Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning

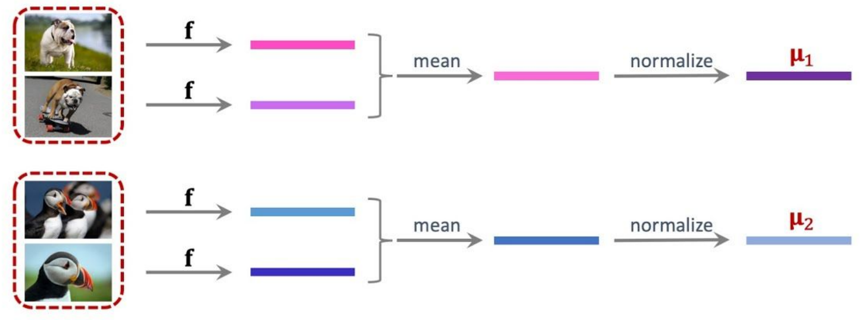
Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Fangzhou Mu, Yin Li, Yingyu Liang

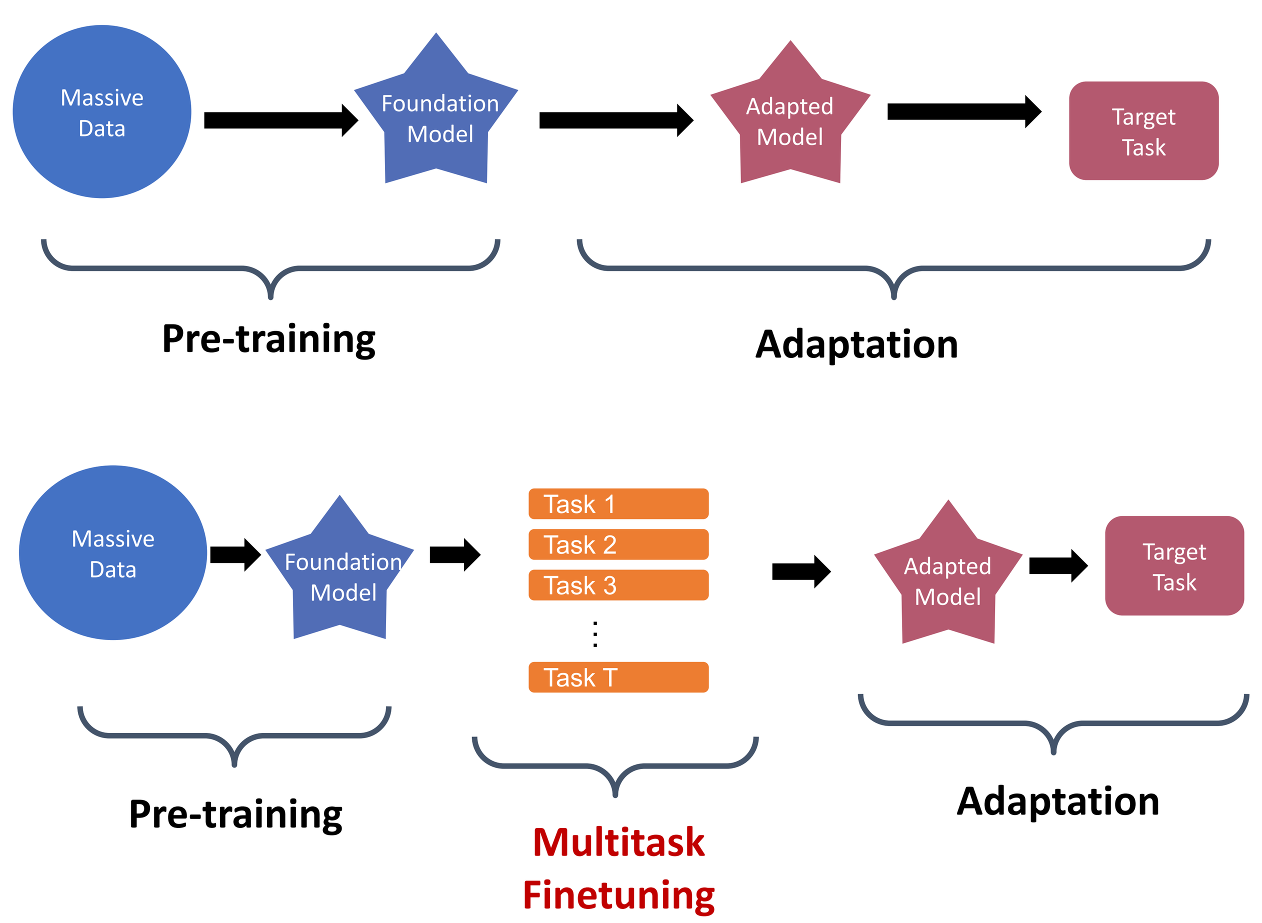
WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

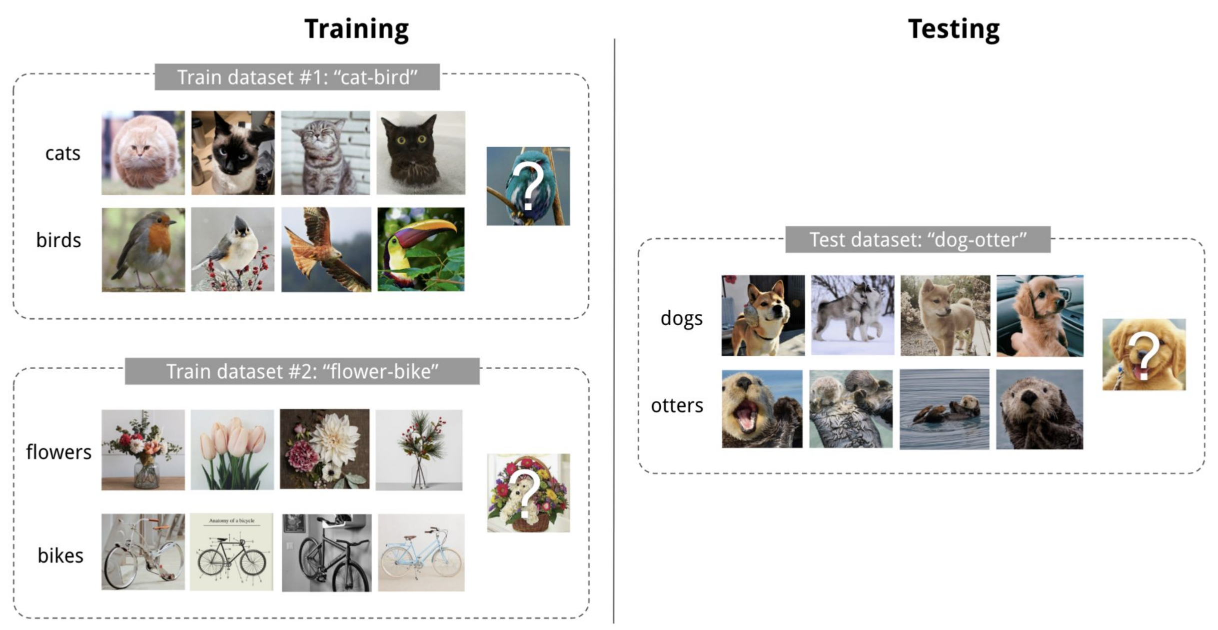
ICLR International Conference On Learning Representations

Motivation

Few-Shot Learning:
Pretraining + Fine Tuning

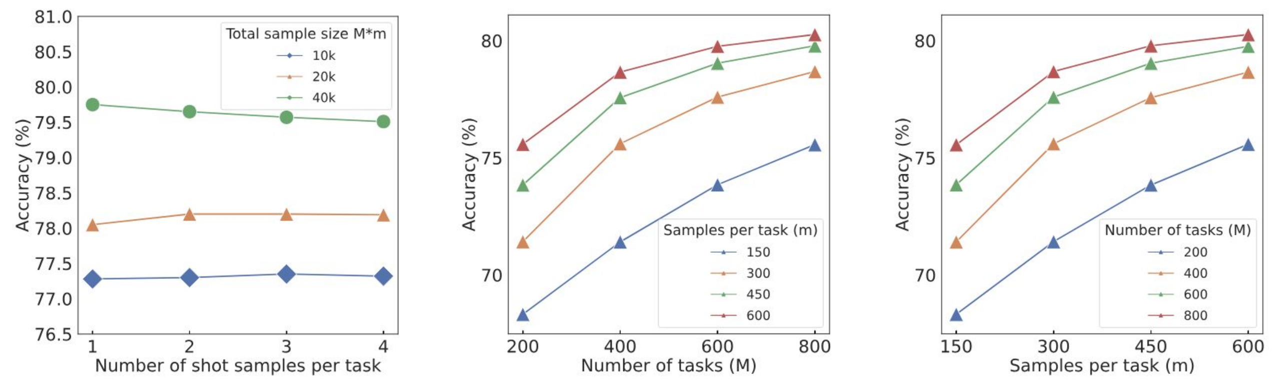






Experiments

Pretrained Method MoCo v3, DINO v2, supervised pretraining
Model ViT-S, ViT-B, ResNet50
Dataset miniImageNet, tieredImageNet, DomainNet, Meta-dataset

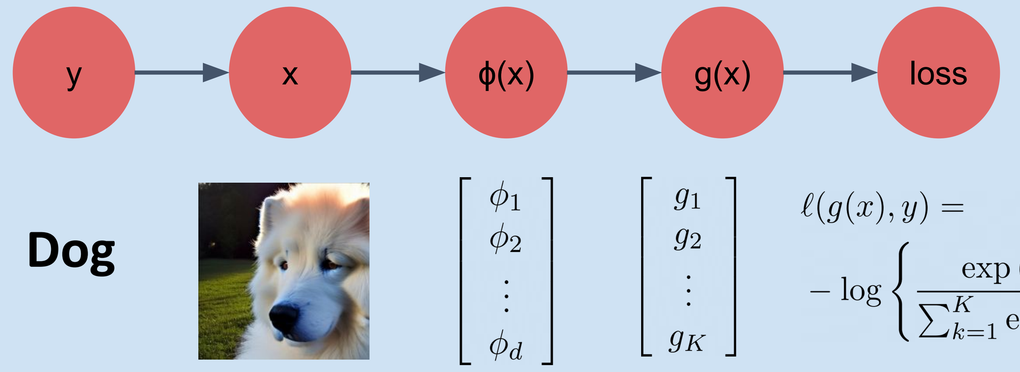


pretrained	backbone	method	miniImageNet 1-shot	miniImageNet 5-shot	tieredImageNet 1-shot	tieredImageNet 5-shot	DomainNet 1-shot	DomainNet 5-shot
MoCo v3	ViT-B	Adaptation	75.33 (0.30)	92.78 (0.10)	62.17 (0.36)	83.42 (0.23)	24.84 (0.25)	44.32 (0.29)
		Standard FT	75.38 (0.30)	92.80 (0.10)	62.28 (0.36)	83.49 (0.23)	25.10 (0.25)	44.76 (0.27)
		Ours	**80.62** (0.26)	**93.89** (0.09)	**68.32** (0.35)	**85.49** (0.22)	**32.88** (0.29)	**54.17** (0.30)
	ResNet50	Adaptation	68.80 (0.30)	88.23 (0.13)	55.15 (0.34)	76.00 (0.26)	27.34 (0.27)	47.50 (0.28)
		Standard FT	68.85 (0.30)	88.23 (0.13)	55.23 (0.34)	76.07 (0.26)	27.43 (0.27)	47.65 (0.28)
		Ours	**71.16** (0.29)	**89.31** (0.12)	**58.51** (0.35)	**78.41** (0.25)	**33.53** (0.30)	**55.82** (0.29)
DINO v2	ViT-S	Adaptation	85.90 (0.22)	95.58 (0.08)	74.54 (0.32)	89.20 (0.19)	52.28 (0.39)	72.98 (0.28)
		Standard FT	86.75 (0.22)	95.76 (0.08)	74.84 (0.32)	89.30 (0.19)	54.48 (0.39)	74.50 (0.28)
		Ours	**88.70** (0.22)	**96.08** (0.08)	**77.78** (0.32)	**90.23** (0.18)	**61.57** (0.40)	**77.97** (0.27)
	ViT-B	Adaptation	90.61 (0.19)	97.20 (0.06)	82.33 (0.30)	92.90 (0.16)	61.65 (0.41)	79.34 (0.25)
		Standard FT	91.07 (0.19)	97.32 (0.06)	82.40 (0.30)	93.07 (0.16)	61.84 (0.39)	79.63 (0.25)
		Ours	**92.77** (0.18)	**97.68** (0.06)	**84.74** (0.30)	**93.65** (0.16)	**68.22** (0.40)	**82.62** (0.24)
Supervised pretraining on ImageNet	ViT-B	Adaptation	94.06 (0.15)	97.88 (0.05)	83.82 (0.29)	93.65 (0.13)	28.70 (0.29)	49.70 (0.28)
		Standard FT	95.28 (0.13)	98.33 (0.04)	86.44 (0.27)	94.91 (0.12)	30.93 (0.31)	52.14 (0.29)
		Ours	**96.91** (0.11)	**98.76** (0.04)	**89.97** (0.25)	**95.84** (0.11)	**48.02** (0.38)	**67.25** (0.29)
	ResNet50	Adaptation	81.74 (0.24)	94.08 (0.09)	65.98 (0.34)	84.14 (0.21)	27.32 (0.27)	46.67 (0.28)
		Standard FT	84.10 (0.22)	94.81 (0.09)	74.48 (0.33)	88.35 (0.19)	34.10 (0.31)	55.08 (0.29)
		Ours	**87.61** (0.20)	**95.92** (0.07)	**77.74** (0.32)	**89.77** (0.17)	**39.09** (0.34)	**60.60** (0.29)

Problem Setup

Hidden representation data model

- Class $y \in \mathcal{C}$ over distribution $y \sim \eta$, sample $x \sim \mathcal{D}(y)$
- $\phi \in \Phi$ hypothesis class of representation functions: e.g. ResNet, ViT
- $g(x) = W\phi(x)$ as prediction logits of class



$$\begin{bmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_d\end{bmatrix} \quad \begin{bmatrix}g_1\\ g_2\\ \vdots\\ g_K\end{bmatrix} \quad \ell(g(x),y) = -\log\left\{\frac{\exp\left(g(\mathbf{x})_y\right)}{\sum_{k=1}^{K}\exp\left(g(\mathbf{x})_k\right)}\right\}$$

- **Contrastive Loss:** $\mathcal{L}_{con-pre} = \mathbb{E}_{x,y}\left[-\log\left(\frac{e^{\phi(x)^\top\phi(x^+)}}{e^{\phi(x)^\top\phi(x^+)}+e^{\phi(x)^\top\phi(x^-)}}\right)\right]$
- **Supervised Loss:** $\mathcal{L}_{sup-pre}(\phi) = \min_W \mathbb{E}_{x,y}[\ell(W\phi(x),y)]$

$\Big\}\ \mathcal{L}_{\text{pre}}(\phi)$

- Task $\mathcal{T} = (y_1,\ldots,y_K) \subseteq \mathcal{C}$, loss w.r.t a task: $\mathcal{L}_{\sup}(\mathcal{T},\phi) = \min_W \mathbb{E}_{x,y}[\ell(W\phi(x),y)]$
- **Multitask Finetuning: *M* tasks** $\{\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_M\}$**, each task with *m* sample** $\mathcal{S}_i := \{(x_j^i, y_j^i) : j \in [m]\}$

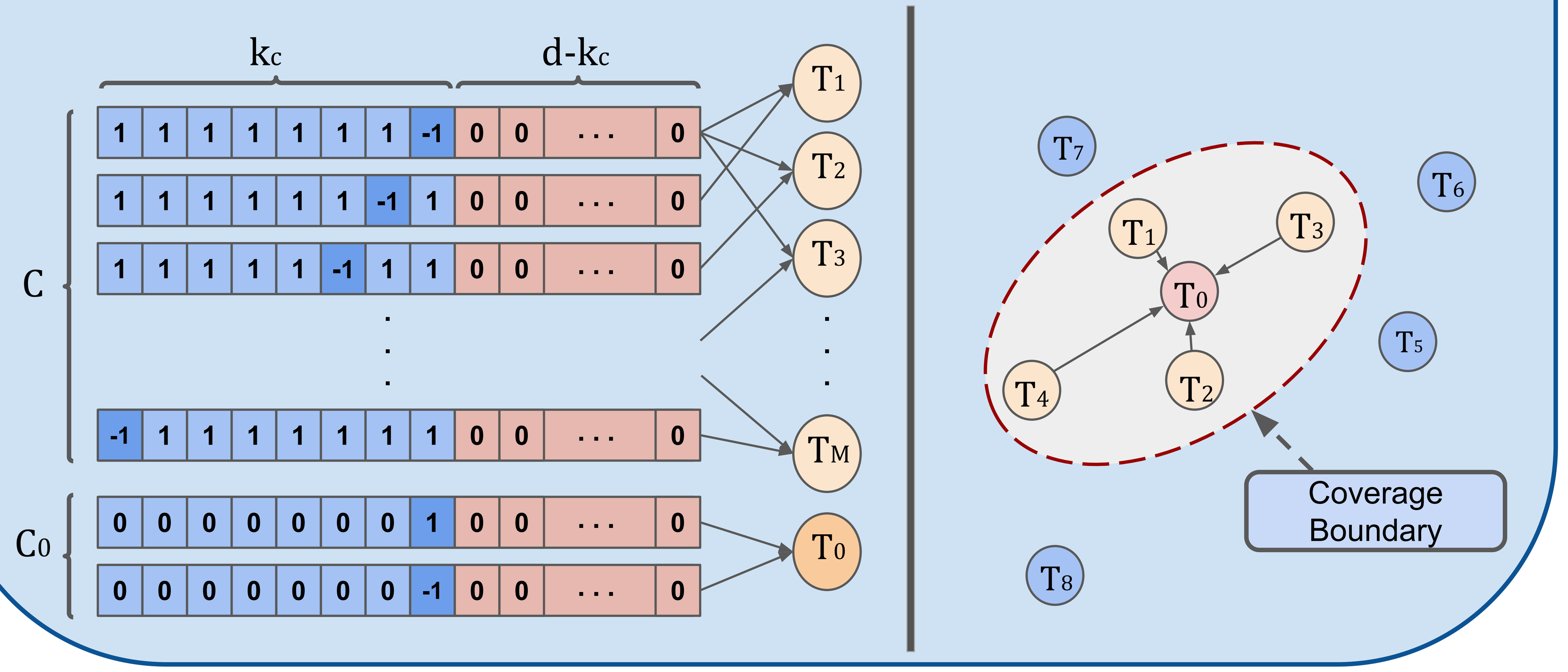
$$\min_{\phi\in\Phi}\frac{1}{M}\sum_{i=1}^{M}\widehat{\mathcal{L}}_{\sup}(\mathcal{T}_i,\phi), \quad \text{where } \widehat{\mathcal{L}}_{\sup}(\mathcal{T}_i,\phi) := \min_{W_i\in\mathbb{R}^d}\frac{1}{m}\sum_{j=1}^{m}\ell\left(W_i^\top\phi\left(x_j^i\right),y_j^i\right)$$

- Suppose target task is $\mathcal{T}_0$
- Let $\phi^* \in \Phi$ denote the model with the lowest target task loss
- Denote error on target task $\mathcal{E}(\phi) = \mathcal{L}_{\sup}(\mathcal{T}_0,\phi) - \mathcal{L}_{\sup}(\mathcal{T}_0,\phi^*)$

Theoretical Analysis

Definition 1 (Diversity and Consistency (Informal))
Consider the latent feature space. **Diversity** refer to the coverage of the finetuning tasks on the target task. **Consistency** refer to similarity.

Theorem (Multitask finetuning loss (Informal))
Suppose in pretraining we have empirical pretraining loss $\hat{\mathcal{L}}_{\text{pre}}(\hat{\phi}) \le \epsilon_0$
The error will be $\mathcal{E}(\hat{\phi}) \le \mathcal{O}(\epsilon_0)$. After sufficient multitask finetuning and get ϕ', the error will be $\mathcal{E}(\phi') \le \mathcal{O}(\alpha\epsilon_0)$ with high probability. The finetuning sample complexity will be $\Omega\left(\frac{1}{\alpha\epsilon_0}\right)$.



Practical Solution

Pretrained	Selection	INet	Omglot	Acraft	CUB	QDraw	Fungi	Flower	Sign	COCO
CLIP	Random	56.29	65.45	31.31	59.22	36.74	31.03	75.17	33.21	30.16
	No Con.	60.89	72.18	31.50	66.73	40.68	35.17	81.03	37.67	34.28
	No Div.	56.85	73.02	32.53	65.33	40.99	33.10	80.54	34.76	31.24
	Selected	**60.89**	**74.33**	**33.12**	**69.07**	**41.44**	**36.71**	**80.28**	**38.08**	**34.52**
DINOv2	Random	83.05	62.05	36.75	93.75	39.40	52.68	98.57	31.54	47.35
	No Con.	83.21	76.05	36.32	93.96	50.76	53.01	98.58	34.22	47.11
	No Div.	82.82	79.23	36.33	93.96	55.18	52.98	98.59	35.67	44.89
	Selected	**83.21**	**81.74**	**37.01**	**94.10**	**55.39**	**53.37**	**98.65**	**36.46**	**48.08**
MoCo v3	Random	59.66	60.72	18.57	39.80	40.39	32.79	58.42	33.38	32.98
	No Con.	59.80	60.79	18.75	40.41	40.98	32.80	59.55	34.01	33.41
	No Div.	59.57	63.00	18.65	40.36	41.04	32.80	58.67	34.03	33.67
	Selected	**59.80**	**63.17**	**18.80**	**40.74**	**41.49**	**33.02**	**59.64**	**34.31**	**33.86**

Table 1: Results evaluating our task selection algorithm on Meta-dataset using ViT-B backbone. No Con.: Ignore consistency. No Div.: Ignore diversity. Random: Ignore both consistency and diversity.

Take-Home Message

We provide the theoretical justification and practical solution of multitask finetuning for adapting pretrained foundation models to downstream tasks with limited labels.

Key Intuition

1. Pre-training uses unlabeled and noisy data for general purpose learning, where the model learns general structure rather than task-specific knowledge. Its performance on a specific task may not be perfect.
2. Despite the target data is limited, we have a clear understanding of the target task and its associated data.
   - We actively select extra data from a relevant source that covers target data characteristic features.
   - We then design specialized tasks for multitask finetuning to equip the model to address the specific types of target tasks effectively.