 28], or need the use of complex recurrent neural network (RNN) architectures [31, 38], or are based on complicated inference steps [7]. In our work, the model is simple and fast and does not need any additional process such as fine tuning. Moreover, we avoid the complexity of recurrent networks, and the issues involved in ensuring the adequacy of their memory. Instead our proposed approach is defined entirely with feed forward convolution neural networks.

The metric based few-shot learning has attracted a lot of interests recently [38, 33, 34, 19]. The basic idea is to learn a metric which can map similar samples close and dissimilar ones distant in the metric space so that a query can be easily classified. Various metric based methods such as siamese networks [5], matching networks [38], prototypical networks [33], and relation networks [34] have been proposed. They differ in their ways of learning the metric. For instance, very recently the relation network [34] proposed to replace the fixed metric learning part (e.g., Euclidean distance) of the previous works with a deep metric for comparing the relation between images.

The success of metric based methods relies on learning a discriminative metric space. The proposed method in this paper can be categorized as the metric learning based framework. To reach the full potential of metric based few-shot learning, we augment the classification loss with a structure-based deep metric learning regularization which enforces the model to map the samples in the support set to well separated clusters in the embedding space. Unlike the metric learning methods based on contrastive [5] or triplet [32, 27] loss that are defined in terms of data pairs or triplets, our approach takes into account the global structure of the embedding space. In fact, the structured margin term in the loss function measures the quality of clustering the data by taking into account the relationship between all the data points in the mini batch at once (instead of data pairs or triplets). Furthermore, this deep learning based metric learning framework does not require the training data to be preprocessed in rigid paired or triplet format and uses a structured prediction framework [37, 12] to ensure that the score of the ground truth clustering assignment is higher than the score of any other clustering assignment.

Taking advantage of contextual information in the support set is critical in episode-based few-shot learning models. A framework for context modeling in the support set was proposed in [38] based on a bi-directional LSTM. However, as the number of classes and shots increases, the model is required to learn longer and more complex dependencies, which negatively affects both generalization and efficiency.

Furthermore, it imposes an arbitrary ordering on the support set by using bi-directional LSTM (i.e., the embedding changes if we shuffle the support set samples). Moreover, the meta-learner architecture proposed in [21] combines temporal convolutions (which aggregate contextual information from past) with causal attention which pinpoints to specific pieces of information. Also, in [14] the Graph Neural Networks (GNNs) are used to show the importance of modeling the relationship between the query and support set in solving the few-shot classification problem.

In this paper, we propose a simpler but effective context-aware query embedding framework based on attention mechanism and 1-D CNN for taking into account the context of the support set and its relationship (i.e., task) with query embedding. The proposed query encoder makes the query embedding task-dependent which helps learning a meta-learner with higher generalization power.

3. Method

In this section we first describe the meta-learning based few-shot classification. We then elaborate on components of our proposed model including structured support set embedding and context-aware query embedding modules.

3.1. Few-Shot Classification

The meta-learning based few-shot classification is defined based on episodic training. The idea behind the episodic paradigm is to simulate the few-shot task that will be encountered at test time. In each training iteration, an episode is formed by randomly selecting N_C classes from the training set with K labeled samples from each class to act as the support set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, where $m = K \times N_C$ and a query set $\mathcal{Q} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_Q}$ of different examples from the same N_C classes. Each $\mathbf{x}_i \in \mathbb{R}^D$ is an input vector of dimension D and $y_i \in \{1, 2, \dots, N_C\}$ is a class label. Training on such episodes is done by feeding the support set $\mathcal{S}$ to the model and updating the model's parameters to minimize the loss of its predictions for the examples in the query set $\mathcal{Q}$. This form of training allows the model to extract transferable knowledge based on different classification tasks seen in the episodes so the model can exploit this knowledge in testing stage to classify the query samples coming from new unseen classes.

In the proposed model, we employ a few-shot learning structure based on episodic training as in Prototypical Networks [33] which uses the support set $\mathcal{S}$ to extract a prototype $\mathbf{c}_j \in \mathbb{R}^N$ from each class $j = 1, \dots, N_C$ through an embedding function $f_\phi(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}^N$, where ϕ is the learnable parameters of the neural network. Each prototype is defined as the mean vector of the embedded support points belonging to its class:

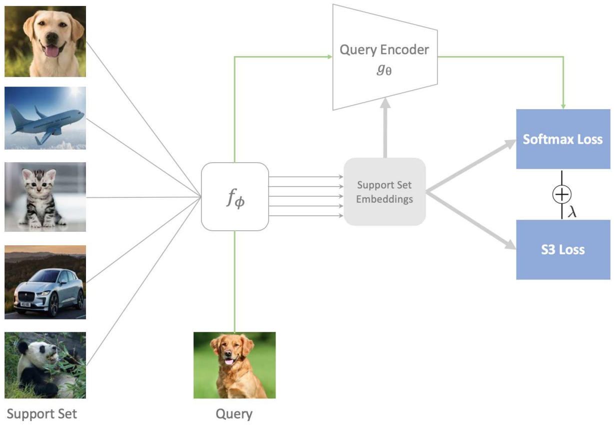

Figure 1. Model architecture for 5-way, 1-shot classification.

$$\mathbf{c}_j = \frac{1}{|\mathcal{S}_j|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_j} f_\phi(\mathbf{x}_i), \qquad (1)$$

where $i = 1, \dots, K$. The samples in the query set are then classified based on their distance to the prototype of each class and a distribution over classes for a query point $\mathbf{x}_q$ is defined based on a softmax over distances to the prototypes in the embedding space [33]:

$$p_\phi(y = j|\mathbf{x}_q) = \frac{\exp(-d(\mathbf{c}_j, f_\phi(\mathbf{x}_q)))}{\sum_{j'} \exp(-d(\mathbf{c}_{j'}, f_\phi(\mathbf{x}_q)))} \qquad (2)$$

It has been shown in [33], that the prototype in Eq. 1 yields cluster representatives with the prototype as the cluster center and there is one cluster per class when a Bregman divergence such as squared Euclidean distance is used.

In order to learn more discriminative embeddings for the few-shot learning task, in this paper we propose to impose a constraint based on structured margin on support set, to explicitly enforce the class separation in embedding space based on the global structure of the support set. Furthermore, a context-aware query embedding module is proposed to create task-dependent query features and also to pull the feature embedding of the query towards corresponding class prototype in support set. The model architecture is shown in Figure 1.

3.2. Structured Support Set Embedding

Similar to metric-based few-shot learning methods [33, 38, 34], our model learns a nonlinear embedding function $f_\phi(\mathbf{x})$, parameterized as a neural network, that maps examples into a space where examples from the same class are close and those from different classes are far apart. The embedded point $f_\phi(\mathbf{x})$ is then classified by a classifier, e.g., the softmax classifier. In this paper, our objective is to learn highly discriminative features with the joint supervision of softmax loss and Structured Support Set (S3) loss as follows:

$$\mathcal{L} = \mathcal{L}_{\text{softmax}} + \lambda \times \mathcal{L}_{S3}, \tag{3}$$

where $\mathcal{L}_{\text{softmax}}$ is:

$$\begin{aligned}\mathcal{L}_{\text{softmax}} = \frac{1}{N_Q} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{Q}_j} \Big[& d(\mathbf{c}_j, g_\theta(f_\phi(\mathbf{x}_i))) + \\ & \log \sum_{j'} \exp(-d(\mathbf{c}_{j'}, g_\theta(f_\phi(\mathbf{x}_i)))) \Big],\end{aligned} \tag{4}$$

which is simply defined based on the average negative log-probability of the correct class assignments, for all query examples. $g_\theta(.)$ is the context-aware query embedding function to be described in the next section and λ is a scalar used for balancing the two loss functions. $\mathcal{L}_{S3}$ is the Structured Support Set (S3) loss which guides the training by enforcing a margin $\Delta(\mathbf{y}, \hat{\mathbf{y}})$ based on the global structure of the support set as follows:

$$\begin{aligned}&\mathcal{L}_{S3}(X, f_\phi) = \\ &\Big[F(X, \hat{\mathbf{y}}; f_\phi) + \gamma \Delta(\mathbf{y}, \hat{\mathbf{y}}) - F(X, \mathbf{y}; f_\phi) \Big]_+,\end{aligned} \tag{5}$$

$$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \text{AMI}(\mathbf{y}, \hat{\mathbf{y}}), \tag{6}$$

where $[z]_+ = \max(z, 0)$ and $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is the set of samples in the support set and $\hat{\mathbf{y}}$ and $\mathbf{y}$ are the predicted and ground-truth support set labeling assignments, respectively. This loss encourages the model to learn an embedding function f_ϕ such that the ground truth labeling score for the support set $F(X, \mathbf{y}; f_\phi)$ is greater than the score for any other label assignments of the set $F(X, \hat{\mathbf{y}}; f_\phi)$, at least by the structured margin $\Delta(\mathbf{y}, \hat{\mathbf{y}})$. $\mathcal{L}_{S3}$ can be considered as a generalization of the triplet loss which takes into account the whole structure of the support set instead of only three samples. F is defined as a scoring function that encourages the embeddings of samples in each class to be as close as possible to the prototype of that class and reduces the intra-class distance between embeddings of each class and results in a compact feature representation of that class around its prototype as follows:

$$F(X, \hat{\mathbf{y}}; f_\phi) = -\sum_{\mathbf{x}_i \in X} \min_j ||f_\phi(\mathbf{x}_i) - \mathbf{c}_j||_2^2, \tag{7}$$

where $\mathbf{x}_i$ is the ith data sample (e.g., image) in the support set and $j = 1, \ldots, N_C$.

The structured margin has been used in structure prediction problems such as structured SVM [8], structured KNN [26], etc., where the problem involves predicting structured objects. In our problem the structured output is defined as the labeling configuration of the support set. We define the structured margin $\Delta(\mathbf{y}, \hat{\mathbf{y}})$ to measure the quality of the label assignment of the support set as in Eq. 6, where AMI is the Adjusted Mutual Information and is defined as:

$$\text{AMI}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{MI(\mathbf{y}, \hat{\mathbf{y}}) - E\{MI(\mathbf{y}, \hat{\mathbf{y}})\}}{1/2(H(\mathbf{y}) + H(\hat{\mathbf{y}})) - E\{MI(\mathbf{y}, \hat{\mathbf{y}})\}}, \tag{8}$$

where MI is the mutual information which is a non-negative quantity which quantifies the information shared by the two label sets (i.e., clusterings), $E\{MI(\mathbf{y}, \hat{\mathbf{y}})\}$ is the expected value of the MI, and H is the entropy. The AMI takes a value of 1 when the two sets are identical and 0 when the MI between two sets equals the value expected due to chance alone. In fact, AMI is an adjustment of the MI score to account for chance. Furthermore, AMI has several other important properties, such as being a metric and a normalized measure, and using the nominal $[0, 1]$ range better than other normalized variants. Our experiments show that using AMI leads to better recognition accuracy compared to normalized MI and other similarity measures.

Training of the model is performed by minimizing the average loss, iterating over training episodes and performing a gradient descent update for each. For calculating the gradient of the $\mathcal{L}_{S3}$ loss, we use the algorithm based on approximate sub-gradients [23, 36] with a simplifying assumption of considering the mean of the embeddings of each class as the prototype (centroid) of each cluster. For more details about approximate sub-modular optimization for structured prediction please refer to [36, 20]. All parameters of our model lie in the embedding function and by using the combination of softmax loss and structured margin loss, the model learns a discriminative embedding function with two key learning objectives, inter-class dispersion and intra-class compactness.

3.3. Context-Aware Query Embedding

The goal of this part of the model is to create task-dependent query embeddings and pull them towards their class prototypes based on the task context in each episode. Task-dependent query features help the meta-learner to learn a more effective distribution over the tasks. Let $f_\phi(\mathbf{x}_q)$ be the embedding of a query image taken from the CNN and $\mathbf{c}_j$ be the prototype of the jth class. For each sample in the query set $\mathcal{Q}$, a context vector $\mathbf{v}_q$ is extracted from the support set based on the similarity of the query embedding and the prototypes in the support set. The context vector is calculated using a content-based attention mechanism as follows:

$$a(\mathbf{c}_j, f_\phi(\mathbf{x}_q)) = \frac{\exp(-d(\mathbf{c}_j, f_\phi(\mathbf{x}_q)))}{\sum_{n=1}^{N_C} \exp(-d(\mathbf{c}_n, f_\phi(\mathbf{x}_q)))}, \tag{9}$$

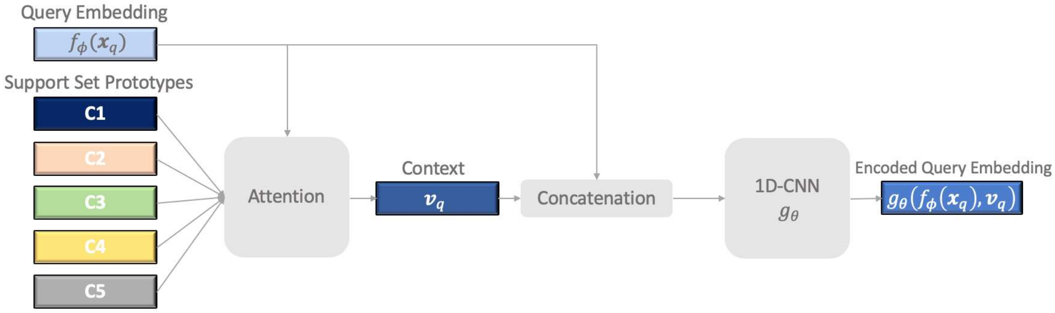

Figure 2. Context-aware query embedding architecture. In this example the query embedding $f_\phi(\mathbf{x}_q)$ and the top prototype $\mathbf{c}_1$ in the support set belong to class 1. Change of blue color of the query embedding shows how the encoder pulls this feature towards the prototype of class one (i.e., $\mathbf{c}_1$) by incorporating the task context in episodes. In general, during training, the non-linear function g_θ learns how to modify the query embedding based on support set context to achieve optimum classification performance.

$$\mathbf{v}_q = \sum_{j=1}^{N_C} a(\mathbf{c}_j, f_\phi(\mathbf{x}_q))\mathbf{c}_j, \tag{10}$$

where a represents the attention weight and d is, again, the Euclidean distance. The more similar a query embedding to a prototype of a class, the larger is the attention weight of that prototype in context vector. The content-based attention has the property that the context vector $\mathbf{v}_q$ will not be sensitive to the order of the prototypes in the support set since it is the weighted sum of them. In other words, the similarity information retrieved from the support set would not change if we randomly shuffle the prototypes in the support set. After calculating the context vector for each query member, $f_\phi(\mathbf{x}_q)$ and $\mathbf{v}_q$ get concatenated and go through a 1-D convolutional block. The convolutional block consists of batch normalization, ReLU activations and pooling. The output of the query encoder is $g_\theta(f_\phi(\mathbf{x}_q), \mathbf{v}_q)$ where θ is the trainable parameter of the encoder (i.e., 1-D CNN). Figure 2 illustrates the details of query embedding module by a toy example of 5-way classification task. The non-linear function g_θ is trained to infer the relationship between query and support set and modify the query feature to increase the discrimination power of the model.

3.4. Zero-Shot Learning and Semi-Supervised Adaptation

In Zero-Shot Learning (ZSL) we are given a class attribute vector $\mathbf{r}_j$ for each class instead of the support set of training data in the few-shot learning setting. In order to have our proposed model to work in zero-shot setting we define the prototype $\mathbf{c}_j = f_{\phi_2}(\mathbf{r}_j)$ to be the embedding of the attribute vector (different from the query embedding f_ϕ), since its modality is different from query images. Classification is performed, as in the few-shot scenario, by finding the nearest class prototype for an embedded query point.

Another capability of the proposed model in this paper is to adapt to semi-supervised classification in testing stage. Specifically, in the semi-supervised scenario, the model needs to adapt to tasks which contain both labeled and unlabeled samples. We assume that we have access to a few labeled examples and many unlabeled examples from the classes in the support set. Since our model is able to generate highly distinguishable feature embeddings in form of separate clusters, unlabeled samples are clustered to the corresponding classes in test time. The prototypes are estimated at test time using the labeled and unlabeled samples and then the query samples are classified based on the nearest prototype. We will show that taking advantage of unlabeled samples can improve the few-shot classification accuracy.

4. Experiments and Results

For fair comparison we follow the same experiment setting as in most recent few-shot learning works [33, 34, 38]. We evaluate our approach on three related tasks: few-shot classification on miniImagenet [38] and Omniglot [17], zero-shot classification on Caltech-UCSD Birds-200-2011 (CUB) [39], and semi-supervised few-shot adaptation on miniImagenet. Following [33, 34, 38], we utilize four convolutional blocks for the embedding module to make the experiments comparable. Specifically, each convolutional block comprises a 64-filter 3 × 3 convolution, batch normalization layer [11], a ReLU nonlinearity, and a 2 × 2 max-pooling layer. When applied to the 28 × 28 Omniglot images this architecture results in a 64-dimensional embedding. We use the same encoder for embedding (i.e., f_ϕ) of both support set and query. The query embedding is modified further by a 1-D CNN in query encoder.

The 1-D CNN in the query embedding module has a convolutional block, batch normalization, and non-linear activation ReLU. As the input tensors are one-dimensional representations of the query images, the convolutional filters are one dimensional of size 1 × 3. Our model is trained end-to-end via SGD with Adam [15]. We use an initial

Table 1. miniImageNet few-shot classification. Results are accuracies averaged over 600 test episodes and with 95% confidence intervals where reported.

Model	**Fine Tune**	**5-way Acc.**	
		1-shot	5-shot
MATCHING NETS [38]	N	43.5%	55.3%
META NETS [22]	N	49.2%	-
MAML [9]	Y	48.7 %	63.1%
META-LEARN LSTM [28]	N	43.4%	63.1%
RELATION NET [34]	N	50.4%	65.3%
EGNN [14]	N	-	66.85%
PROTOTYPICAL NETS [33]	N	49.4%	68.2%
CDFS (ours)	N	**54.7**%	**75.8**%

learning rate of 10^{-3} and cut the learning rate in half every 2000 episodes. We observe that the classification performance of our model remains largely stable across a wide range of small λ values, so we fix it to 0.005. Also, the γ in Eq. 7 is set to 0.01. Optimizing the loss in our model does not need any complex selection of the training samples such as pairs or triplets. Consequently, the learning of our CNN based model is more efficient than methods based on contrastive or triplet loss and is easy to implement. We implement our model using the TensorFlow [1] framework on an Intel Xeon CPU and a NVIDIA TITAN X GPU.

4.1. Few-Shot Learning

We perform experiments for few-shot classification on miniImagenet [38] and Omniglot [17] datasets as follows.

4.1.1 *mini*ImageNet

The *mini*Imagenet dataset, consists of $60,000$ RGB images with 100 classes, each having 600 examples and we resize input images to 84×84. 64, 16, and 20 classes are used for training, validation and testing, respectively. During training, there are 80 and 75 images in one episode of 5-way 1-shot and 5-way 5-shot setting. In fact, the 5-way 1-shot setting contains 15 query images, and 5-way 5-shot setting has 10 query images for each of the N_C classes in each training episode. Few-shot classification accuracies on *mini*Imagenet are shown in Table 1. All results are averaged over 600 test episodes and are reported with 95% confidence intervals. The proposed method achieves state-of-the-art result in both 1-shot and 5-shot settings without any fine-tuning (Table 1).

4.1.2 Omniglot

Omniglot dataset contains 1623 characters (classes) from 50 different alphabets. There are 20 samples in each class, drawn by different people. For this experiment all input images are resized to 28×28. Following previous few-shot classification works, we augment new classes through 90, 180 and 270 rotations of existing data and use 1200 original classes plus rotations for training and remaining 423 classes plus rotations for testing. The few-shot classification accuracy on Omniglot is computed by averaging over 1000 randomly generated episodes from the testing set. During training, the 5-way 1-shot contains 19 query images, the 5-way 5-shot has 15 query images, the 20-way 1-shot has 10 query images and the 20-way 5-shot has 5 query images in each episode. The total number of samples in each episode for different settings during training is show in Table 3. During testing, there are one and five query images per class for the 1-shot and 5-shot experiments, respectively.

The results of 5-way and 20-way classification for 1-shot and 5-shot classification are shown in Table 2. The best-performing methods are highlighted. The proposed method achieves state-of-the-art performance under 20-way experiments setting and competitive results for 5-way classification. For 5-way 5-shot setting almost all methods perform perfectly since it is a rather easy classification task. Since the results for Omniglot dataset are saturated, this dataset is not suitable for evaluation and we just report the results for completeness and focus more on other datasets for the rest of experiments.

4.2. Zero-Shot Classification

We use the Caltech-UCSD Birds (CUB) 200-2011 dataset in order to evaluate our proposed method for zero-shot learning. The CUB dataset contains $11,788$ images of 200 bird species. We divide the classes into 100 training, 50 validation, and 50 test. For images we use 1024-D features extracted by applying GoogLeNet [35] pre-trained on ImageNet. We also augment images using the procedure in [33]. For class attribute for zero-shot setting the 312-dimensional attribute vectors provided with the CUB dataset are used. These attributes encode various characteristics of the bird species such as their color, shape, and feather patterns.

We use an MLP network on top of both the 1024-dimensional image features and the 312-dimensional attribute vectors to produce a 1024-dimensional output space. We normalize the class prototypes to be of unit length, since

Table 2. Omniglot few-shot classification. Results are accuracies averaged over 1000 test episodes and with 95% confidence intervals where reported.

Model	**Fine Tune**	**5-way Acc.**		**20-way Acc.**	
		1-shot	5-shot	1-shot	5-shot
MANN [31]	N	82.8%	94.9%	-	-
CONVOLUTIONAL SIAMESE NETS [16]	N	96.7%	98.4%	88.0%	96.5%
CONVOLUTIONAL SIAMESE NETS [16]	Y	97.3%	98.4%	88.1%	97.0%
MATCHING NETS [38]	N	98.1%	98.9%	93.8%	98.5%
MATCHING NETS [38]	Y	97.9%	98.7%	93.5%	98.7%
SIAMESE NETS WITH MEMORY [13]	N	98.4%	99.6%	95.0%	98.6%
NEURAL STATISTICIAN [6]	N	98.1%	99.5%	93.2%	98.1%
META NETS [22]	N	99.0%	-	97.0%	-
MAML [9]	Y	98.7%	**99.9**%	95.8%	98.9%
RELATION NET [34]	N	99.6%	99.8%	97.6%	99.1%
PROTOTYPICAL NETS [33]	N	98.8%	99.7%	96.0%	98.9%
CDFS (ours)	N	**99.7**%	**99.9**%	**99.2**%	**99.5**%

Table 3. Number of samples in episodes in different few-shot classification setting for Omniglot dataset during training.

Experiment	num. of queries	num. of support set samples	num. episode samples
5-way 1-shot	19	5	100
5-way 5-shot	15	25	100
20-way 1-shot	10	20	220
20-way 5-shot	5	100	200

Table 4. Zero-shot classification accuracies on CUB-200.

Model	**Feature Ext.**	**50-way Acc.**
SJE [2]	GoogLeNet	50.1
ESZSL [30]	GoogLeNet	47.2
SSE-RELU [40]	VGG-19	30.4
JLSE [41]	VGG-19	42.1
SYNC-STR[4]	GoogLeNet	54.5
SEC-ML [3]	VGG-19	43.3
REL. NET [34]	N-GoogLeNet	62.0
PROTO.NETS [33]	GoogLeNet	54.6
CDFS (ours)	GoogLeNet	**58.1**
CDFS (ours)	N-GoogLeNet	**65.1**

the attribute vectors come from a different modality than the images. Training episodes were constructed with 50 classes and 10 query images per class. The embeddings were optimized via SGD with Adam at a fixed learning rate of 10^{-4}. The result of zero-shot learning is shown in Table 4. The second column demonstrate the type of feature extractor used for extracting image features. [34] uses a different feature extractor (N-GoogLeNet). In order to make the results comparable, we also employ the same backbone and experiment setting as in [34] and report the result in the last row of Table 4. As it can be observed from Table 4, our proposed method is the best performing approach in zero-shot setting.

4.3. Semi-supervised Adaptation

We assume that we have access to a few labeled examples (i.e., five example per class) and many unlabeled examples from the same classes in the support set. Since our model is able to generate highly distinguishable feature embeddings in form of separate clusters, unlabeled samples are clustered to the corresponding classes in test time. The prototypes are estimated at test time using the labeled and unlabeled samples and then the query samples are classified based on the nearest prototype. We use *mini*Imagenet for this experiment and the 5-way 1-shot setting contains 15 query images, and 5-way 5-shot setting has 10 query images for each of the classes in each training episode. Table 5 shows how the number of unlabeled examples at test time affects the classification accuracy of the trained model. The results indicate that more unlabeled samples yield better performance, however, experiments show that with increasing the number of unlabeled samples over 40 samples, the improvement plateaus.

Table 5. 5-way testing accuracy using CDFS method on *mini*Imagenet for the semi-supervised scenario for different number of unlabeled samples per class (n).

n	**1-shot**	**5-shot**
5	55.1	76.6
10	55.9	77.3
20	57.2	78.8
40	58.9	79.1

4.4. Ablation Study

In order to evaluate the effect of context-aware query encoder and the $S3$ loss, we perform the following ablation study. The first experiment setting is training and testing the model without using the query encoder which we denote by CDFS-NoQE. The second scenario is to remove the $S3$ loss during training which we denote by CDFS-NoS3. For this experiment the *mini*Imagenet data set is used in 5-way 5-shot setting and all the experimental parameters are the same as the previous experiments. It can be observed from Table 6 that removing either the $S3$ regularization or the query encoder causes the performance to drop, since it reduces the discriminative power of the model. However, removing the $S3$ loss has more negative effect on accuracy and causes a significant drop in accuracy which shows the importance of this regularization in performance of the model. Table 6 confirms the effectiveness of the proposed query and support set feature encoding frameworks *specially when they are used together.*

Table 6. Ablation study to evaluate the effect of S3 loss and query encoder in the CDFS model on miniImagenet.

Model	**miniImagenet (5-way 5-shot)**
CDFS-NoQE	72.6
CDFS-NoS3	70.0
CDFS (full)	**75.8**

5. Conclusion

In this paper, we introduced a simple but effective few-shot learning model which can produce highly discriminative embedding space using the combination of proposed query and support set feature manipulation frameworks. By removing the softmax loss and defining each episode as one set without a query, the proposed approach can be considered as a few-shot clustering method which learns a deep non-linear metric in order to learn to cluster the data in few-shot setting. The future work is to extend the proposed framework to unsupervised few-shot classification by following the idea of *learning to cluster* proposed in this work.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th Symposium on Operating Systems Design and Implementation 16)*, pages 265–283, 2016.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[3] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, pages 730–746. Springer, 2016.

[4] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[5] S. Chopra, R. Hadsell, Y. LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546, 2005.

[6] H. Edwards and A. Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.

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

[8] T. Finley and T. Joachims. Training structural svms when exact inference is intractable. In *Proceedings of the 25th international conference on Machine learning*, pages 304–311. ACM, 2008.

[9] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[10] S. Gidaris and N. Komodakis. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[12] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[13] Ł. Kaiser, O. Nachum, A. Roy, and S. Bengio. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*, 2017.

[14] J. Kim, T. Kim, S. Kim, and C. D. Yoo. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2019.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[17] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[18] K. Lee, S. Maji, A. Ravichandran, and S. Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019.

[19] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2019.

[20] H. Lin and J. A. Bilmes. Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871*, 2012.

[21] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. Meta-learning with temporal convolutions. *arXiv preprint arXiv:1707.03141*, 2(7), 2017.

[22] T. Munkhdalai and H. Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563. JMLR. org, 2017.

[23] H. Oh Song, S. Jegelka, V. Rathod, and K. Murphy. Deep metric learning via facility location. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5382–5390, 2017.

[24] B. Oreshkin, P. R. López, and A. Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 719–729, 2018.

[25] F. Pahde, O. Ostapenko, P. J. Hnichen, T. Klein, and M. Nabi. Self-paced adversarial training for multimodal few-shot learning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 218–226. IEEE, 2019.

[26] M. Pugelj and S. Džeroski. Predicting structured outputs k-nearest neighbours method. In *International Conference on Discovery Science*, pages 262–276. Springer, 2011.

[27] A. Rahimpour, L. Liu, A. Taalimi, Y. Song, and H. Qi. Person re-identification using visual attention. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4242–4246. IEEE, 2017.

[28] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. 2016.

[29] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[30] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[31] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

[32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[33] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[34] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. 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.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[36] S. Tschiatschek, R. K. Iyer, H. Wei, and J. A. Bilmes. Learning mixtures of submodular functions for image collection summarization. In *Advances in neural information processing systems*, pages 1413–1421, 2014.

[37] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104. ACM, 2004.

[38] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[39] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *2011 International Conference on Computer Vision*, pages 2524–2531. IEEE, 2011.

[40] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174, 2015.

[41] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016.