Code

Paper

ICML
International Conference
On Machine Learning

Learning to Learn from APIs: Black-Box Data-Free Meta-Learning

Zixuan Hu[1], Li Shen[2], Zhenyi Wang[3], Baoyuan Wu[4], Chun Yuan[1], Dacheng Tao[5]

1Tsinghua Shenzhen International Graduate School, China; 2 JD Explore Academy, China;
3 State University of New York at Buffalo, USA;
4 the Chinese University of Hong Kong, Shenzhen, China; 5 the University of Sydney, Australia

Tsinghua University

京东探索研究院
JD EXPLORE ACADEMY

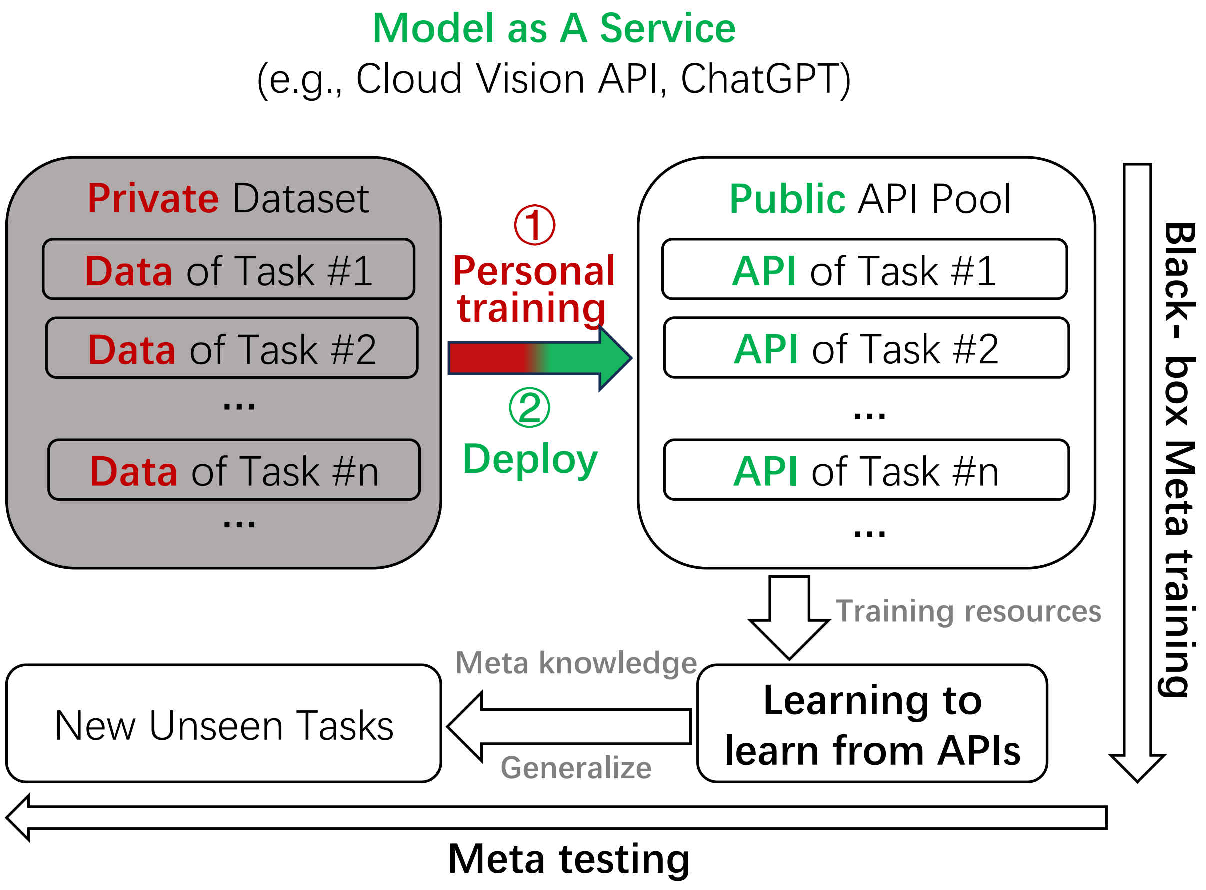
University at Buffalo
The State University of New York

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

THE UNIVERSITY OF SYDNEY

Overview

Learning to Learn from APIs (Black-box Data-free Meta-learning):



Goal:

Learning to learn from APIs or Black-box Data-free Meta-learning aims to enable efficient learning of new unseen tasks by meta-learning the meta-knowledge from a collection of black-box APIs without access to their private training data and with only query access.

Challenges:

- **Data-free:** no access to the original training data
- **Black-box:** with only query access to the APIs and with no prior knowledge of the underlying model architecture and parameters inside each API
- **Privacy-preserving:** no privacy leakage of original training data
- **Model-agnostic:** each API may correspond to arbitrary underlying model architectures and model scale.

Motivation

What information can we obtain from numerous public APIs of different tasks?

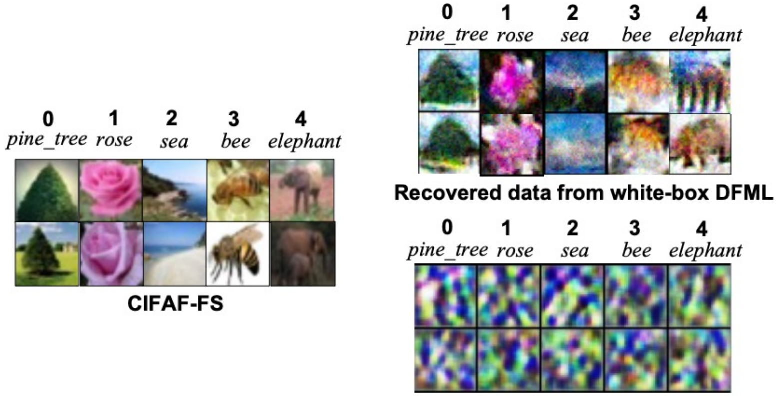
- Model parameters
- Model architectures
- Underlying data knowledge

Model parameters, Model architectures — DRO[1](UAI, 2022)

Underlying data knowledge — PURER[2](CVPR, 2023)

Challenges	DRO[1]	PURER[2]	Ours
Data-free	✅	✅	✅
Black-box	❌	❌	✅
Privacy-preserving	✅	❌	✅
Model-agnostic	❌	✅	✅

Explore underlying data knowledge contained in APIs in a safe (privacy-preserving) way:



CIFAF-FS

Recovered data from white-box DFML

Recovered data from ours

Pseudo Data ← APIs

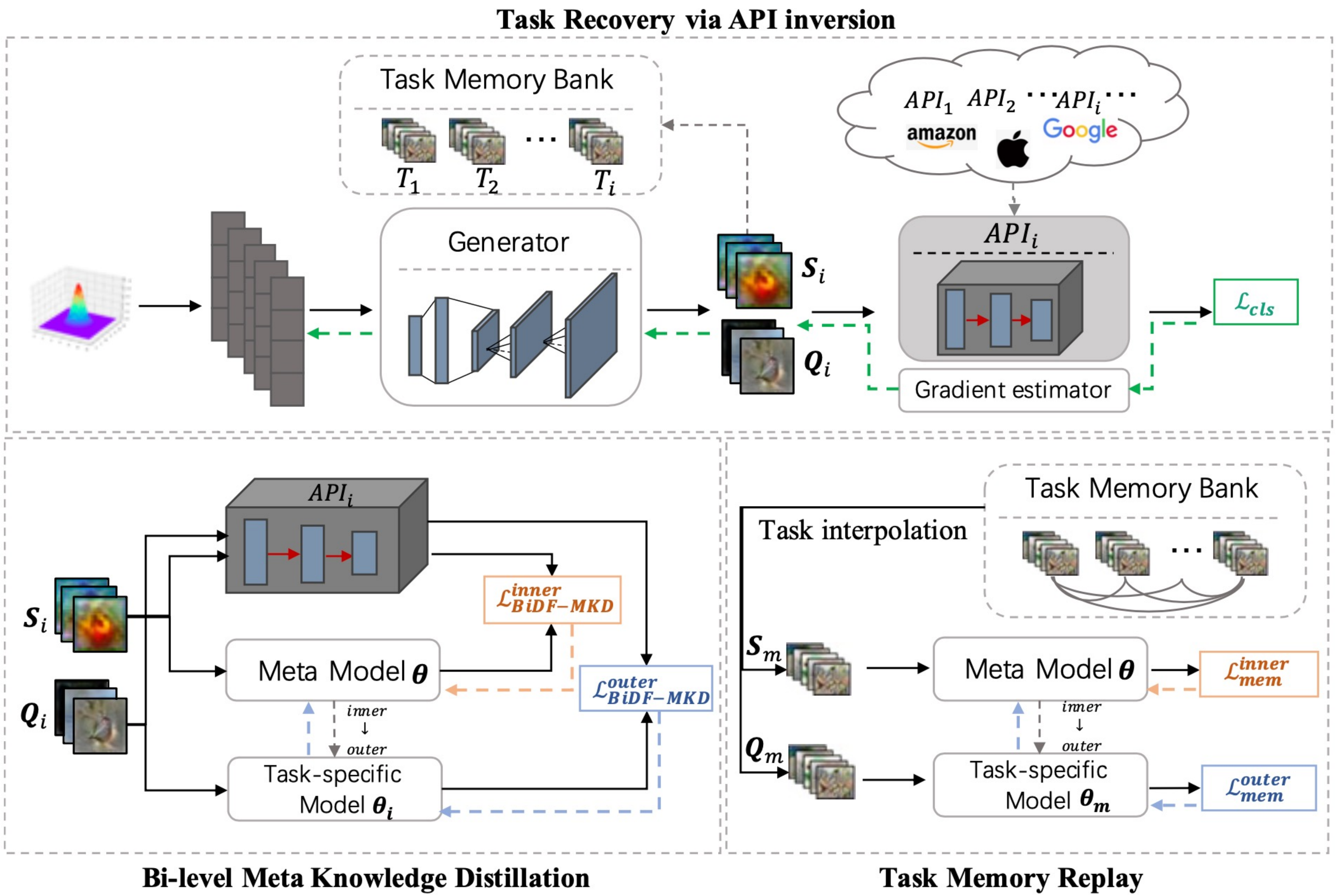
Inverse

Privacy-preserving

[1] Zhenyi Wang, et al. Meta-learning without data via Wasserstein distributionally-robust model fusion. UAI 2022.
[2] Zixuan Hu, et al. Architecture, Dataset and Model-Scale Agnostic Data-free Meta-Learning. CVPR 2023.

Methodology

Overall framework:



Methodology

<u>Pseudo task recovery via API inversion:</u>

1. Objective function

$$\min_{\boldsymbol{Z},\boldsymbol{\theta}_G} \mathcal{L}_{cls}(\hat{\boldsymbol{X}}) = \frac{1}{|\hat{\boldsymbol{X}}|} \sum_{(\hat{\boldsymbol{x}},y)\in(\hat{\boldsymbol{X}},\boldsymbol{Y})} l_{cls}(\hat{\boldsymbol{x}},y), \text{ s.t. } \hat{\boldsymbol{X}} = \mathrm{G}(\boldsymbol{Z};\boldsymbol{\theta}_{\mathrm{G}}),\ \ell_{cls}(\hat{\boldsymbol{x}},y) = CE(A_i(\hat{\boldsymbol{x}}),y)$$

2. Gradient backward propogation

$$\boldsymbol{\theta}_G^{t+1} = \boldsymbol{\theta}_G^t - \eta\nabla_{\boldsymbol{\theta}_G}\mathcal{L}_{cls}$$
$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \eta\nabla_{\boldsymbol{z}}\mathcal{L}_{cls}.$$

$$\nabla_{\boldsymbol{\theta}_G}\mathcal{L}_{cls} = \frac{\partial\mathcal{L}_{cls}}{\partial\boldsymbol{\theta}_G} = \frac{1}{|\hat{\boldsymbol{X}}|}\sum_{\hat{\boldsymbol{x}}\in\hat{\boldsymbol{X}}}\left[\frac{\partial\ell_{cls}}{\partial\hat{\boldsymbol{x}}}\times\frac{\partial\hat{\boldsymbol{x}}}{\partial\boldsymbol{\theta}_G}\right]$$

$$\nabla_{\boldsymbol{z}}\mathcal{L}_{cls} = \frac{\partial\mathcal{L}_{cls}}{\partial\boldsymbol{z}} = \frac{\partial\ell_{cls}}{\partial\hat{\boldsymbol{x}}}\times\frac{\partial\hat{\boldsymbol{x}}}{\partial\boldsymbol{z}}$$

intractable

Zero-order gradient estimation

3. Zero-order optimization by querying APIs

$$\hat{\nabla}_{\hat{\boldsymbol{x}}}\ell_{cls} = \frac{1}{q}\sum_{i=1}^{q}\left[\frac{d_{\hat{\boldsymbol{x}}}}{\mu}\left(\ell_{cls}\left(\hat{\boldsymbol{x}}+\mu\boldsymbol{u}_i,y\right)-\ell_{cls}(\hat{\boldsymbol{x}},y)\right)\boldsymbol{u}_i\right]$$

Methodology

ICML International Conference On Machine Learning

Bi-level meta knowledge distillation for meta-learning:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{BiDf-MKD}^{outer}(\boldsymbol{\theta}) = \sum_{\hat{\boldsymbol{x}} \in \boldsymbol{Q}_i} \ell_{KL}(F(\hat{\boldsymbol{x}}; \boldsymbol{\theta}_i), A_i(\hat{\boldsymbol{x}})),$$

$$\text{s.t. } \boldsymbol{\theta}_{\mathrm{i}} = \min_{\theta} \mathcal{L}_{\mathrm{BiDf-MKD}}^{\mathrm{inner}} \triangleq \min_{\boldsymbol{\theta}} \sum_{\hat{\boldsymbol{x}} \in \boldsymbol{S}_i} \ell_{KL}(F(\hat{\boldsymbol{x}}; \boldsymbol{\theta}), A_i(\hat{\boldsymbol{x}}))$$

Task-memory replay:

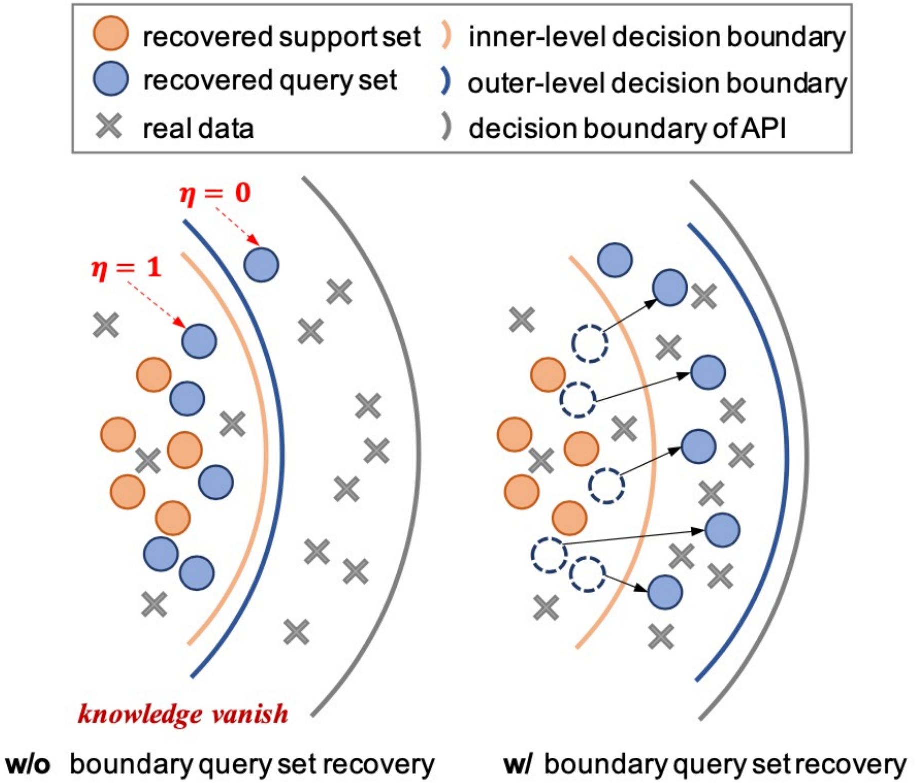
$$\min_{\boldsymbol{\theta}} \mathcal{L}_{mem}^{outer} = \mathcal{L}_{cls}(F(\boldsymbol{Q}_m; \boldsymbol{\theta}_m), \boldsymbol{Y}_{\boldsymbol{Q}_m}),$$

$$\text{s.t. } \boldsymbol{\theta}_{\mathrm{m}} = \min_{\theta} \mathcal{L}_{\mathrm{mem}}^{\mathrm{inner}} \triangleq \min_{\theta} \mathcal{L}_{\mathrm{cls}}(\mathrm{F}(\boldsymbol{S}_{\mathrm{m}}; \boldsymbol{\theta}), \boldsymbol{Y}_{\boldsymbol{S}_{\mathrm{m}}})$$

Methodology

Knowledge vanish issue of data-free meta-learning:

Definition 4.1. The *complete knowledge vanish* of meta-learning occurs when the outer-level optimization can be ignored, namely the mutual information $I(\boldsymbol{\theta}; \boldsymbol{Q}_i | \boldsymbol{\theta}_i, \boldsymbol{S}_i) = 0$ (or $H(\boldsymbol{\theta}|\boldsymbol{\theta}_i, \boldsymbol{S}_i) = H(\boldsymbol{\theta}|\boldsymbol{\theta}_i, \boldsymbol{S}_i, \boldsymbol{Q}_i)$).



Boundary query set recovery:

$$\min_{\boldsymbol{z}, \boldsymbol{\theta}_G} \ \ell_Q(\hat{\boldsymbol{x}}, y) = CE(A_i(\hat{\boldsymbol{x}}), y) - \lambda_Q \cdot \eta \cdot \ell_{KL}(F(\hat{\boldsymbol{x}}; \boldsymbol{\theta}_i), A_i(\hat{\boldsymbol{x}})),$$

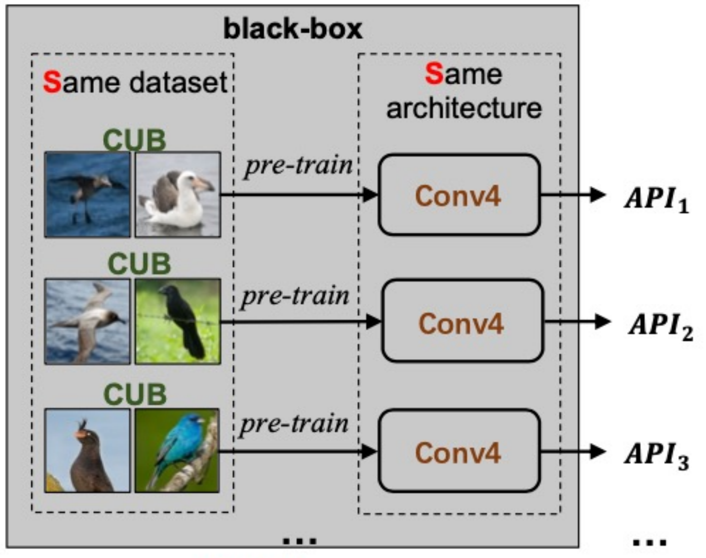
$$\text{s.t. } \hat{\boldsymbol{x}} = \mathrm{G}(\boldsymbol{z}; \boldsymbol{\theta}_{\mathrm{G}}),$$

$$\eta = \mathbb{I}\{\arg\max F(\hat{\boldsymbol{x}}; \boldsymbol{\theta}_i) = \arg\max A_i(\hat{\boldsymbol{x}})\}.$$

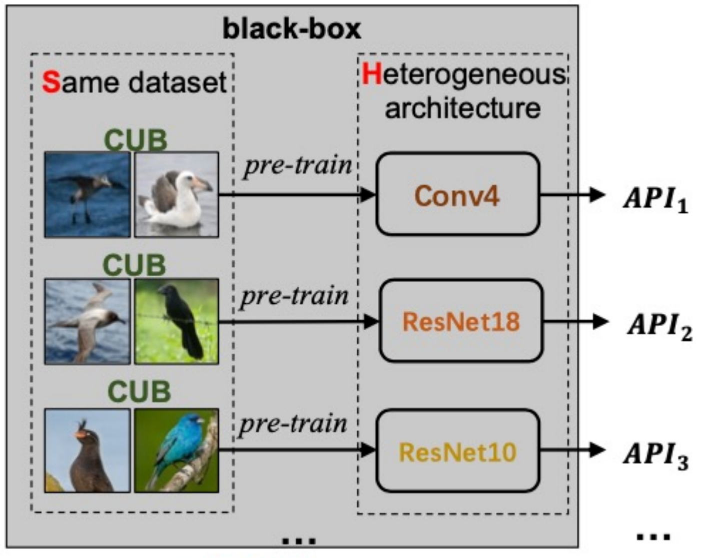
Experiments

Three real-world scenarios:

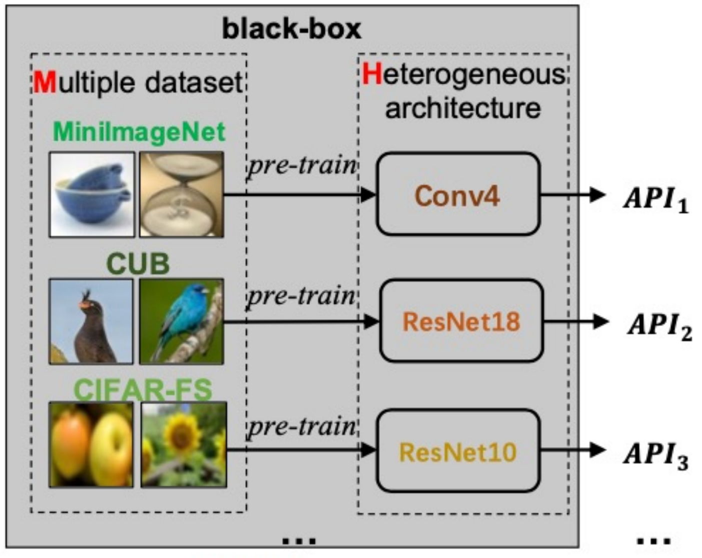
- **API-SS**: All APIs are designed for solving tasks from the **S**ame meta training subset with the **S**ame architecture inside
- **API-SH**: All APIs are designed for solving tasks from the **S**ame meta training subset but with **H**eterogeneous architectures inside
- **API-MH**: All APIs are designed for solving tasks from **M**ultiple meta training subsets with **H**eterogeneous architectures inside



API-SS scenario



API-SH scenario



API-MH scenario

Experiments

Main results:

Table 1. Compare to baselines in API-SS scenario.

API-SS	Method	1-shot	5-shot
CIFAR-FS 5-way	Random	20.35 ± 0.42	20.59 ± 0.45
	Best-API	19.04 ± 0.68	19.04 ± 0.67
	Single-DFKD	20.04 ± 0.63	20.14 ± 0.64
	Distill-Avg	24.24 ± 0.46	27.56 ± 0.51
	Ours	$\mathbf{35.48 \pm 0.67}$	$\mathbf{47.58 \pm 0.74}$
MiniImageNet 5-way	Random	21.20 ± 0.38	21.13 ± 0.37
	Best-API	20.51 ± 0.63	20.39 ± 0.62
	Single-DFKD	20.03 ± 0.60	20.14 ± 0.66
	Distill-Avg	20.53 ± 0.20	21.24 ± 0.24
	Ours	$\mathbf{29.35 \pm 0.60}$	$\mathbf{39.47 \pm 0.64}$
CUB 5-way	Random	21.09 ± 0.38	21.11 ± 0.37
	Best-API	19.99 ± 0.69	19.95 ± 0.70
	Single-DFKD	19.56 ± 0.64	20.06 ± 0.64
	Distill-Avg	21.07 ± 0.25	21.97 ± 0.30
	Ours	$\mathbf{29.10 \pm 0.64}$	$\mathbf{43.43 \pm 0.66}$

Table 3. Compare to baselines in API-MH scenario.

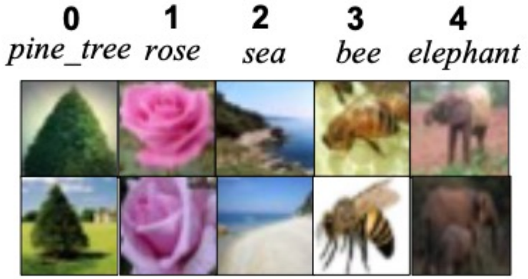
API-MH	Method	1-shot	5-shot
5-way	Random	20.88 ± 0.39	21.00 ± 0.40
	Best-API	19.44 ± 0.65	19.64 ± 0.66
	Single-DFKD	19.04 ± 0.66	19.68 ± 0.64
	Distill-Avg	21.57 ± 0.25	23.11 ± 0.29
	Ours	$\mathbf{32.78 \pm 0.60}$	$\mathbf{40.24 \pm 0.65}$

Table 2. Compare to baselines in API-SH scenario.

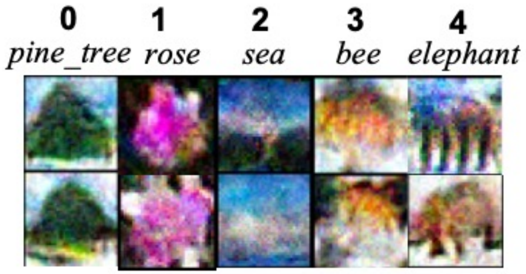
API-SH	Method	1-shot	5-shot
CIFAR-FS 5-way	Random	20.35 ± 0.42	20.59 ± 0.45
	Best-API	19.04 ± 0.68	19.04 ± 0.67
	Single-DFKD	19.56 ± 0.67	20.06 ± 0.60
	Distill-Avg	22.82 ± 0.38	25.91 ± 0.45
	Ours	$\mathbf{35.58 \pm 0.79}$	$\mathbf{46.92 \pm 0.77}$
MiniImageNet 5-way	Random	21.20 ± 0.38	21.13 ± 0.37
	Best-API	20.51 ± 0.63	20.39 ± 0.62
	Single-DFKD	20.11 ± 0.64	20.23 ± 0.66
	Distill-Avg	20.32 ± 0.22	20.67 ± 0.24
	Ours	$\mathbf{30.55 \pm 0.62}$	$\mathbf{39.74 \pm 0.65}$
CUB 5-way	Random	21.09 ± 0.38	21.11 ± 0.37
	Best-API	19.99 ± 0.69	19.95 ± 0.70
	Single-DFKD	20.13 ± 0.66	20.24 ± 0.64
	Distill-Avg	20.46 ± 0.24	21.02 ± 0.26
	Ours	$\mathbf{30.11 \pm 0.58}$	$\mathbf{43.98 \pm 0.64}$

Results

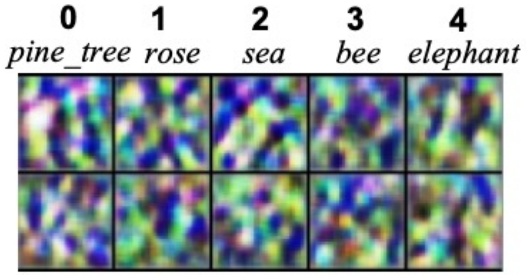
Data privacy:



CIFAF-FS



Recovered data from white-box DFML



Recovered data from ours

Zero-order VS. First-order (black-box VS. white-box):

Table 5. Effectiveness of zero-order gradient estimator. Grey: unfair comparison with white-box DFML.

API-SS	Method	1-shot	5-shot
CIFAR-FS 5-way	FO	37.66 ± 0.75	51.16 ± 0.79
	ZO	35.48 ± 0.67	47.58 ± 0.74
MiniImageNet 5-way	FO	30.66 ± 0.59	42.30 ± 0.64
	ZO	29.35 ± 0.60	39.47 ± 0.64
CUB 5-way	FO	31.62 ± 0.60	44.32 ± 0.69
	ZO	29.10 ± 0.64	43.43 ± 0.66

Results

Effect of the number of APIs:

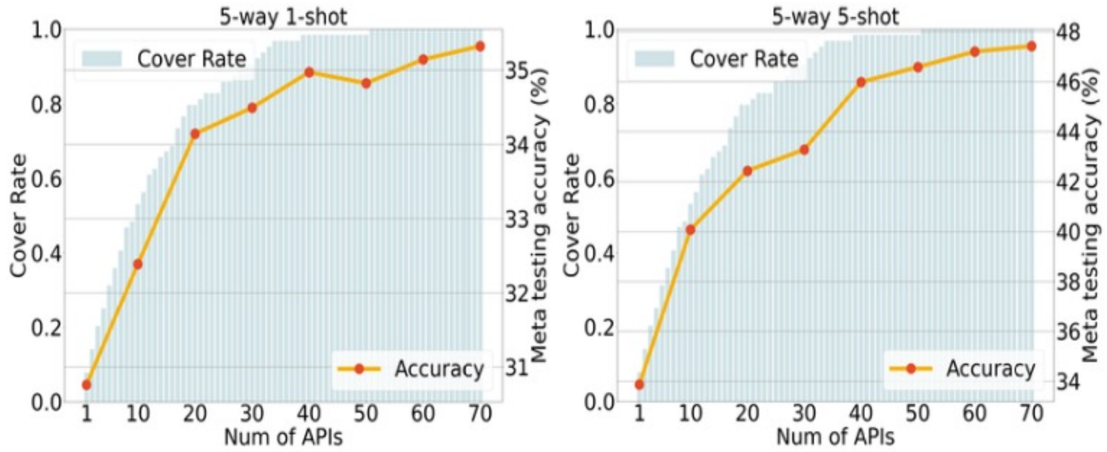


Figure 5. Effect of the number of APIs in API-SS scenario.

Effect of the number of query times:

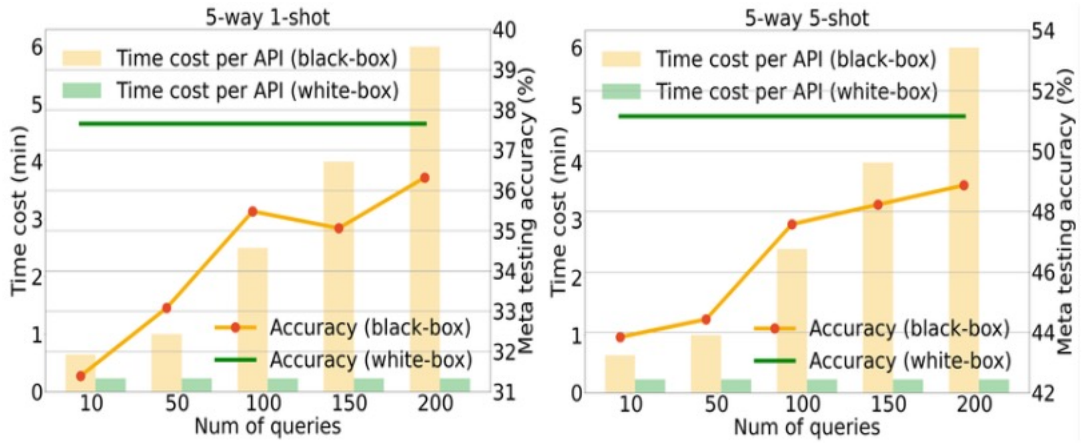


Figure 6. Effect of the number of query times on the accuracy and time cost. Here, white-box DFML provides unfair bounds of accuracy and time cost.

Code

Paper

Learning to Learn from APIs: Black-Box Data-Free Meta-Learning

Zixuan Hu[1], Li Shen[2], Zhenyi Wang[3], Baoyuan Wu[4], Chun Yuan[1], Dacheng Tao[5]

1Tsinghua Shenzhen International Graduate School, China; 2 JD Explore Academy, China;
3 State University of New York at Buffalo, USA;
4 the Chinese University of Hong Kong, Shenzhen, China; 5 the University of Sydney, Australia