Task Affinity with Maximum Bipartite Matching in Few-Shot Learning

Cat P. Le, Juncheng Dong, Mohammadreza Soltani, Vahid Tarokh

Department of Electrical and Computer Engineering, Duke University

The International Conference on Learning Representations (ICLR) 2022

Duke

Motivation

- Few-shot learning is the problem of learning a task given only a few data samples. The additional database is often available for pre-training.
- Our goal is to design a continual learning framework for the few-shot learning.
- We propose a non-commutative task affinity score that is used to identify the related base tasks/classes of data.
- Next, we use the related data for pre-training and then fine-tuning the model with the few-shot data.

Task Definition

- A task is often defined as the function of data samples and the corresponding loss function.
- For classification task, a task is defined as the function of data samples and corresponding labels.
- We represent a task by a well-trained neural network on the data, referred to as an ε-approximation network.

Task Affinity Score

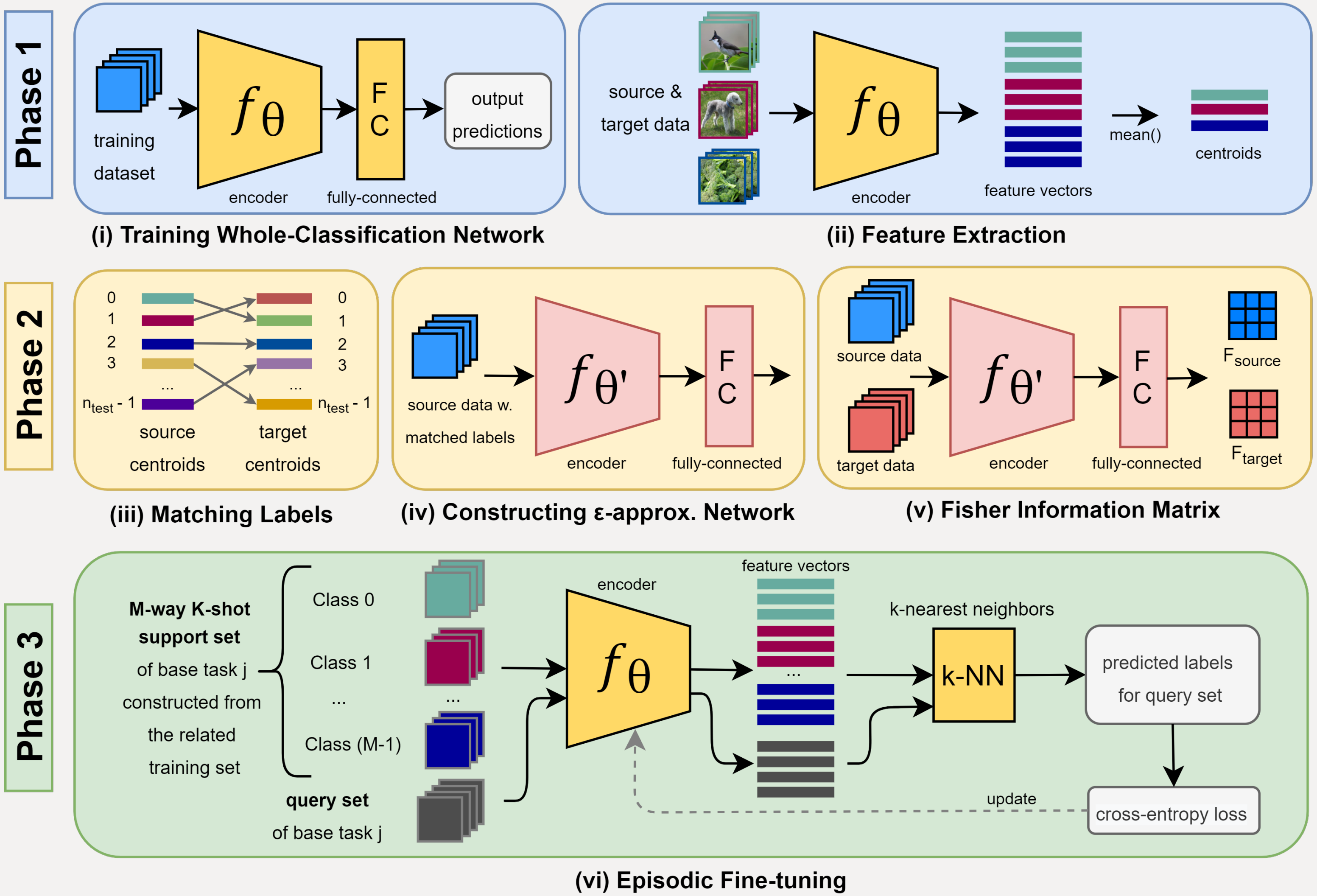
- The task affinity score (TAS) is asymmetric by design, since it is easier to apply comprehensive task to the simple one than vice versa.
- Let (T_a, X_a) be the ***source*** task-dataset. $N_{\theta a}$ is the approx. network.
- $F_{a,a}$ is the Fisher Information matrix of $N_{\theta a}$ with X_a^{query}.
- Let (T_b, X_b) is ***target*** task-dataset.
- $F_{a,b}$ is the Fisher Information matrix of $N_{\theta a}$ with $X_b^{support}$.
- TAS is defined as:

$$s[a,b] := \frac{1}{\sqrt{2}} \left\| F_{a,a}^{\frac{1}{2}} - F_{a,b}^{\frac{1}{2}} \right\|_F$$

Few-shot Framework

- A few-shot task of ***M-way K-shot*** is the classification of M classes, each class has K data points for training.
- Our proposed framework consists of 3 phases:
  1. **Training Whole-Classification Network and Feature Extraction**: training the representor network for the entire database classes and use this network's encoder for feature extraction.
  2. **Task Affinity**: find the most related data classes in the database to the target few-shot task and construct the related dataset.
  3. **Episodic Fine-tuning**: pretrain the few-shot model with related dataset, then episodic fine-tune the model with few-shot target data.

Few-shot Diagram



Phase 1

i. **Training the Whole-Classification Network**: train a neural network with the entire classes in the database.

ii. **Feature Extraction**: given the encoder of the well-trained whole-class network, we extract the feature vectors for each class of data and compute the mean vector, called centroid. This centroid is the embedding vector for the corresponding class of data.

Duke

Phase 2

- We define source tasks (with the same format as the target task) by randomly drawing the classes from the database.

iii. **Matching Labels**: map each source tasks' centroids to the target task's centroids to minimize the cumulative distance between pairs. After matching, we modify the source task's labels to match the target task's labels. This process guarantees the computed distance is label-invariant.

iv. **Constructing ε-approx. Network**: train a neural network to represent the modified label source task.

v. **Fisher Information Matrix**: compute Fisher matrices and TAS from source task to target task.

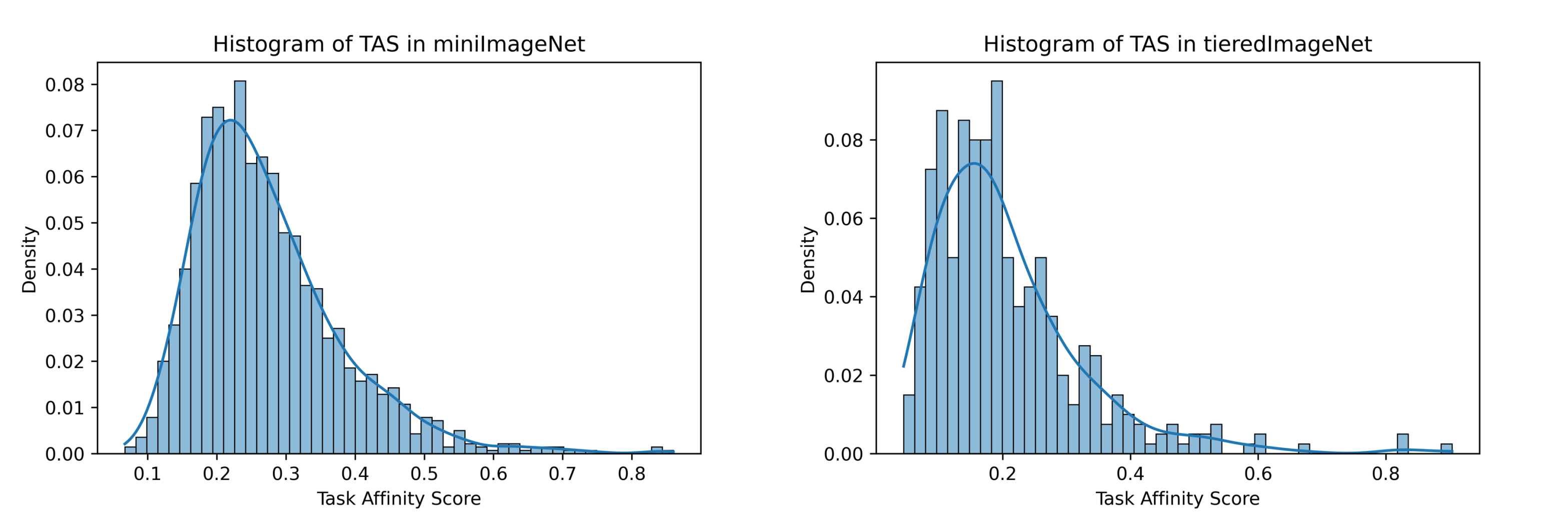
Phase 3

- Repeat the process in Phase 2 for various source tasks. Next, we select the top-N source tasks, and collect the corresponding classes of data.
- Construct the related dataset using the related classes.

vi. **Episodic Fine-Tuning**: construct the few-shot model using the encoder of the whole-class network (from Phase 1) and the k-NN classifier. Randomly generate few-shot base task from the related dataset and fine-tuning the few-shot model using cross entropy loss. Lastly, apply the target data to update the k-NN classifier of the few-shot model.

Experimental Results

- We conducts experiments on miniImageNet and tieredImageNet.
- We compute the TAS for numerous source tasks, which are randomly generated from the dataset.



Results

- We select the top-N closest tasks to construct the related dataset.
- Our few-shot models achieve competitive results given a smaller number of parameters.
- As more incoming tasks arrive, our framework is capable of learning continuously.

Table 1: Comparison of the accuracy against state-of-the art methods for 5-way 1-shot and 5-way 5-shot classification with 95% confidence interval on miniImageNet dataset.

Model	Backbone	Params	1-shot	5-shot
Matching-Net (Vinyals et al., 2016)	ConvNet-4	0.11M	43.56±0.84	55.31±0.73
MAML (Finn et al., 2017)	ConvNet-4	0.11M	48.70±1.84	63.11±0.92
Prototypical-Net (Snell et al., 2017)	ConvNet-4	0.11M	49.42±0.78	68.20±0.66
Simple CNAPS (Bateni et al., 2020)	ResNet-18	11M	53.2±0.90	70.8±0.70
Activation-Params (Qiao et al., 2018)	WRN-28-10	37.58M	59.60±0.41	73.74±0.19
LEO (Rusu et al., 2018)	WRN-28-10	37.58M	61.76±0.08	77.59±0.12
Baseline++ (Chen et al., 2019)	ResNet-18	11.17M	51.87±0.77	75.68±0.63
SNAIL (Mishra et al., 2017)	ResNet-12	7.99M	55.71±0.99	68.88±0.92
AdaResNet (Munkhdalai et al., 2018)	ResNet-12	7.99M	56.88±0.62	71.94±0.57
TADAM (Oreshkin et al., 2018)	ResNet-12	7.99M	58.50±0.30	76.70±0.30
MTL (Sun et al., 2019)	ResNet-12	8.29M	61.20±1.80	75.50±0.80
MetaOptNet (Lee et al., 2019)	ResNet-12	12.42M	62.64±0.61	78.63±0.46
SLA-AG (Lee et al., 2020)	ResNet-12	7.99M	62.93±0.63	79.63±0.47
ConstellationNet (Xu et al., 2020)	ResNet-12	7.99M	64.89±0.23	79.95±0.17
RFS-distill (Tian et al., 2020)	ResNet-12	13.55M	64.82±0.60	82.14±0.43
EPNet (Rodríguez et al., 2020)	ResNet-12	7.99M	65.66±0.85	81.28±0.62
Meta-Baseline (Chen et al., 2021)	ResNet-12	7.99M	63.17±0.23	79.26±0.17
IE-distill[1] (Rizve et al., 2021)	ResNet-12	9.13M	65.32±0.81	83.69±0.52
TAS-simple (ours)	**ResNet-12**	**7.99M**	**64.71**±0.43	**82.08**±0.45
TAS-distill (ours)	**ResNet-12**	**7.99M**	**65.13**±0.39	**82.47**±0.52
TAS-distill[2] (ours)	**ResNet-12**	**12.47M**	**65.68**±0.45	**83.92**±0.55

[1] performs with standard ResNet-12 with Dropblock as a regularizer, [2] performs with wide-layer ResNet-12

More Results

Table 2: Comparison of the accuracy against state-of-the art methods for 5-way 1-shot and 5-way 5-shot classification with 95% confidence interval on tieredImageNet dataset .

Model	Backbone	Params	1-shot	5-shot
MAML (Finn et al., 2017)	ConvNet-4	0.11M	51.67$\pm$1.81	70.30$\pm$0.08
Prototypical-Net (Snell et al., 2017)	ConvNet-4	0.11M	53.31$\pm$0.89	72.69$\pm$0.74
Relation-Net (Sung et al., 2018)	ConvNet-4	0.11M	54.48$\pm$0.93	71.32$\pm$0.78
Simple CNAPS (Bateni et al., 2020)	ResNet-18	11M	63.00$\pm$1.00	80.00$\pm$0.80
LEO-trainval (Rusu et al., 2018)	ResNet-12	7.99M	66.58$\pm$0.70	85.55$\pm$0.48
Shot-Free (Ravichandran et al., 2019)	ResNet-12	7.99M	63.52$\pm n/a$	82.59$\pm n/a$
Fine-tuning (Dhillon et al., 2019)	ResNet-12	7.99M	68.07$\pm$0.26	83.74$\pm$0.18
MetaOptNet (Lee et al., 2019)	ResNet-12	12.42M	65.99$\pm$0.72	81.56$\pm$0.53
RFS-distill (Tian et al., 2020)	ResNet-12	13.55M	71.52$\pm$0.69	86.03$\pm$0.49
EPNet (Rodríguez et al., 2020)	ResNet-12	7.99M	72.60$\pm$0.91	85.69$\pm$0.65
Meta-Baseline (Chen et al., 2021)	ResNet-12	7.99M	68.62$\pm$0.27	83.74$\pm$0.18
IE-distill[1] (Rizve et al., 2021)	ResNet-12	13.55M	72.21$\pm$0.90	87.08$\pm$0.58
TAS-simple (ours)	**ResNet-12**	**7.99M**	**71.98$\pm$0.39**	**86.58$\pm$0.46**
TAS-distill (ours)	**ResNet-12**	**7.99M**	**72.81$\pm$0.48**	**87.21$\pm$0.52**

[1] performs with wide-layer ResNet-12 with Dropblock as a regularizer

Table 3: Comparison of the accuracy against state-of-the art methods for 5-way 1-shot and 5-way 5-shot classification with 95% confidence interval on CIFAR-FS dataset.

			CIFAR-FS	
Model	**Backbone**	**Params**	1-shot	5-shot
MAML (Finn et al., 2017)	ConvNet-4	0.11M	58.90$\pm$1.90	71.50$\pm$1.00
Prototypical-Net (Snell et al., 2017)	ConvNet-4	0.11M	55.50$\pm$0.70	72.00$\pm$0.60
Relation-Net (Sung et al., 2018)	ConvNet-4	0.11M	55.00$\pm$1.00	69.30$\pm$0.80
Prototypical-Net (Snell et al., 2017)	ResNet-12	7.99M	72.20$\pm$0.70	83.50$\pm$0.50
Shot-Free (Ravichandran et al., 2019)	ResNet-12	7.99M	69.20$\pm n/a$	84.70$\pm n/a$
TEWAM (Qiao et al., 2019)	ResNet-12	7.99M	70.40$\pm n/a$	81.30$\pm n/a$
MetaOptNet (Lee et al., 2019)	ResNet-12	12.42M	72.60$\pm$0.70	84.30$\pm$0.50
RFS-simple (Tian et al., 2020)	ResNet-12	13.55M	71.50$\pm$0.80	86.00$\pm$0.50
RFS-distill (Tian et al., 2020)	ResNet-12	13.55M	73.90$\pm$0.80	86.90$\pm$0.50
IE-distill[1] (Rizve et al., 2021)	ResNet-12	9.13M	75.46$\pm$0.86	88.67$\pm$0.58
TAS-simple (ours)	**ResNet-12**	**7.99M**	**73.47$\pm$0.42**	**86.82$\pm$0.49**
TAS-distill (ours)	**ResNet-12**	**7.99M**	**74.02$\pm$0.55**	**87.65$\pm$0.58**
TAS-distill[2] (ours)	**ResNet-12**	**12.47M**	**75.56$\pm$0.62**	**88.95$\pm$0.65**

[1] performs with standard ResNet-12 with Dropblock as a regularizer, [2] performs with wide-layer ResNet-12

Table 4: Comparison of the accuracy against state-of-the art methods for 5-way 1-shot and 5-way 5-shot classification with 95% confidence interval on FC-100 dataset.

			FC-100	
Model	**Backbone**	**Params**	1-shot	5-shot
Prototypical-Net (Snell et al., 2017)	ConvNet-4	0.11M	35.30$\pm$0.60	48.60$\pm$0.60
Prototypical-Net (Snell et al., 2017)	ResNet-12	7.99M	37.50$\pm$0.60	52.50$\pm$0.60
TADAM (Oreshkin et al., 2018)	ResNet-12	7.99M	40.10$\pm$0.40	56.10$\pm$0.40
MetaOptNet (Lee et al., 2019)	ResNet-12	12.42M	41.10$\pm$0.60	55.50$\pm$0.60
RFS-simple (Tian et al., 2020)	ResNet-12	13.55M	42.60$\pm$0.70	59.10$\pm$0.60
RFS-distill (Tian et al., 2020)	ResNet-12	13.55M	44.60$\pm$0.70	60.90$\pm$0.60
IE-distill[1] (Rizve et al., 2021)	ResNet-12	9.13M	44.65$\pm$0.77	61.24$\pm$0.75
TAS-simple (ours)	**ResNet-12**	**7.99M**	**43.10$\pm$0.67**	**60.65$\pm$0.62**
TAS-distill (ours)	**ResNet-12**	**7.99M**	**44.62$\pm$0.70**	**61.46$\pm$0.65**

[1] performs with standard ResNet-12 with Dropblock as a regularizer

Conclusion

- We propose a non-commutative task affinity.
- We design a few-shot learning framework that has memory and is capable of selective learning from the related data.
- Our few-shot model achieves competitive performance while using a small number of parameters.
- Additionally, this model is capable of continuous few-shot learning.