PROJECTIVE SUBSPACE NETWORKS FOR FEW SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalization from limited examples, usually studied under the umbrella of meta-learning, equips learning techniques with the ability to adapt quickly in dynamical environments and proves to be an essential aspect of lifelong learning. In this paper, we introduce the *Projective Subspace Networks (PSN)*, a deep learning paradigm that learns non-linear embeddings from limited supervision. In contrast to previous studies, the embedding in PSN deems samples of a given class to form an affine subspace. We will show that such modeling leads to robust solutions, yielding competitive results on supervised and semi-supervised few-shot classification. Moreover, our PSN approach has the ability of end-to-end learning. In contrast to previous works, our projective subspace can be thought of as a richer representation capturing higher-order information datapoints for modeling new concepts.

1 INTRODUCTION

Supervised learning with deep architectures, though achieving remarkable results in many areas, requires large amount of annotated data. Various studies show that many deep learning techniques in computer vision, speech recognition and natural language understanding, to name a few methods stated in Hinton et al. (2012); Krizhevsky et al. (2012), will fail to produce reliable models that generalize well if limited data annotations are available. Aside from the labor associated with annotating data, precise annotation can become ill-posed in some cases. One prime example of such a difficulty is object detection labeling requires annotating bounding boxes of objects as explained in Alexe et al. (2012).

In contrast to the current trend in deep learning, humans can learn new concepts from only a few examples. This in turn provides humans with lifelong learning abilities. Inspired by such learning abilities, several approaches are developed to study learning from limited examples Lake et al. (2015); Lazaridou et al. (2017); Vinyals et al. (2016); Triantafillou et al. (2017); Xu et al. (2017); Finn et al. (2017); Ravi & Larochelle (2017); Wang et al. (2018); Mishra et al. (2018); Qiao et al. (2018); Neill & Buitelaar (2018). In machine learning, the diverse ideas in this context include embedding features through metric learning (Koch et al. (2015), Vinyals et al. (2016)), optimization technique(Finn et al. (2017), Ravi & Larochelle (2017)), and generative models(Fei-Fei et al. (2006), Lake et al. (2015)).

In this work, we propose a deep model that learns new concepts from limited data to address two challenging learning problems; namely: (**i**) few-shot classification and (**ii**) semi-supervised few-shot learning. The goal of few-shot classification is to learn a model that can discriminate a given query by comparing to a few of samples (a.k.a. the support set). An example is to classify a motorcycle by viewing some different types of motorcycles (or other vehicles). The goal of semi-supervised few-shot learning is to additionally benefit from unlabeled data to boost the performance of the model.

Our method, coined Projective Subspace Networks or PSN for short, learns non-linear embeddings using subspaces, in a sense that samples of a class are modeled as a low-dimensional affine subspace. The use of subspaces to model images and sets has a long history in computer vision and machine learning. For example, it has been proved that the set of all reflectance functions (the mapping from surface normals to intensities) produced by Lambertian objects lie close to a low-dimensional linear subspace(Basri & Jacobs (2003)).

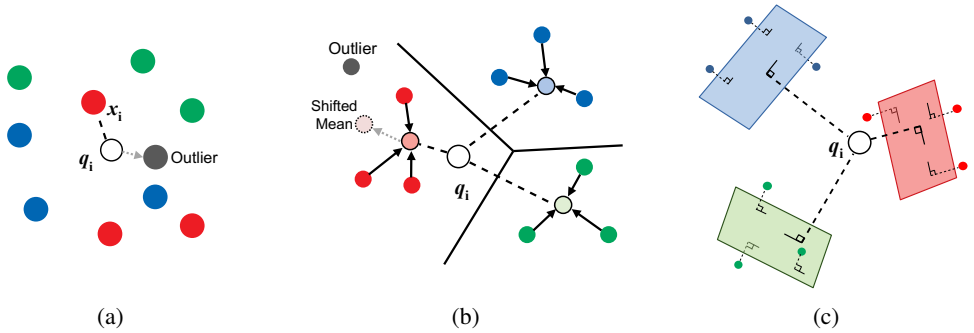

Figure 1: Feature embedding in (a) Matching Networks (Vinyals et al. (2016)), (b) Prototypical Networks(Snell et al. (2017)), and (c) our PSN method. For both Matching and Prototypical Networks, an outlier (shown in gray) can lead to misclassification. This is because in the Matching Networks, pair-wise similarities are used for inference. Prototypical Networks improve upon Matching Networks by making use of class-specific averages to reduce the effect of outliers. Nevertheless, class averages are not immune to outliers. In contrast, in our proposal, each class is represented by a low-dimensional subspace. Our empirical evaluations suggest that such modeling is more discriminative and robust.

In spite of its intriguing properties, to the best of our knowledge, the subspace modeling has never been used to address few-shot learning problems previously. This makes our paper unparalleled to previous studies (*e.g.*, Vinyals et al. (2016); Snell et al. (2017)) where non-linear embeddings shaped the core of the solution (see Fig. 1 for a conceptual comparison).

We empirically observed that embeddings tailored towards capturing the structure of each class through low-dimensional affine subspaces could lead to discriminative models. Interestingly, such models can be built with minimum overheads and without opting for advanced methods in subspace creation (*e.g.*, such as the notion of sparsity). Our conjecture here is that subspaces are less sensitive to outliers as compared to other embedding techniques for few-shot learning.

Our contributions in this work are:

i. Few-shot learning is formulated as an embedding problem through subspaces. We rely on a well-established concept stating that samples of a class (and hence variations such as pose and illumination) can be effectively captured by a low-dimensional affine space.

ii. Adaptation from few-shot learning to semi-supervised learning is performed with a refinement through soft-assignment and its robustness is reflected in our experiments.

iii. We also introduce an evaluation mechanism to assess the generalization ability over unseen classes.

2 Related Work

In this section, we briefly review the literature on few-shot learning and subspace clustering. Few-shot learning was originally introduced to imitate human learning capabilities in classification. Some of the early works made use of generative models and similarity learning to capture the variation within parts and geometric configurations of objects(Fei-Fei et al. (2006); Lake et al. (2015); Torralba et al. (2007)). As of late, few-shot problems are mainly addressed through meta-learning techniques. For example, in Koch et al. (2015); Vinyals et al. (2016); Snell et al. (2017); Garcia & Bruna (2018), the problem of few-shot learning is formulated as non-linear embedding or representation learning.

The closest approach to our proposed work is the Prototypical Networks model(Snell et al. (2017)). It uses the random choice of episode images which form the prototype center per class. In addition, recent metric learning approaches use complex architectures and pipelines in convolutional networks by Gidaris & Komodakis (2018) and Wang et al. (2018). In contrast, our projective subspace concept is extremely simple by design as it constitutes just a single layer of the network.

Another common meta-learning approach to few-shot classification uses meta-learning and manipulates gradient updates to train the model parameters. Ravi & Larochelle (2017) utilized Long-Short Term Memory (LSTM)-based learner to optimize the model parameters and their gradients. Moreover, the Model Agnostic Meta-Learner(MAML) proposed by Finn et al. (2017) used gradients per task as well as a meta-gradient from combined tasks, and, as a result, it outperformed LSTM-based learner. Mishra et al. (2018) employed ResNet to model few-shot classification as a temporal solution with aggregation, thus, in meta-learning context, their model can refer to the past experience.

Our proposed work is also related to clustering methods such as K-means, K-modes, and K-prototypes which are commonly used in grouping representations. However, they preserve the information in dimensions that may be irrelevant for our task. Thus, we project the feature representations into low-dimensional spaces. Feature transformations such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) are typically employed to reduce dimensionality of feature representations *e.g.*, in neural networks (Chan et al. (2015); Sun et al. (2017)). The literature of subspace for clustering analysis and classification is immense, thus, we present only works closely related to our paper. The subspace clustering is an effective tool that can group high-dimensional data as explained by Kriegel et al. (2009) in their review. Moreover, the recent progress of subspace clustering methods such as the works by Ji et al. (2017) and Peng et al. (2016) has opened the possibility to extend subspace clustering to end-to-end training of neural networks. In order to perform classification by the use of subspaces, some proposed methods calculate the distance from the hyperplanes to the query point as proposed by Basri et al. (2011) and Vidal (2011) which can be seen as a variant of K-nearest neighbors.

3 Problem Set-Up

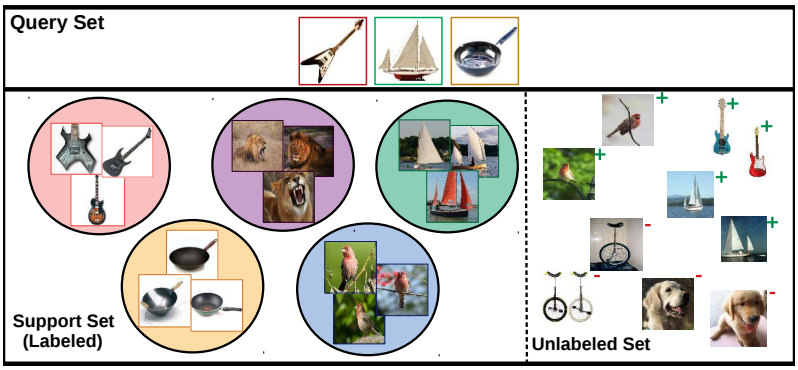

Figure 2: A visual representation of an episode consists of the support set (labeled) S and the query set Q. The samples from the same class are represented as a group inside the circle. The query set is sampled from the same classes as the support set. In semi-supervised few-shot setting, the support set is supplemented with unlabeled data $\mathcal{R}$ composed of samples from the *distractors* classes as well as the classes of the labeled data. The plus and minus signs describe the unlabeled data from the classes in the support set and *distractors*, respectively.

We start by defining the terminology used in few-shot learning. The problem of N-way K-shot classification (*e.g.*, 5-way 1-shot) is defined as classifying queries belonging to N classes by seeing only K samples from each class. For example, in 5-way 1-shot setting, the model needs to identify a query among five classes by seeing only one sample from each class. To obtain a reliable model, so-called *episodes* are used in training. An episode $\mathcal{T}_i$ consists of two sets, the support set S and the query set Q. The system is then trained by minimizing a classification loss over episodes, simulating the scenarios it will encounter at test time. This episode setting is the same as proposed by Vinyals et al. (2016).

A related problem is semi-supervised few-shot setting learning where unlabeled data is provided to the model. In the literature, various configurations are considered for semi-supervised few-shot learning (*e.g.*, Garcia & Bruna (2018); Boney & Ilin (2017); Ren et al. (2018)). In this work, we

follow the challenging protocol in Ren et al. (2018) where so-called *distractors* are introduced. Here, an episode includes the support set S, query set Q, and unlabeled set $\mathcal{R}$. The support (labeled) S and query Q sets are configured as in few-shot learning. Additionally, an unlabeled set $\mathcal{R}$ is provided to assist the classification task within an episode. In the unlabeled set, there are examples from two different sources: the support classes and the *distractor* classes. As the name implies, examples from *distractor* classes are irrelevant to the classification task and represent classes outside the support set.

4 Projective Subspace Networks

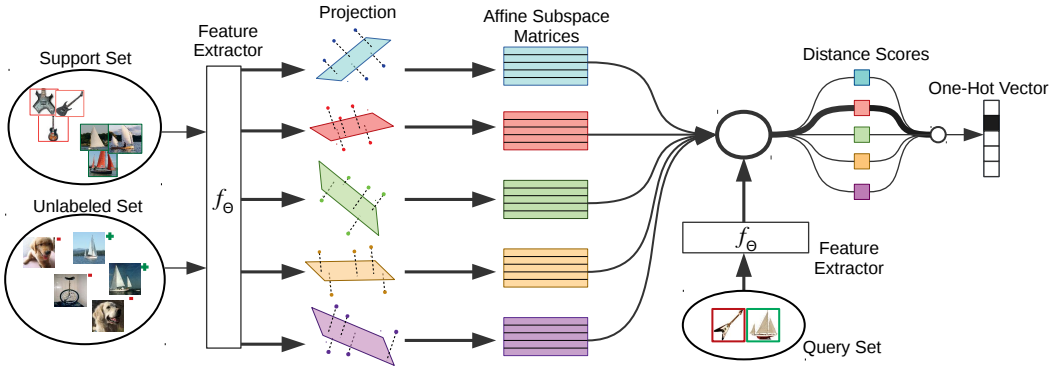

Figure 3: Projective Subspace Networks Architecture

In what follows, we introduce our PSN approach. The whole framework is trained end-to-end (see Fig. 3) which enables PSN to be used with various deep architectures. As an example, in § 5, we use PSN on top of the WideResNets(Zagoruyko & Komodakis (2016)) for few-shot classification.

We start by introducing our notations for the (N-way, K-shot) few-shot learning. Each episode or task $\mathcal{T}_i$ is composed of the support set $S = \{(\boldsymbol{x}_{1,1}, c_{1,1}), (\boldsymbol{x}_{1,2}, c_{1,2}), \cdots, (\boldsymbol{x}_{N,K}, c_{N,K})\}$ and the query set $Q = \{\boldsymbol{q}_1, \cdots, \boldsymbol{q}_{N\times M}\}$. Here, $\boldsymbol{x}_{i,j}$ denotes the j-th sample from class i and $c_{i,j} \in \{1, \cdots, N\}$. In the semi-supervised setting, there is an unlabeled set $\mathcal{R} = \{\boldsymbol{r}_1, .., \boldsymbol{r}_U\}$ within an episode. We propose to model points by subspaces $\{\boldsymbol{Z}_i\}_{i=1}^N$. Each subspace $\boldsymbol{Z}_i$ has a basis represented by $\mathbb{R}^{D\times n} \ni \boldsymbol{P}_i = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_n]; n \leq D$, with $\boldsymbol{P}_i^\top \boldsymbol{P}_i = \mathbf{I}_n$.

4.1 PSN for Few-Shot Classification

Let $f_\Theta : \mathcal{X} \to \mathbb{R}^D$ be a mapping from the input space $\mathcal{X}$ to some D-dimensional representation realized by a neural network. Our goal is to learn Θ, *i.e.*, the embedding function in a way that the resulting space is suitable for subspace representation. For simplicity, we assume that every class in an episode can be described by just one subspace. Extension to multiple subspaces per class is straightforward though. Define $\boldsymbol{\mu}_k$ as the mean of class k in the embedded space. That is,

$$\boldsymbol{\mu}_k = \frac{1}{K} \sum_{i,\ c_i=k} f_\Theta(\boldsymbol{x}_i). \tag{1}$$

A basis for the subspace representing class k can be obtained by Singular Value Decomposition (SVD). To be specific, we define $\boldsymbol{X}_k = [\boldsymbol{x}_{k,1} - \boldsymbol{\mu}_k, \cdots, \boldsymbol{x}_{k,K} - \boldsymbol{\mu}_k]$. Applying truncated SVD on $\boldsymbol{X}_k$ provides us with $\boldsymbol{P}_k$. We emphasize that more involved techniques to obtain robust subspaces from $\boldsymbol{X}_k$ can potentially improve the PSN. Nevertheless, our goal is to assess whether the concept of subspace modeling for few-shot learning is justified or not and thus we opt for truncated SVD in our implementation. Now a query $\boldsymbol{q}_j$ can be projected onto $\boldsymbol{P}_k$ which yields:

$$\boldsymbol{y}_{j,k} = \boldsymbol{P}_k^\top (f_\Theta(\boldsymbol{q}_j) - \boldsymbol{\mu}_k). \tag{2}$$

The distance from the query to $\boldsymbol{P}_k$ is:

$$d_{j,k} = \|f_\Theta(\boldsymbol{q}_j) - \boldsymbol{\mu}_k - \boldsymbol{P}_k \boldsymbol{y}_{j,k}\|. \tag{3}$$

We define the probability of the query assigned to class k using a softmax function as:

$$p_\Theta(c = k|\boldsymbol{q}_j) = \frac{\exp\big(-d^2_{j,k}\big)}{\sum_{k'} \exp\big(-d^2_{j,k'}\big)}. \tag{4}$$

Now, we can minimize the negative log of Eqn. 4 to obtain Θ. To train the whole framework, backpropagation through SVD is required which is available in modern deep learning packages such as PyTorch(Paszke et al. (2017)). Algorithm 1 explains the steps of training our PSN. The code will be released online on Github[1].

Algorithm 1 Train Projective Subspace Networks

Input: Each episode $\mathcal{T}_i$ with $S = \{(\boldsymbol{x}_{1,1}, c_{1,1}), \cdots, (\boldsymbol{x}_{N,K}, c_{N,K})\}$ and $Q = \{\boldsymbol{q}_1, ..., \boldsymbol{q}_{N\times M}\}$

$\Theta_0 \leftarrow$ random initialization

for t in $\{\mathcal{T}_1, ..., \mathcal{T}_{N_\mathcal{T}}\}$ **do**

  $\mathcal{L}_t \leftarrow 0$

  for k in $\{1, ..., N\}$ **do**

    $\boldsymbol{X} \leftarrow \boldsymbol{S}_k$ ▷ Get examples in the support set from class k

    $\boldsymbol{\mu}_k \leftarrow \frac{1}{K}\sum_{\boldsymbol{x}\in\boldsymbol{X}} f_\Theta(\boldsymbol{x})$ ▷ Mean from the support set

    $\boldsymbol{\mu}_k \leftarrow \dfrac{K\boldsymbol{\mu}_k + \sum_i m_i f_\Theta(\boldsymbol{r}_i)}{K + \sum_i m_i}$ ▷ Refined mean(only semi-supervised learning)

    $\tilde{\boldsymbol{X}} \leftarrow [\boldsymbol{x}_i - \boldsymbol{\mu}_k, ..., \boldsymbol{x}_K - \boldsymbol{\mu}_k]$

    $[\mathcal{U}, \Sigma, \mathcal{V}^\top] \leftarrow \text{SVD}(\tilde{\boldsymbol{X}})$ ▷ Matrix factorization using SVD

    $\boldsymbol{P}_k \leftarrow \mathcal{U}_{1,\dots,n}$ ▷ Truncate the matrix

    for $\boldsymbol{q_j}$ in Q_k **do**

      $\boldsymbol{y}_{j,k} \leftarrow \boldsymbol{P}_k^\top(f_\Theta(\boldsymbol{q_j}) - \boldsymbol{\mu}_k)$ ▷ Query projection

      $d_{j,k} \leftarrow \|f_\Theta(\boldsymbol{q_j}) - \boldsymbol{\mu}_k - \boldsymbol{P}_k\boldsymbol{y}_{j,k}\|$ ▷ Distance calculation

      $p_{j,k} \leftarrow \dfrac{\exp(-d^2_{j,k})}{\sum_{k'}\exp(-d^2_{j,k'})}$ ▷ Softmax on distance scores

    end for

  end for

  $\mathcal{L}_t \leftarrow \frac{1}{N^2M}\sum_k\sum_j -\log\left(p_{j,k}\right)$

  Update Θ using $\nabla_\Theta\mathcal{L}_t$

end for

4.2 PSN FOR SEMI-SUPERVISED FEW-SHOT LEARNING

In what follwos, we extend the model developed in § 4.1 to address semi-supervised few-shot learning. In doing so, we need to take advantage of the unlabeled data to fit better subspaces to our data. We achieve this by refining the center of each class according to

$$\tilde{\boldsymbol{\mu}}_k = \frac{K\boldsymbol{\mu}_k + \sum_i m_i f_\Theta(\boldsymbol{r_i})}{K + \sum_i m_i}\ , \tag{5}$$

with

$$m_i = \frac{\exp(-\|f_\Theta(\boldsymbol{r}_i) - \boldsymbol{\mu}_k)\|^2)}{\sum_{k'}\exp(-\|f_\Theta(\boldsymbol{r}_i) - \boldsymbol{\mu}_{k'})\|^2)}\ , \tag{6}$$

where m_i is the soft-assignment score for unlabeled samples. To work at the presence of *distractors*, we propose to use a fake class with zero mean. We empirically observed that such a simple modification to the means can improve the results without the need of refining the SVD step.

[1] https://github.com/anonymousauthor/PSN

Models	5-way 5-shot	20-way 5-shot
Matching Nets (Vinyals et al. (2016))	$55.31 \pm 0.73\%$	$22.69 \pm 0.20\%$
MAML (Finn et al. (2017))	$63.11 \pm 0.92\%$	$19.29 \pm 0.29\%$
Optimization FS (Ravi & Larochelle (2017))	$60.60 \pm 0.71\%$	$26.06 \pm 0.25\%$
Meta-SGD (Li et al. (2017))	$64.03 \pm 0.94\%$	$28.92 \pm 0.35\%$
Prototypical Nets (Snell et al. (2017))	$65.49 \pm 0.25\%$	$37.23 \pm 0.21\%$
PSN	$\mathbf{66.62 \pm 0.69\%}$	$\mathbf{38.26 \pm 0.23\%}$

Table 1: Few-shot classification results on Mini-ImageNet using 4-convolutional stages with 95% confidence intervals.

5 Experiments

Below we contrast and assess our method against state-of-the-art techniques on two challenging datasets, namely Mini-ImageNet(Ravi & Larochelle (2017)) and Tiered-ImageNet(Ren et al. (2018)).

Mini-ImageNet. The Mini-ImageNet(Ravi & Larochelle (2017))[2] contains 60,000 images of the ImageNet(Russakovsky et al. (2015)) datasets. Images in the Mini-ImageNet are of size 84×84 and represent 100 classes with 64, 16, and 20 classes being used for training, validation, and testing, respectively.

Tiered-ImageNet. This dataset is also derived from ImageNet but contains a broader set of classes compared to the Mini-ImageNet. There are 351 classes from 20 different categories for training, 97 classes from 6 different categories for validation, and 160 classes from 8 different categories for testing. In contrast to the Mini-ImageNet, the training and test sets in Tiered-ImageNet represent distinct classes. Moreover, in designing the Tiered-ImageNet, the problem of few-shot learning with unlabeled data was taken into account and the labeled data is only within a small percentage when performing semi-supervised learning. In this large dataset, learning from the labeled data is still sufficient to produce reasonable representations even though the unlabeled data is set in huge portion *e.g.*, 90%.

5.1 Few-shot Learning

We follow the general practice and evaluate our method on the Mini-ImageNet dataset when it comes to few-shot learning and classification. Various procedures such as pre-training using 64 classes (*e.g.*, Qiao et al. (2018)) and training with more classes/way have shown to increase the overall accuracy. Nevertheless, we avoid such procedures deliberately as we are mainly interested in contrasting the core idea, *i.e.*, the role of subspaces in few-shot learning. So, we trained on 5-way 5-shot and applied the same classification task setup during testing. The CNN architecture is the same as the one used in Snell et al. (2017) with 4-convolutional stages. We also use WideResNets(Zagoruyko & Komodakis (2016)) with 16 depth, 6 widening factor, and 0.3 dropout rate to compare with ResNets(He et al. (2016)) solution reported by Mishra et al. (2018). The feature dimensions from both architectures are 1600 and 384 respectively. We used ADAM(Kingma & Ba (2015)) for optimizing our model and set the learning rate to 0.001 and cut it to half every 2K episodes for both architectures.

Results. By design, the method cannot accommodate learning with exactly one example, hence, the comparison here is provided only for 5-shot in 5-way and 20-way. In every episode, the query set contains 15 samples from each class. Our method outperforms the previous methods with 4-convolutional stages shown in Table 1. We also implemented WideResNets(Zagoruyko & Komodakis (2016)) using our method and obtained the performance for 5-way 5-shot: $\mathbf{69.92 \pm 0.64}\%$ and 20-way 5-shot: $\mathbf{41.84 \pm 0.24}\%$ that can outperform ResNets solution proposed by Mishra et al. (2018) with $68.88 \pm 0.92\%$ in 5-way and 5-shot classification task.

[2]We note that another version of Mini-ImageNet dataset was considered in Vinyals et al. (2016). But the split from Ravi & Larochelle (2017) has been widely used for performance measurements because the Mini-ImageNet split by Vinyals et al. (2016) had not been available previously.

Models	Mini-ImageNet 5-way 5-shot		Tiered-ImageNet 5-way 5-shot	
PN, *Supervised*	$59.08 \pm 0.22\%$		$66.15 \pm 0.22\%$	
PSN, *Supervised*	$\mathbf{63.43 \pm 0.61}\%$		$\mathbf{68.72 \pm 0.49}\%$	
	w/o *Distractors*	**w/ *Distractors***	**w/o *Distractors***	**w/ *Distractors***
PN-SSL, Non-Masked	$64.59 \pm 0.28\%$	$63.55 \pm 0.28\%$	$70.25 \pm 0.31\%$	$68.32 \pm 0.22\%$
PN-SSL, Masked	$64.39 \pm 0.24\%$	$62.96 \pm 0.14\%$	$69.88 \pm 0.20\%$	$69.08 \pm 0.25\%$
PSN, Semi-Supervised	$\mathbf{68.12 \pm 0.67}\%$	$\mathbf{66.10 \pm 0.66}\%$	$\mathbf{71.15 \pm 0.67}\%$	$\mathbf{69.15 \pm 0.51}\%$

Table 2: Semi-supervised few-shot classification results on the Mini-ImageNet and Tiered-ImageNet with 40% and 10% labeled data, respectively. We show the classification results with and without *distractors*. We compare our results to Prototypical Networks on semi-supervised learning (PN-SSL) with soft K-means (non-masked) and masked K-means (masked), as proposed by Ren et al. (2018).

5.2 Semi-Supervised Few-shot Learning

In this experiment, the embedding architecture has 4-convolutional layers as Prototypical Networks(Snell et al. (2017)). The episode composition for labeled or support set and query set is similar to the few-shot learning classification task, but there is an additional unlabeled set provided in each episode. Our model is trained on 100K episodes for Mini-ImageNet and Tiered-ImageNet with 40% and 10% of labeled data, respectively. We used the ADAM solver(Kingma & Ba (2015)), the set the learning rate to 0.001 with the weight decay and cut the rate to half every 10K episodes. We trained in two settings: (i) $supervised$ setting, where only labeled data is taken into account, and (ii) semi-supervised setting for which the unlabeled set is also used. The unlabeled set is composed of the examples from the classes in the support set and *distractor* classes. The number of supporting classes and *distractor* classes is set to five for training and testing. In the training stage, the number of examples in the unlabeled set is 50 consisting of five examples from each class. In the testing stage, the unlabeled set consists of 20 examples from each class. We also define the query set to have 20 examples per class for testing purpose.

Results. In these experimental results, the performance is counted over 600 episodes. The results are averaged over 10 random splits of labeled and unlabeled sets. The *supervised* experiment shows that our method learns robust feature embedding from a small portion of labeled data.

With the help of soft-assignment over unlabeled datapoints, the *semi-supervised* experiment detailed in Table 2 is demonstrated to outperform Prototypical Networks for Semi-Supervised Learning (SSL) proposed by Ren et al. (2018).

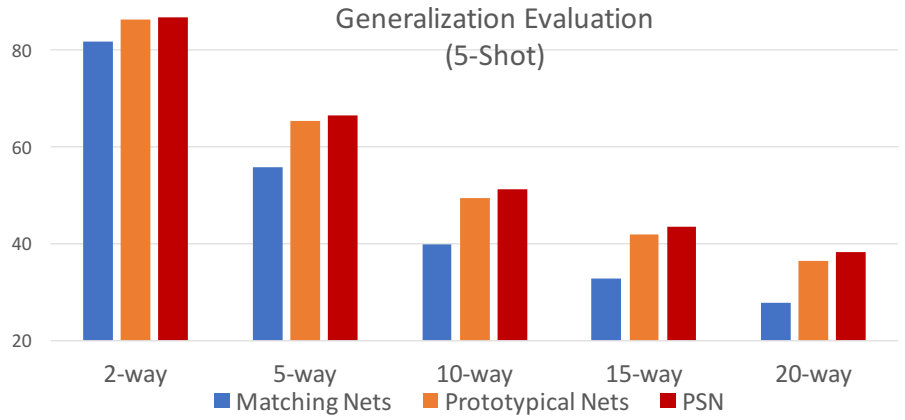

Figure 4: The generalization evaluation on Mini-ImageNet for 2-, 5-, 10-, 15- and 20-way (trained in 5-way 5-shot setting).

5.3 EVALUATION ON GENERALIZATION

In what follows, we evaluate few-shot learning in terms of its capacity to generalize on large number of unseen classes. Transfer learning approaches are typically trained and tested with the equal numbers of classes. Below, we assess whether the model trained on low number of classes can generalize well to classification tasks involving large number of classes. In the evaluation below, the model is trained in 5-way 5-shot setting. For the testing stage, we set the number of examples to 5 (5-shot setting) and evaluate the model given 2-way, 5-way, 10-way, 15-way, 20-way setting on the Mini-ImageNet test set. We argue that this is a more realistic evaluation scenario as the number of classes typically increases as new data arrives. In Figure 4, we compare our method with Prototypical Networks and Matching Networks. The above plot shows that the gap between the MatchingNet and the PSN increases in favour of the PSN as we increase K of K-way setting.

6 DISCUSSION

We have compared our PSN approach to Prototypical Networks proposed by Snell et al. (2017) as their architecture is related to ours, *i.e.*, they make use of the support set to obtain a refined representation. Moreover, the Prototypical Network makes use of the class means and can be easily incorporated in our testbed.

However, our method proposes a non-linear embedding for which the features from the same class span a unique subspace. Therefore, we argue that PSN, capturing higher-order information about a given class, is more robust than representing a given class by the mean of datapoints (first-order representation).

Even though PSN involves SVD in the training stage, our method is fast to run and easy to implement. We note that the number of parameters to learn is equal to the number of parameters of the Prototypical Networks. Our approach can be easily combined with more complicated design choices to obtain better results or solve a variety of few-shot learning problems.

Computational Complexity. Below we compare the complexity of our approach and Prototypical Networks. For Prototypical Networks, the complexity of running their algorithm amounts to $\mathcal{O}(DK)$, where K is the number of samples and D is the feature dimensionality. Our PSN approach is somewhat slower given the full-rank SVD for which the complexity amounts to $\mathcal{O}(\min(D^2K, DK^2))$. Thus, the speed of PSN will be somewhat affected if the number of training samples K is large. However, fast approximate algorithms for SVD, *i.e.*, described by Menon & Elkan (2011), can be used in place of the full SVD.

7 CONCLUSIONS

This paper presents the feature embedding via projection into subspaces called Projective Subspace Networks. PSN can adapt quickly to incoming tasks and update the model parameters accordingly. This meta-learning design is not dedicated to a specific neural networks architecture. Thus, our approach is applicable to a wide-range of problems.

The representations learned via PSN are highly expressive as demonstrated by few-shot supervised and semi-supervised experiments. Each class is represented by a unique subspace which contains higher-order class information compared to a mere mean-based class representation (first-order representation) akin to Prototypical Networks. Moreover, we speculate that classification with the use of subspaces is more robust to outliers than the mean-based representation.

The experiments show that our method generalizes well to larger number of samples K and attains the state-of-the-art results. In semi-supervised setting, PSN also achieves notable performance with as little as 40% and 10% labeled data for Mini-ImageNet and Tiered-ImageNet, respectively. Our experiments show that PSN benefits from unlabeled data which increases the final classification performance.

REFERENCES

Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 34:2189–2202, 2012.

R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 25:218–233, 2003.

Ronen Basri, Tal Hassner, and Lihi Zelnik-Manor. Approximate nearest subspace search. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 33:266–278, 2011.

Rinu Boney and Alexander Ilin. Semi-supervised few-shot learning with prototypical networks. *arXiv preprint arXiv:1711.10856*, 2017.

Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing(TIP)*, 24:5017–5032, 2015.

Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 28:594–611, 2006.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning(ICML)*, 2017.

Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. In *International Conference on Learning Representations(ICLR)*, 2018.

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems(NIPS)*, 2017.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations(ICLR)*, 2015.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning Deep Learning 2015 Workshop*, 2015.

Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data(TKDD)*, 3:1:1–1:58, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems(NIPS)*, pp. 1097–1105, 2012.

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

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, 41 Suppl 4:677–705, 2017.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

Aditya Krishna Menon and Charles Elkan. Fast algorithms for approximating the singular value decomposition. *ACM Trans. Knowl. Discov. Data(TKDD)*, 5:13:1–13:36, 2011.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations(ICLR)*, 2018.

James O' Neill and Paul Buitelaar. Few shot transfer learning betweenword relatedness and similarity tasks using a gated recurrent siamese network. 2018.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*, 2017.

Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *International Joint Conference on Artificial Intelligence(IJCAI)*, pp. 1925–1931, 2016.

Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L. Yuille. Few-shot image recognition by predicting parameters from activations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations(ICLR)*, 2017.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2018.

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

Jake Snell, Kevin Swersky, and Zemel Richard. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems(NIPS)*, pp. 4077–4087, 2017.

Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *International Conference on Computer Vision(ICCV)*, pp. 3820–3828, 2017.

A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI)*, 29: 854–869, 2007.

Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. In *Advances in Neural Information Processing Systems(NIPS)*, pp. 2255–2265, 2017.

R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28:52–68, 2011.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems(NIPS)*, pp. 3630–3638, 2016.

Yu-Xiong Wang, Ross Girshick, Martial Herbert, and Bharath Hariharan. Low-shot learning from imaginary data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Z. Xu, L. Zhu, and Y. Yang. Few-shot object recognition from machine-labeled web images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference(BMVC)*, 2016.