TAFSSL: Task-Adaptive Feature Sub-Space Learning for few-shot classification

Moshe Lichtenstein Prasanna Sattigeri Rogerio Feris Raja Giryes Leonid Karlinsky

Abstract

The field of Few-Shot Learning (FSL), or learning from very few (typically 1 *or* 5*) examples per novel class (unseen during training), has received a lot of attention and significant performance advances in the recent literature. In this paper we propose yet another simple technique that is important for the few shot learning performance - a search for a compact feature sub-space that is discriminative for a given few-shot test task. We show that the Task-Adaptive Feature Sub-Space Learning (TAFSSL) can significantly boost the performance in FSL scenarios when some additional unlabeled data accompanies the novel few-shot task, be it either the set of unlabeled queries (transductive FSL) or some additional set of unlabeled data samples (semi-supervised FSL). Specifically, we show that on the challenging miniImageNet and tieredImageNet benchmarks, TAFSSL can improve the current state-of-the-art in both transductive and semi-supervised FSL settings by more than* 5%*, while increasing the benefit of using unlabeled data in FSL to above* 10% *performance gain.*

1. Introduction

The great success of Deep Learning (DL) methods to solve complex computer vision problems can be attributed in part to the emergence of large labeled datasets and strong parallel hardware. Yet in many practical situations, the amount of data and/or labels available for training or adapting the DL model to a new target task is prohibitively small. In extreme cases, we might be interested in learning from as little as one example per novel class. This is the typical scenario of Few-Shot Learning (FSL), a very active and exciting research topic of many concurrent works. Many methods have been proposed for effective pre-training of the FSL methods backbones. However, little attention has been given to adapting the feature spaces resulting from these backbones to the novel classes few-shot tasks during test time. It has been shown that some moderate gains can be obtained from using the few training examples (support set) of the novel tasks to fine-tune the backbones. It has also been shown that label propagation and clustering operating in the pre-trained backbone's original feature space provide some gains for FSL with additional unlabeled data (transductive and semi-supervised).

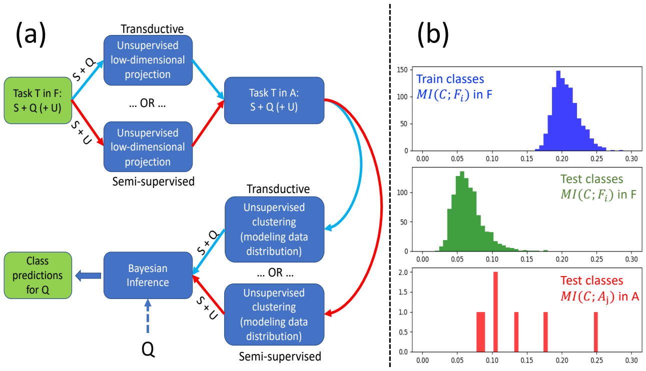


Figure 1: **(a) TAFSSL overview:** red and blue pathways are for semi-supervised and transductive FSL respectively. T - few-shot task; S - support set; Q - query set; U - optional set of additional unlabeled examples (semi-supervised FSL); F - original feature space; A - task adapted feature sub-space. **(b) Improved SNR in A:** the normalized (by min entropy) Mutual Information (MI) between either train or test classes and the features in F (of dimension 1024) or in A (7-dim) motivates to use A over F.

However, slight adaptation of the backbone's feature space to a given task, using few iterations of fine-tuning on the support set or other techniques, might not be sufficient to bridge over the generalization gap introduced by the FSL backbone observing completely novel classes unseen during training (as confirmed by the relatively moderate performance gains obtained from these techniques). Intuitively, we could attribute this in part to many of the feature space dimensions (feature vector entries) becoming 'useless' for a given set of novel classes in the test few-shot task. Indeed, every feature vector entry can be seen as a certain 'pattern detector' which fires strongly when a certain visual pattern is observed on the input image. The SGD (or other) backbone training is making sure all of these patterns are discriminative for the classes used for pre-training. But, due to likely over-fitting, many of these patterns are base classes specific, and do not fire for the novel test classes. Hence, their corresponding feature vector entries will mainly produce 'noise values' corresponding to 'pattern not observed'. In other words, the ratio of feature vector entries that can be used for recognition of novel classes to ones which mainly output 'noise' significantly decreases for test few-shot task

(Figure 1b). And it is unlikely that small modifications to the feature space recovers a significant portion of the 'noise producing' feature entries. The high level of noise in the feature vectors intuitively has significant adverse implications on the performance of the FSL classifier operating on this vector, especially the popular distance based classifiers like nearest-neighbor [11, 10] are affected. In light of this intuition, we conjecture that for a significant performance boost, we need to concentrate our efforts on the so-called *Task-Adaptive Feature Sub-Space Learning (TAFSSL)* - seeking sub-spaces of the backbone's feature space that are discriminative for the novel classes of the test few-shot task and which are 'as noise free as possible', that is most of the sub-space dimensions indeed 'find' the patterns they represent in the images of the novel categories belonging to the task (Figure 1a).

2. Related work

In many practical applications, in addition to the labeled support set, we have additional unlabeled data accompanying the few-shot task. In transductive FSL [2, 6, 4, 7] we assume the set of task queries arrives in a bulk and we can simply use it as a source of unlabeled data, allowing query samples to 'learn' from each other. In [2] the query samples are used in fine-tuning in conjunction with entropy minimization loss in order to maximize the certainty of their predictions. In semi-supervised FSL [5, 8, 1, 6] the unlabeled data comes in addition to the support set and is assumed to have a similar distribution to the target classes (although some unrelated samples noise is also allowed). In the LST [5] self-labeling and soft attention are used on the unlabeled samples intermittently with fine-tuning on the labeled and self-labeled data. Similarly to LST, [8] updates the class prototypes using k-means like iterations initialized from the PN prototypes. Their method also includes down-weighting the potential distractor samples (likely not to belong to the target classes) in the unlabeled data. In [1] unlabeled examples are used through soft-label propagation. In [9] semi-supervised few-shot domain adaptation is considered. In [3, 6, 4] graph neural networks are used for sharing information between labeled and unlabeled examples in semi-supervised [3, 6] and transductive [4] FSL setting. Notably, in [6] a Graph Construction network is used to predict the task specific graph for propagating labels between samples of semi-supervised FSL task.

3. FSSL and TAFSSL

Let a CNN backbone $\mathcal{B}$ pre-trained for FSL on a (large) dataset $\mathcal{D}_b$ with a set of base (training) classes $\mathcal{C}_b$. Denote by $\mathcal{B}(x) \in \mathcal{F} \subset \mathbf{R}^m$ to be a feature vector corresponding to an input image x represented in the feature space $\mathcal{F}$ by the backbone $\mathcal{B}$. Under this notation, we define the goal of linear Feature Sub-Space Learning (FSSL) to find an 'optimal' (for a certain task) linear sub-space $\mathcal{A}$ of $\mathcal{F}$ and a linear mapping W of size $r \times m$ (typically with $r \ll m$) such that:

$$\mathbf{R}^r \supset \mathcal{A} \ni A = W \cdot \mathcal{B}(x) \tag{1}$$

is the new representation of an input image x as a vector A in the feature sub-space $\mathcal{A}$ (spanned by rows of W).

Now, consider an n-way + k-shot few-shot test task $\mathcal{T}$ with a query set Q, and a support set: $S = \{s_i^j | 1 \leq i \leq n, 1 \leq j \leq k, \mathcal{L}(s_i^j) = i\}$, where $\mathcal{L}(x)$ is the class label of image x, so in S we have k training samples (shots) for each of the n classes in the task $\mathcal{T}$. Using the PN [10] paradigm we assume $k = 1$ (otherwise support examples of the same class are averaged to a single class prototype) and that each $q \in Q$ is classified using Nearest Neighbor (NN) in $\mathcal{F}$:

$$CLS(q) = \underset{i}{\operatorname{argmin}} ||\mathcal{B}(s_i^1) - \mathcal{B}(q)||^2 \tag{2}$$

Then, in the context of this given task $\mathcal{T}$, we can define linear Task-Adaptive FSSL (TAFSSL) as a search for a linear sub-space $\mathcal{A}_\mathcal{T}$ of the feature space $\mathcal{F}$ defined by a $\mathcal{T}$-specific projection matrix $W_\mathcal{T}$, such that the probability:

$$\frac{exp(-\tau \cdot ||W_\mathcal{T} \cdot (\mathcal{B}(s^1_{\mathcal{L}(q)}) - \mathcal{B}(q))||^2)}{\sum_i exp(-\tau \cdot ||W_\mathcal{T} \cdot (\mathcal{B}(s^1_i) - \mathcal{B}(q))||^2)} \tag{3}$$

of predicting q to belong to the same class as the 'correct' support $s^1_{\mathcal{L}(q)}$ is maximized, while of course the true label $\mathcal{L}(q)$ is unknown at test time (here τ in eq. 3 is a temperature parameter, we used $\tau = 1$).

With only few labeled samples in the support set S, we cannot expect to effectively learn the $W_\mathcal{T}$ projection to the sub-space $\mathcal{A}_\mathcal{T}$ using SGD on S. Yet, when unlabeled data accompanies the task $\mathcal{T}$ (Q in transductive FSL, or an additional set of unlabeled samples U in semi-supervised FSL), we can use this data to find such $W_\mathcal{T}$ that: (a) the dimensions of $\mathcal{A}_\mathcal{T}$ are 'disentangled', meaning their pairwise independence is maximized; (b) after the 'disentanglement' dimensions that 'exhibit the least noise' are kept.

Luckily, simple classical methods can be used for TAFSSL approximating the requirements (a) and (b). Both Principle Component Analysis (PCA) and Independent Component Analysis (ICA) applied in $\mathcal{F}$ on the set of samples: $S \cup Q$ (transductive FSL) or $S \cup U$ (semi-supervised FSL) can approximate (a). PCA under the approximate joint Gaussianity assumption of $\mathcal{F}$, and ICA under approximate non-Gaussianity assumption. In addition, if after the PCA rotation we keep the directions with maximal variance, thus biasing towards those that are *not* uni-modal (noise features are usually uni-modal). Similarly, the dimensions that are chosen by ICA and exhibit stronger departure from Gaussianity are more likely not to be noise.

TAFSSL summary. To summarize, both PCA and ICA are good simple approximations for TAFSSL using unlabeled data and therefore we simply use them to perform the 'unsupervised low-dimensional projection' in the first step of our proposed approach (Figure 1a).

4. Clustering

After applying PCA or ICA based TAFSSL, the feature noise levels are usually significantly reduced (Figure 1b) making the task-adapted feature sub-space $\mathcal{A}_{\mathcal{T}}$ of the original feature space $\mathcal{F}$ to be much more effective for clustering. We propose two clustering-based algorithms, the Bayesian K-Means (BKM) and Mean-Shift Propagation (MSP), summarized in Algorithms 1 and 2 respectively. They are used to perform the 'unsupervised clustering' + 'bayesian inference' steps of our approach (Figure 1a).

Algorithm 1 Bayesian K-Means (BKM)

```
Cluster the samples of task $\mathcal{T}$ ($Q \cup S$ or $U \cup S$ in transductive or semi-supervised FSL respectively) into $k$ clusters, associating each to $c_k$ - the centroid of cluster $k$.
for each $s \in S$, $q \in Q$, and $k$ do
    $P(cluster(q) = k) = \frac{\exp(-||q-c_k||^2)}{\sum_j \exp(-||q-c_j||^2)}$
    $P(cluster(s) = k) = \frac{\exp(-||s-c_k||^2)}{\sum_j \exp(-||s-c_j||^2)}$
    $P(\mathcal{L}(q) = i|cluster(q) = k) = \sum_{\mathcal{L}(s)=i} \frac{\exp(-||q-s||^2) \cdot P(cluster(s)=k)}{\sum_{t \in S} \exp(-||q-t||^2) \cdot P(cluster(t)=k)}$
$P(\mathcal{L}(q) = i) = \sum_k P(\mathcal{L}(q) = i|cluster(q) = k) \cdot P(cluster(q) = k)$ =0
```

Algorithm 2 Mean-Shift Propagation (MSP)

```
Initialize:
Compute prototypes: $\{p_i = \frac{1}{k} \cdot \sum_{s \in S, \mathcal{L}(s)=i} s\}$, where $k$ is # of shots in task $\mathcal{T}$
for N times do
    Compute $P(\mathcal{L}(x) = i) = \frac{\exp(-||x-p_i||^2)}{\sum_j \exp(-||x-p_j||^2)}$, $\forall x \in Q \cup S$ (or $x \in U \cup S$)
    Compute predictions $c(x) = \text{argmax}_i \, P(\mathcal{L}(x) = i)$
    $K_i = \sum_x \mathbb{1}_{(c(x)=i) \wedge (P(\mathcal{L}(x)=i)>T)}$, where $T$ is a threshold parameter
    K = $min_i\{K_i\}$
    Compute the new prototypes: $\{p_i = \frac{1}{K} \cdot \sum_{x \in \hat{S}_i} x\}$, where $\hat{S}_i$ are the top $K$ samples that have $c(x) = i$ sorted in decreasing order of $P(\mathcal{L}(x) = i)$
return labels $c(q), \forall q \in Q$
```

5. Results

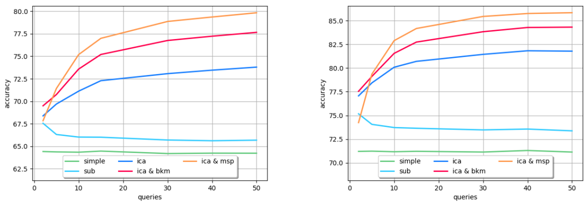
We have evaluated our approach on the popular few-shot *mini*ImageNet and *tiered*ImageNet classification benchmarks. We used the standard evaluation protocols, exactly as in corresponding (compared) works. The results of the transductive and semi-supervised FSL evaluation, together with comparison to previous methods, are summarized in tables 1 and 2 respectively. As can be seen, the top performing of our proposed TAFSSL based approaches (ICA+MSP) consistently outperforms all the previous SOTA by more then 10% (transductive) and 8% (semi-supervised) in the more challenging 1-shot setting. All the 0.95 confidence intervals are below 0.5% and are not reported. The tests are performed on $10,000$ random 5-way episodes with 15 queries each (unless otherwise specified). In all experiments not involving BKM, the class probabilities were computed using the NN classifier to the class prototypes. Simple = Simple-Shot [11], Sub = [11] with subtracting task means.

	Mini 1-shot	Mini 5-shot	Tiered 1-shot	Tiered 5-shot
Simple shot [11]	64.30	81.48	71.26	86.59
TPN [6]	55.51	69.86	59.91	73.30
TEAM [7]	60.07	75.90	-	-
EGNN + trans. [4]	-	76.37	-	80.15
Trans. Fine-Tuning [2]	65.73	78.40	73.34	85.50
PCA	70.53	80.71	80.07	86.42
ICA	72.10	81.85	80.82	86.97
BKM	72.05	80.34	79.82	85.67
PCA + BKM	75.11	82.24	83.19	87.83
ICA + BKM	75.79	82.83	83.39	88.00
MSP	71.39	82.67	76.01	87.13
PCA + MSP	76.31	84.54	84.06	89.13
ICA + MSP	**77.06**	**84.99**	**84.29**	**89.31**

Table 1: Transductive setting

	$\lvert U \rvert$	Mini 1-shot	Mini 5-shot	Tiered 1-shot	Tiered 5-shot
TPN [6]	360	52.78	66.42	-	-
PSN [1]	100	-	68.12	-	71.15
TPN [6]	1170	-	-	55.74	71.01
LST [5]	30	65.00	-	75.40	-
SKM [8]	100	62.10	73.60	68.60	81.00
TPN [6]	100	62.70	74.20	72.10	83.30
LST [5]	50	-	77.80	-	83.40
LST [5]	100	70.10	78.70	77.70	85.20
ICA	30	72.00	81.31	80.24	86.57
ICA	50	72.66	81.96	80.86	87.03
ICA	100	72.80	82.27	80.91	87.14
ICA + BKM	30	75.70	83.59	82.97	88.34
ICA + BKM	50	76.46	84.36	83.51	88.81
ICA + BKM	100	76.83	84.83	83.73	88.95
ICA + MSP	30	78.55	84.84	85.04	88.94
ICA + MSP	50	79.58	85.41	85.75	89.32
ICA + MSP	100	80.11	85.78	86.00	89.39

Table 2: Semi supervised setting.



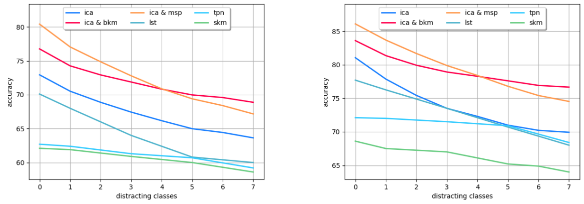
(a) mini (b) tired

Figure 2: Number of queries in transductive FSL setting.

5.1. Ablation study

Number of queries in transductive FSL. In transductive FSL the size of Q in the test episodes affects the performance. Figure 2 shows results varying the number of queries from 2 to 50. As can be seen from the figure, already for as little as 5 queries a substantial gap can be observed between ICA+MSP and the best of the baselines.

Out of distribution noise in unlabeled data. The unlabeled data may become contaminated with samples "unrelated" to the few-shot task $\mathcal{T}$. This type of noise is typically evaluated using additional random samples from "distracting" classes added to the unlabeled set. Figure 3 compares our ICA-based TAFSSL to SOTA semi-supervised FSL methods varying the number of distracting classes between 0 and 7, seeing a consistent 8% accuracy gain.



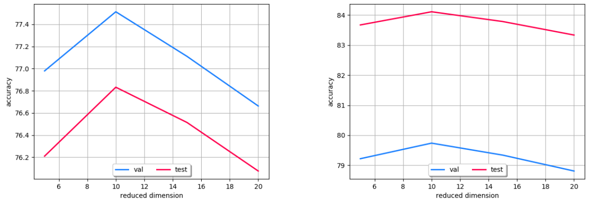
(a) mini

(b) tired

Figure 3: Unlabeled data noise in semi-supervised.

	Backbone	Mini 1-shot	Tiered 1-shot
TPN [6]	Conv-4	55.51	59.91
TPN [6]	ResNet-12	59.46	-
Transductive Fine-Tuning [2]	WRN	65.73	73.34
PCA + MSP	Conv-4	56.63	60.27
PCA + MSP	ResNet-10	70.93	76.27
PCA + MSP	ResNet-18	73.73	80.60
PCA + MSP	WRN	73.72	81.61
PCA + MSP	DenseNet	76.31	84.06

Table 3: Backbones comparison.



(a) mini

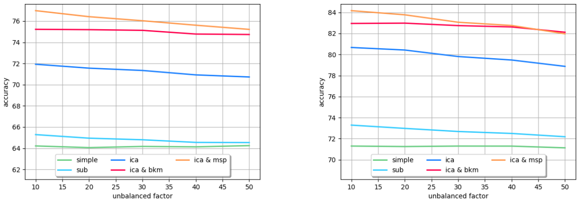
(b) tiered

Figure 4: ICA dimension vs accuracy.

The number of TAFSSL sub-space dimensions. In figure 4 we explore the effect of the number of chosen dimensions in ICA-based TAFFSL. As observed, the optimal number of dimensions is 10, which is consistent between both test and validation sets and the two benchmarks.

Backbone architectures. In Table 3 we different backbones to evaluate the performance of PCA+MSP in 1-shot transductive FSL. As expected, larger backbones produce better performance for the TAFSSL approach and greater gains (8%) over SOTA (using same backbone).

Unbalanced test classes in unlabeled data. In all previous transductive FSL works, the test tasks (episodes) were balanced in terms of the number of queries corresponding to each of the test classes. In practical applications there is no guarantee that the bulk of queries sent for offline evaluation will be balanced in terms of classes. Figure 5 evaluates our approach and baselines under varying levels of query set skew. The level of skew was controlled through having random $15+Uniform([0, R])$ query samples per class. Varying R from 10 to 50 (50 = factor 4 between # queries per class) shows our TAFSSL is robust to lack of balance.



(a) mini

(b) tiered

Figure 5: Unbalanced transductive FSL.

6. Summary and Conclusions

In this paper we have presented Task Adaptive Feature Sub-Space Learning (TAFSSL) achieving large margin improvements over transductive and semi-supervised FSL state-of-the-art. Potential future work directions include incorporating TAFSSL into the meta-training (pre-training) process; exploring non-linear TAFSSL; and exploring the benefits of TAFSSL in cross-domain few-shot learning.

References

[1] Authors Anonymous. Projective Sub-Space Networks For Few-Sot Learning. In ICLR 2019 OpenReview. 2, 3

[2] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A Baseline For Few-Shot Image Classification. Technical report, 2019. 2, 3, 4

[3] Victor Garcia and Joan Bruna. Few-Shot Learning with Graph Neural Networks. arXiv:1711.04043, pages 1–13, 2017. 2

[4] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-Labeling Graph Neural Network for Few-shot Learning. Technical report. 2, 3

[5] Xinzhe Li, Qianru Sun, Yaoyao Liu, Shibao Zheng, Qin Zhou, Tat-Seng Chua, and Bernt Schiele. Learning to Self-Train for Semi-Supervised Few-Shot Classification. 6 2019. 2, 3

[6] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning To Propagate Labels: Transductive Propagation Networ For Few-Shot Learning, 2019. 2, 3, 4

[7] Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive Episodic-Wise Adaptive Metric for Few-Shot Learning. 2019. 2, 3

[8] 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. ICLR, 3 2018. 2, 3

[9] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised Domain Adaptation via Minimax Entropy. In ICCV, 4 2019. 2

[10] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks for Few-shot Learning. In NIPS, 2017. 2

[11] Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. SimpleShot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning. 11 2019. 2, 3