Unsupervised Few-Shot Image Classification by Learning Features into Clustering Space

Shuo Li, Fang Liu, Zehua Hao, Kaibo Zhao, and Licheng Jiao

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Joint International Research Laboratory of Intelligent Perception and Computation, School of Artificial Intelligent, Xidian University, Xi'an, 710071, P.R. China

Abstract

Most few-shot image classification methods are trained based on tasks. Usually, tasks are built on base classes with a large number of labeled images, which consumes large effort. Unsupervised few-shot image classification methods do not need labeled images, because they require tasks to be built on unlabeled images. In order to efficiently build tasks with unlabeled images, we propose a novel single-stage clustering method: Learning Features into Clustering Space (LF2CS), which first set a separable clustering space by fixing the clustering centers and then use a learnable model to learn features into the clustering space. Based on our LF2CS, we put forward an image sampling and c-way k-shot task building method. With this, we propose a novel unsupervised few-shot image classification method, which jointly learns the learnable model, clustering and few-shot image classification. Experiments and visualization show that our LF2CS has a strong ability to generalize to the novel categories. From the perspective of image sampling, we implement four baselines according to how to build tasks. We conduct experiments on the Omniglot, miniImageNet, tieredImageNet, and CIFARFS datasets based on the Conv-4 and ResNet-12 backbones. Experimental results show that ours outperform the state-of-the-art methods.

Motivation

There are two common unsupervised ways to build tasks from the auxiliary dataset:

1) CSS-based methods (Comparative Self-Supervised) use data augmentations to obtain another view of the images to construct the image pairs, and then use the image pairs to build tasks;
2) Clustering-based methods, by a clustering algorithm, divide the images into clusters to obtain the pseudo-labels, and then use the pseudo-labels to build tasks.

Since CSS-based methods build tasks based on the different views, these methods are concise and effective, but the diversity of tasks may be poor. Also, since Clustering-based methods build tasks based on the clusters, the diversity of tasks may be good, but these methods usually contain multiple steps.

To combine the advantages of both methods, we propose a single-stage clustering method: Learn Features into Clustering Space (LF2CS).

Notations and Problem Statement

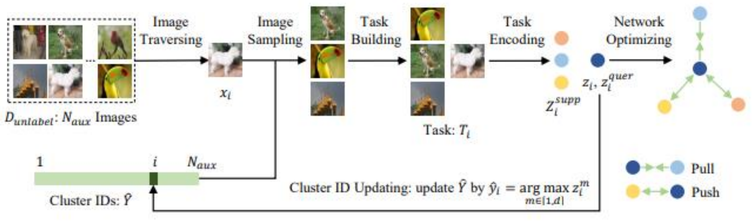
In few-shot learning (FSL), an auxiliary dataset $D_{aux} = \{(x_i, y_i)\}_{i=1}^{N_{aux}}$ is used to train a learnable model, and then applies the model to a target dataset $D_{target} = \{\{(x_i, y_i)\}_{i=1}^{N_{label}}, \{x_j\}_{j=1}^{N_{unlabel}}\}$ (where $N_{label} \ll N_{unlabel}$), so that the target dataset with few labeled images can be classified. D_{aux} and D_{target} are also known as the base classes and the novel classes, respectively. D_{aux} contains a large number of labeled images. Unsupervised few-shot learning (UFSL) aims to train a model using an unlabeled auxiliary dataset $D_{unlabel} = \{x_i\}_{i=1}^{N_{aux}}$ which contains a large number of unlabeled images, and then also applies the model to D_{target}. Task-based FSL organizes images into tasks in the form of c-way k-shot. Task T_i is composed of the support set T_i^{supp}, the query set T_i^{quer} and the label y_i^{task}:

$$T_i = \{T_i^{supp}, T_i^{quer}, y_i^{task}\} = \{\{(x_j, y_j)\}_{j=1}^{c \times k}, \{x_i\}, y_i^{task}\},$$

where $y_i^{task} \in \{0,1\}^{c \times k}$, T_i^{supp} contains c classes and each class has k labeled images. In our work, T_i^{quer} contains only one image x_i. Under the setting of supervised FSL, since the image label is available on the base classes, the task T_i is built in an supervised way. Under the setting of UFSL, the task T_i needs to be built in an unsupervised way.

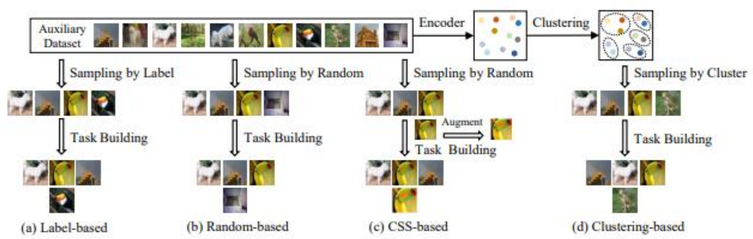
Our Approach

By fixing the clustering center matrix to the identity matrix, our LF2CS first sets a separable clustering space, and then uses a learnable model to learn features into the clustering space. By this way, LF2CS does not require any other images when determining the cluster id of each image. Based on LF2CS, we put forward an image sampling and task building method, and thus propose a clustering-based unsupervised few-shot image classification method, as shown in the following figure. First, each image x_i is fed into a random initialized network to obtain the feature z_i, and LF2CS is used to obtain the initial cluster id $\hat{y}_i$, so that the cluster ids $\hat{Y}$ of all images can be initialized. Then, for each image x_i, $c \times k$ images are sampled based on the cluster ids $\hat{Y}$, in which k images have the same cluster id as the image x_i, so a task can be built based on these $c \times k$+1 images. Finally, encoders are used to extract features of the images in the task, and the new cluster id $\hat{y}_i$ of the image x_i is calculated by LF2CS and updated to the cluster ids $\hat{Y}$.



Our method first generate cluster ids of images, then samples images to build tasks, and realizes joint learning of feature extraction, clustering and FSL. After the task is built, any task-based few-shot learning method can be used to optimize the network, such as MatchNet. We optimize the network based on MatchNet and use the softmax over cosine to measure similarity between features. FSL can be optimized by Mean Squared Error loss between p_i^{task} and the task label y_i^{task}, and LF2CS can be optimized by Cross Entropy loss between the feature z_i and the cluster id $\hat{y}_i$. The optimization of our method can be treated as a multi-task learning problem. Therefore, we jointly learn LF2CS and FSL, and our method increases the similarity of features with the same cluster id and reduces the similarity of features with different cluster ids.

Four Baselines



From the view of sampling, we implement four baselines according to how to build tasks: (a) Label-based baseline, which is a supervised baseline. Since the images have category labels, two of the four sampled images belong to the same category. (b) Random-based baseline, in which four images are randomly sampled, and the label of task is randomly determined. (c) CSS-based baseline, in which three images are randomly sampled, and then one of the images is selected to obtain another view through data augmentation. (d) Clustering-based baseline, in which first all images are divided into multiple clusters by a clustering algorithm, and then four images are selected with cluster ids as labels.

Please refer to https://github.com/xidianai/LF2CS for the code.

Experiments

We conduct experiments on the Omniglot, miniImageNet, tieredImageNet, and CIFARFS datasets. We mainly use 5-way 1-shot and 5-way 5-shot few-shot image classification accuracies to evaluate our method in our work. The larger the value of 5-way 1-shot and 5-way 5-shot few-shot image classification accuracies, the better the performance. We report the results on Omniglot in Table 1, miniImageNet in Table 2, tieredImageNet in Table 3, and CIFARFS in Table 4. Compared with other unsupervised baselines, the method of building tasks by randomly selecting images (Random-based) has the worst accuracies and our LF2CS has the best accuracies in all cases.

Table 1. Few-shot image classification accuracies of 5-way and 20-way on Omniglot. All accuracies are averaged over 1000 test episodes and are reported with 95% confidence intervals. The backbone of all methods is Conv-4. Bold represent the best values.

	Method	5-way Acc.		20-way Acc.	
		1-shot	5-shot	1-shot	5-shot
FSL	MatchNet [46]	98.10±-%	98.90±-%	93.80±-%	98.50±-%
	MAML [12]	98.70±0.40%	99.90±0.10%	95.80±0.30%	98.90±0.20%
	ProtoNet [44]	98.80±-%	99.70±-%	96.00±-%	98.90±-%
	Meta-SGD [31]	99.50±-%	99.90±-%	95.90±-%	99.00±-%
	RelationNet [45]	99.60±0.20%	99.80±0.10%	97.60±0.20%	99.10±0.10%
	Baselines: Label	97.92±0.22%	99.47±0.08%	92.84±0.21%	97.68±0.10%
UFSL	CACTUs [17]	68.84±-%	87.78±-%	48.09±-%	73.36±-%
	UMTRA [21]	83.80±-%	95.43±-%	74.25±-%	92.12±-%
	LASIUM [22]	83.26±0.55%	95.29±0.22%	-	-
	ProtoTransfer [35]	88.00±-%	96.48±-%	72.27±-%	89.08±-%
	AAL [2]	88.40±0.75%	98.00±0.32%	70.20±0.86%	88.30±1.22%
	ULDA [39]	91.00±0.42%	98.14±0.15%	78.05±0.31%	94.08±0.13%
	UFLST [18]	97.03±-%	99.19±-%	91.28±-%	97.37±-%
	Baselines: Random	58.50±0.70%	71.73±0.59%	33.94±0.31%	47.14±0.32%
	Baselines: CSS	83.97±0.56%	94.65±0.26%	65.25±0.34%	84.74±0.22%
	Baselines: Clustering	83.14±0.62%	91.67±0.36%	61.52±0.36%	77.05±0.28%
	Ours: LF2CS	**97.31±0.25%**	**99.32±0.10%**	**91.72±0.22%**	**97.65±0.09%**

Table 2. 5-way 1-shot and 5-way 5-shot few-shot image classification accuracies on the miniImageNet dataset. All accuracies are averaged over 1000 test episodes and reported with 95% confidence intervals. Bold represent the best values.

	Method	Backbone	5-way 1-shot	5-way 5-shot
FSL	MatchNet [46]	Conv-4	46.60±-%	60.00±-%
	ProtoNet [44]	Conv-4	49.42±0.78%	68.20±0.66%
	RelationNet [45]	Conv-4	50.44±0.82%	65.32±0.70%
	MAML [12]	Conv-4	48.70±1.84%	63.11±0.92%
	MetaOptnet [25]	ResNet-12	62.64±0.61%	78.63±0.46%
	Baselines: Label	Conv-4	52.53±0.62%	65.98±0.53%
	Baselines: Label	ResNet-12	60.06±0.69%	71.76±0.58%
UFSL	UFLST [18]	Conv-4	33.77±0.70%	45.03±0.73%
	AAL [2]	Conv-4	37.67±0.39%	40.29±0.68%
	CACTUs [17]	Conv-4	39.90±0.74%	53.97±0.70%
	UMTRA [21]	Conv-4	39.93±-%	50.73±-%
	LASIUM [22]	Conv-4	40.19±0.58%	54.56±0.55%
	ULDA [39]	Conv-4	40.63±0.61%	56.18±0.59%
	CSSL-FSL [26]	ResNet-50	48.53±1.26%	63.13±0.87%
	No-Labels [6]	ResNet-50	50.10±0.20%	60.10±0.20%
	Baselines: Random	Conv-4	25.29±0.39%	28.86±0.39%
	Baselines: CSS	Conv-4	41.00±0.56%	52.17±0.54%
	Baselines: Clustering	Conv-4	41.01±0.59%	51.52±0.55%
	Ours: LF2CS	Conv-4	**48.32±0.64%**	**61.52±0.52%**
	Baselines: Random	ResNet-12	29.92±0.46%	36.82±0.48%
	Baselines: CSS	ResNet-12	45.42±0.60%	58.37±0.54%
	Baselines: Clustering	ResNet-12	48.48±0.63%	61.10±0.58%
	Ours: LF2CS	ResNet-12	**53.14±0.62%**	**67.36±0.50%**

Table 3. 5-way 1-shot and 5-way 5-shot few-shot image classification accuracies on the tieredImageNet dataset. All accuracies are averaged over 1000 test episodes and reported with 95% confidence intervals. Bold represent the best values.

	Method	Backbone	5-way 1-shot	5-way 5-shot
FSL	ProtoNet [44]	Conv-4	53.31±0.89%	72.69±0.74%
	RelationNet [45]	Conv-4	54.48±0.93%	71.32±0.78%
	MAML [12]	Conv-4	51.67±1.81%	70.30±1.75%
	MetaOptnet [25]	ResNet-12	65.99±0.72%	81.56±0.53%
	Baselines: Label	Conv-4	56.09±0.72%	68.86±0.60%
	Baselines: Label	ResNet-12	64.67±0.76%	76.71±0.57%
UFSL	ULDA [39]	Conv-4	41.77±0.65%	56.78±0.63%
	Baselines: Random	Conv-4	24.81±0.36%	28.80±0.40%
	Baselines: CSS	Conv-4	40.74±0.59%	52.72±0.58%
	Baselines: Clustering	Conv-4	42.60±0.62%	55.11±0.57%
	Ours: LF2CS	Conv-4	**49.15±0.65%**	**62.54±0.58%**
	Baselines: Random	ResNet-12	31.58±0.50%	38.82±0.53%
	Baselines: CSS	ResNet-12	43.13±0.62%	56.36±0.56%
	Baselines: Clustering	ResNet-12	44.93±0.64%	57.53±0.59%
	Ours: LF2CS	ResNet-12	**53.16±0.66%**	**66.59±0.57%**

Table 4. Few-shot image classification accuracies on the CIFARFS dataset. All accuracies are averaged over 1000 test episodes and reported with 95% confidence intervals. Bold represent the best values.

	Method	Backbone	5-way 1-shot	5-way 5-shot
FSL	ProtoNet	Conv-4	55.50±0.70%	72.60±0.60%
	RelationNet	Conv-4	55.00±1.00%	69.30±0.80%
	MAML	Conv-4	58.90±1.90%	71.50±1.00%
	MetaOptnet	ResNet-12	72.08±0.70%	85.00±0.50%
	Baselines: Label-based	Conv-4	65.65±0.74%	76.75±0.57%
	Baselines: Label-based	ResNet-12	67.14±0.76%	77.46±0.54%
UFSL	No-Labels	ResNet-50	53.00±0.20%	62.50±0.20%
	Baselines: Random-based	Conv-4	30.87±0.47%	38.03±0.49%
	Baselines: CSS-based	Conv-4	42.59±0.65%	55.64±0.59%
	Baselines: Clustering-based	Conv-4	44.72±0.66%	58.21±0.57%
	Ours: LF2CS	Conv-4	**51.52±0.72%**	**66.82±0.57%**
	Baselines: Random-based	ResNet-12	34.25±0.55%	44.48±0.55%
	Baselines: CSS-based	ResNet-12	46.46±0.67%	61.39±0.59%
	Baselines: Clustering-based	ResNet-12	44.88±0.67%	58.50±0.59%
	Ours: LF2CS	ResNet-12	**55.04±0.72%**	**70.62±0.57%**

Conclusion

In our work, we propose a novel single-stage clustering method: Learning Features into Clustering Space (LF2CS), which fixes the cluster center matrix to the identity matrix, thereby setting a strongly separable clustering space, and then learns features into the clustering space. Based on this, we put forward an image sampling and task building method, and with this, we propose an unsupervised few-shot image classification method. Experimental results and visualization show that our LF2CS has a strong ability to generalize to the novel categories. Based on Conv-4 and ResNet-12, we conduct experiments on four FSL datasets, and our method achieves the state-of-the-art results.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.62076192), Key Research and Development Program in Shaanxi Province of China (No.2019ZDLGY03-06), the State Key Program of National Natural Science of China (No.61836009), in part by the Program for Cheung Kong Scholars and Innovative Research Team in University (No.IRT 15R53), in part by The Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No.B07048), in part by the Key Scientific Technological Innovation Research Project by Ministry of Education, the National Key Research and Development Program of China.