Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning

Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Fangzhou Mu, Yin Li, Yingyu Liang
University of Wisconsin - Madison

IBM Research Seminar

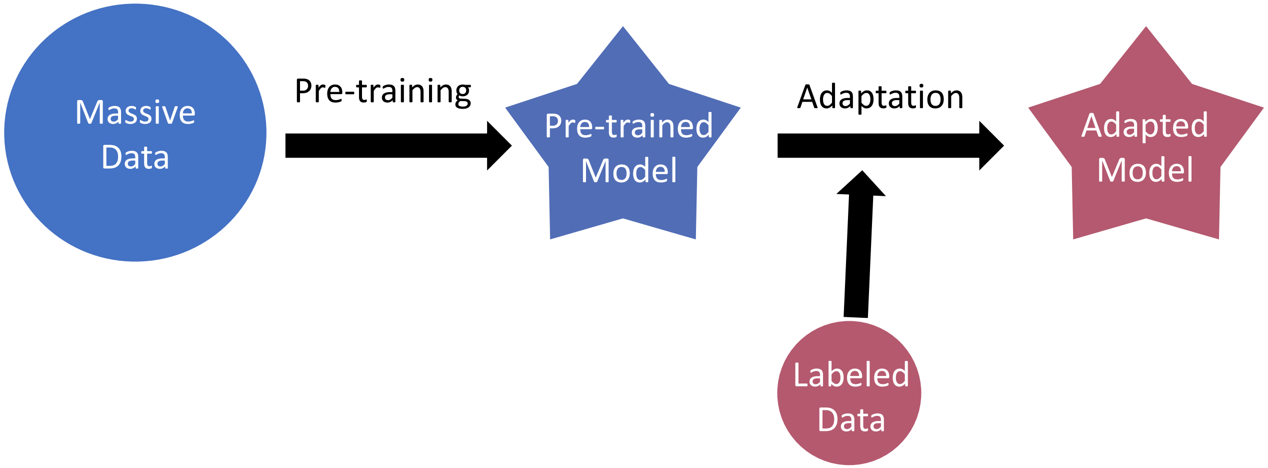
ICLR

New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning $\Longrightarrow$ pre-training + adaptation

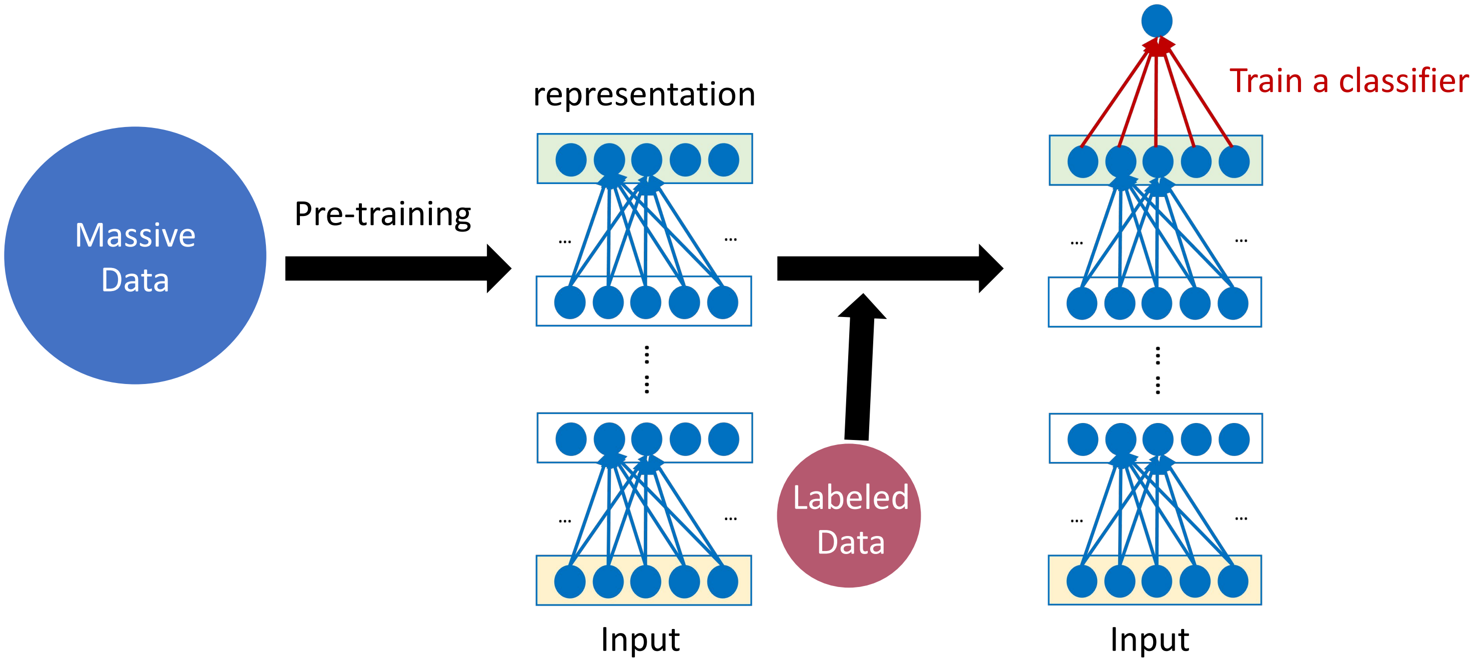
New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ⟹ pre-training + adaptation



New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ⟹ pre-training + adaptation



New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ⟹ pre-training + adaptation

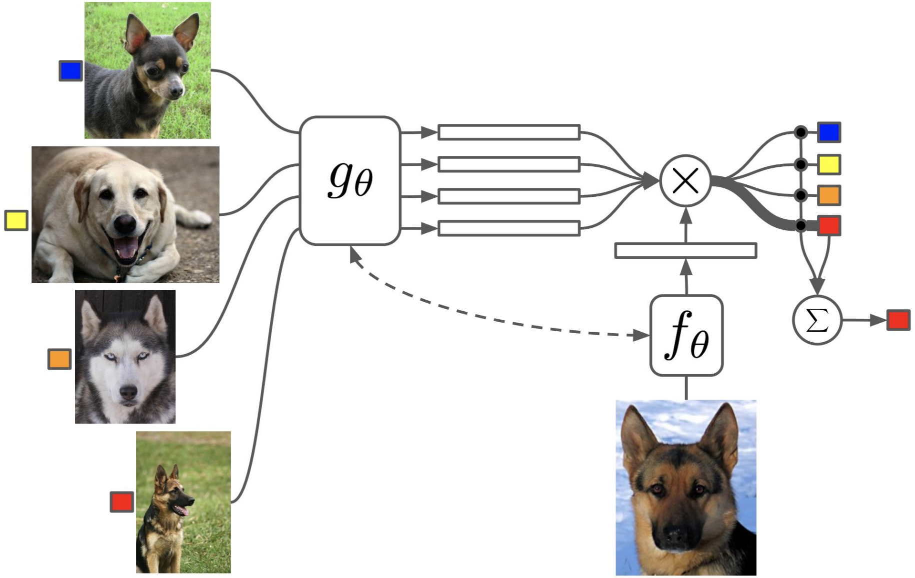


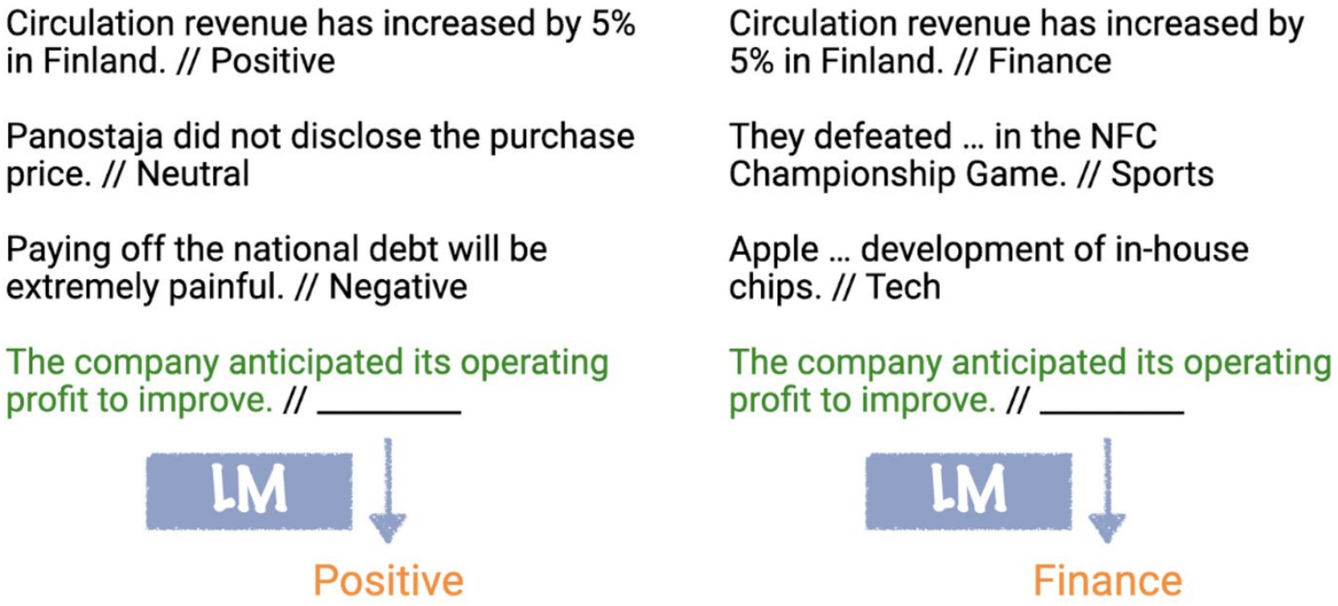
Figure 1: Matching Networks architecture

Adaptation of a pre-trained image encoder

Figures from: *Matching Networks for One Shot Learning, 2017.*

New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ⟹ pre-training + adaptation

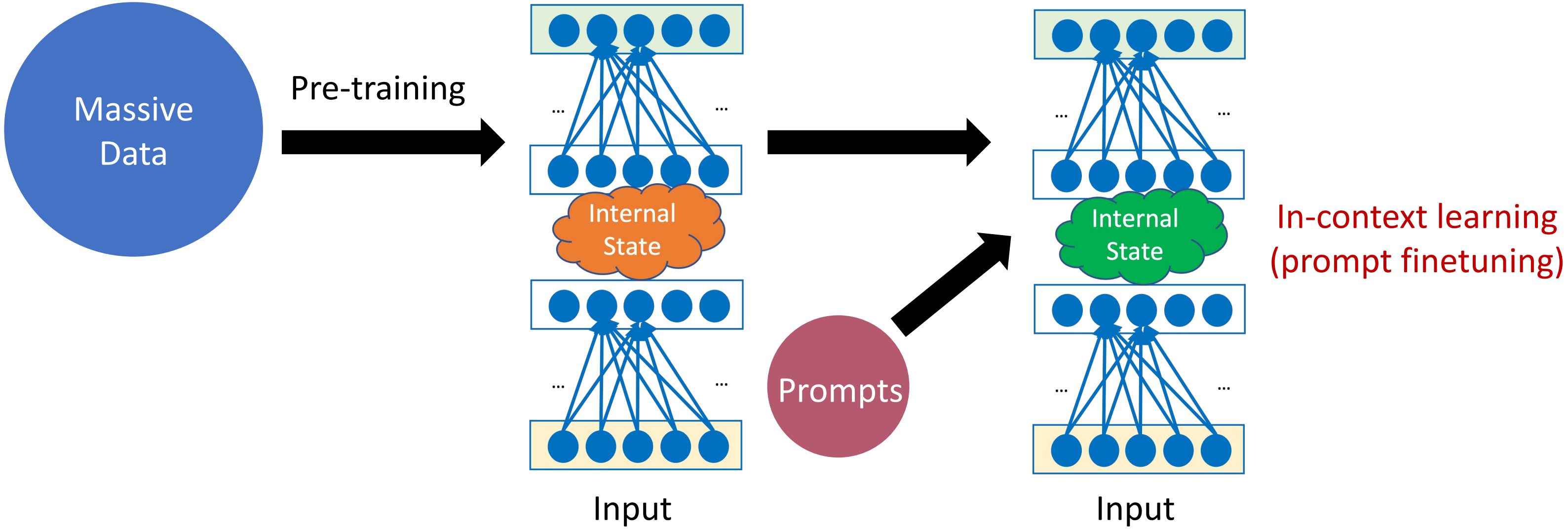


Adaptation of a pre-trained language decoder

Figures from: *How does in-context learning work? A framework for understanding the differences from traditional supervised learning, 2022.*

New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ⟹ pre-training + adaptation

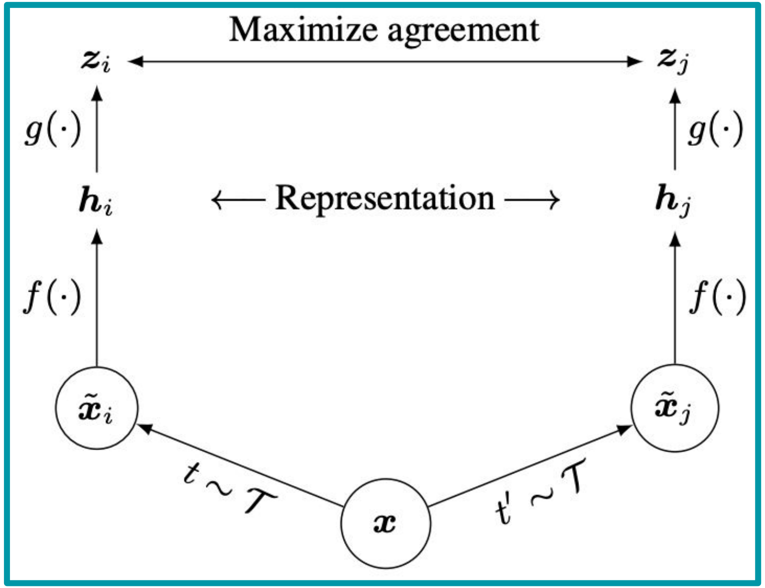


What does pre-training look like?

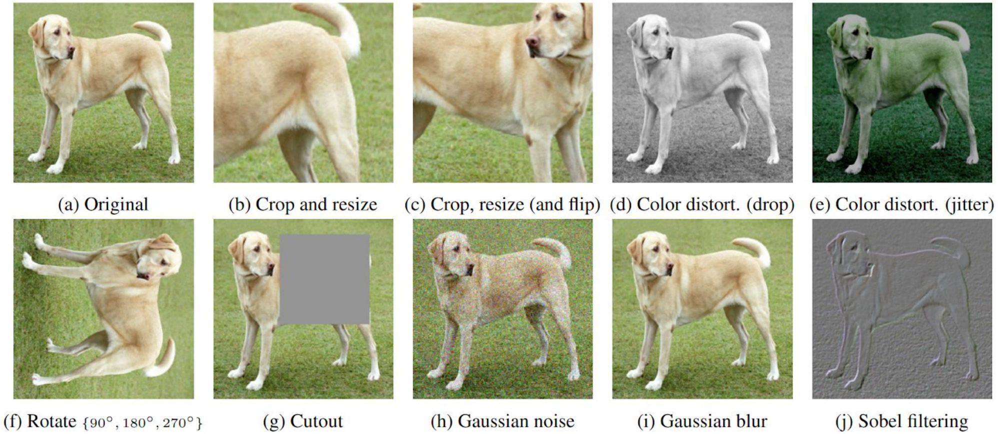
- Supervised learning

- Self-supervised learning:
    - Next sentence prediction (BERT)
    - Masked language prediction (BERT, RoBERTa)
    - Auto-regressive language modeling (GPT, Llama)
    - Contrastive learning (SimCLR, SimCSE, CLIP, DINO)

Intro - Contrastive Learning



$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$$



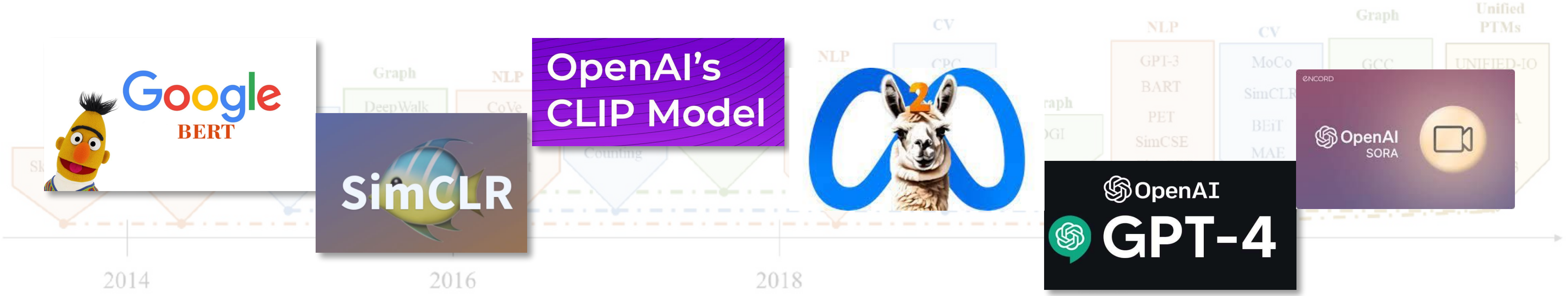
SimCLR - (Image, Image)
No need labels

Figures from: *A Simple Framework for Contrastive Learning of Visual Representations, 2020*

Image Data Augmentation

Figures from: *A Simple Framework for Contrastive Learning of Visual Representations, 2020*

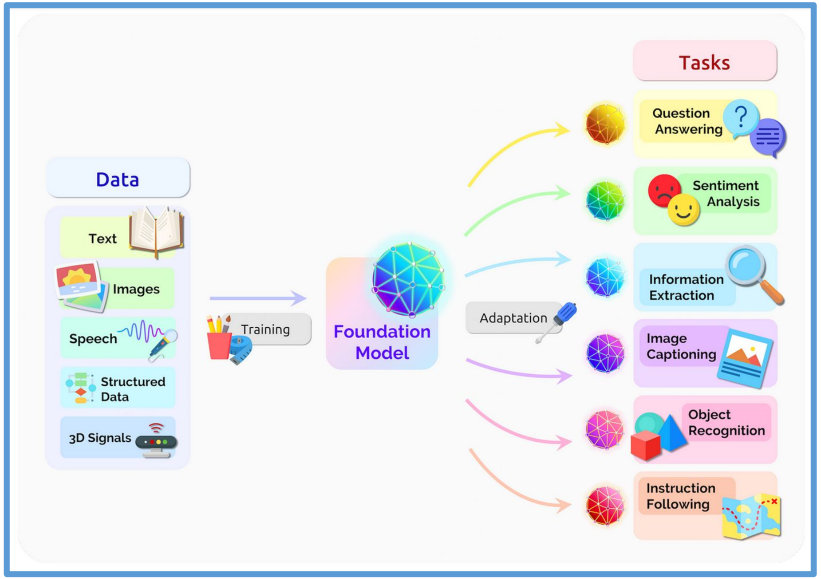
Intro - Foundation Model



The history and evolution of foundation models

Figures from: *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT, 2023.*

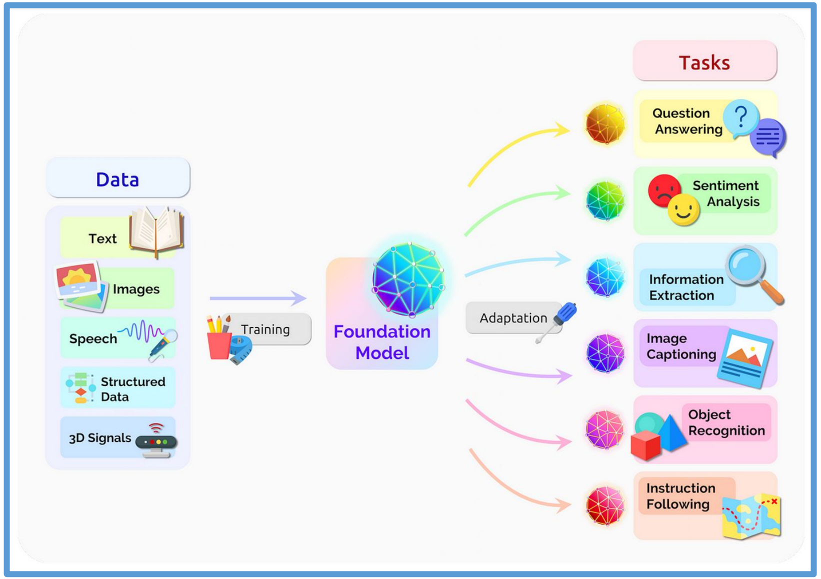
Intro - Foundation Model

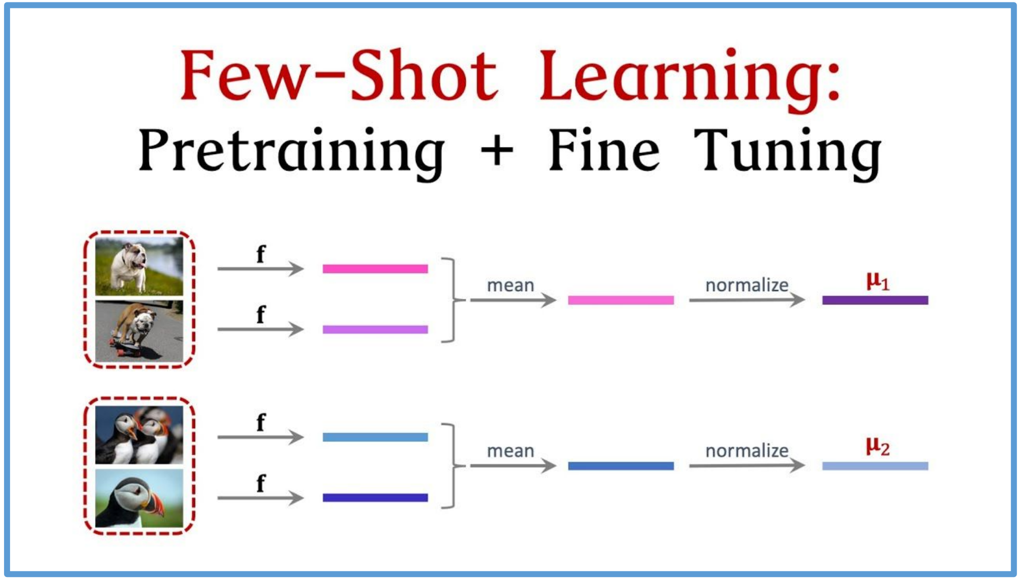


Universality

Figures from: *On the opportunities and risks of foundation models, 2021.*

Intro - Foundation Model





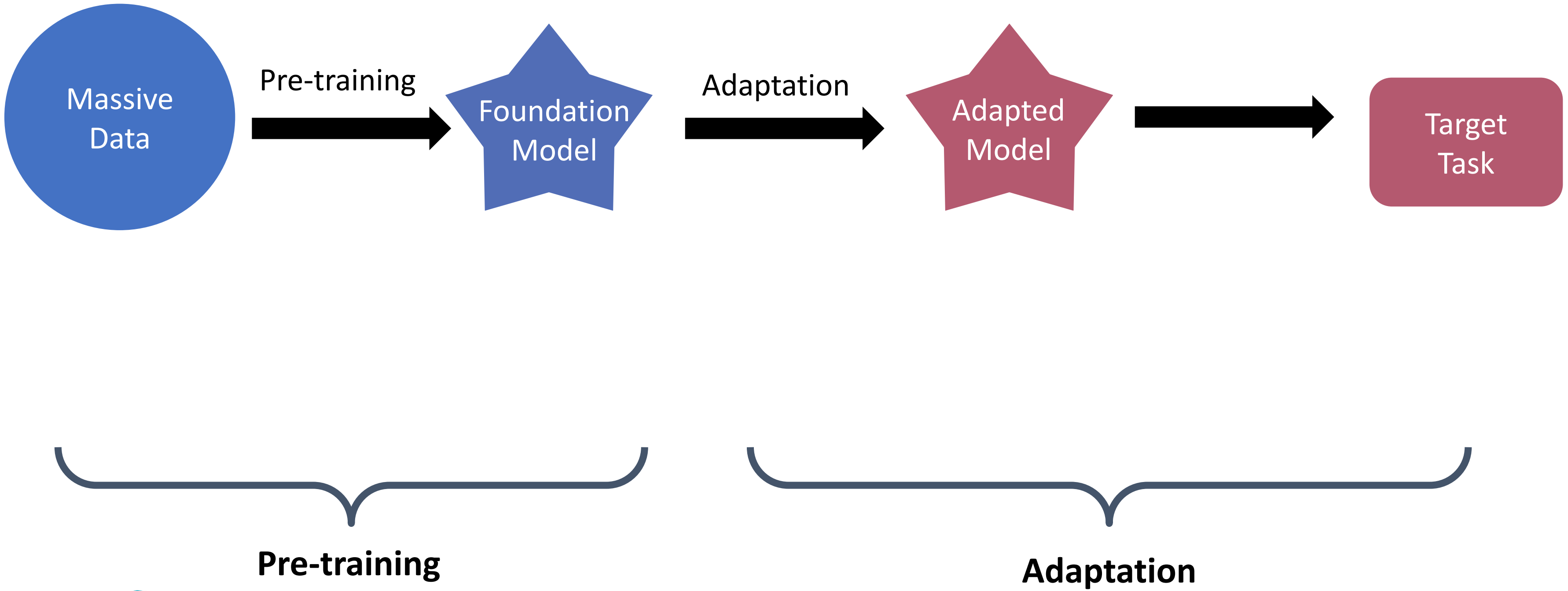
Universality

Figures from: *On the opportunities and risks of foundation models, 2021.*

Label Efficiency

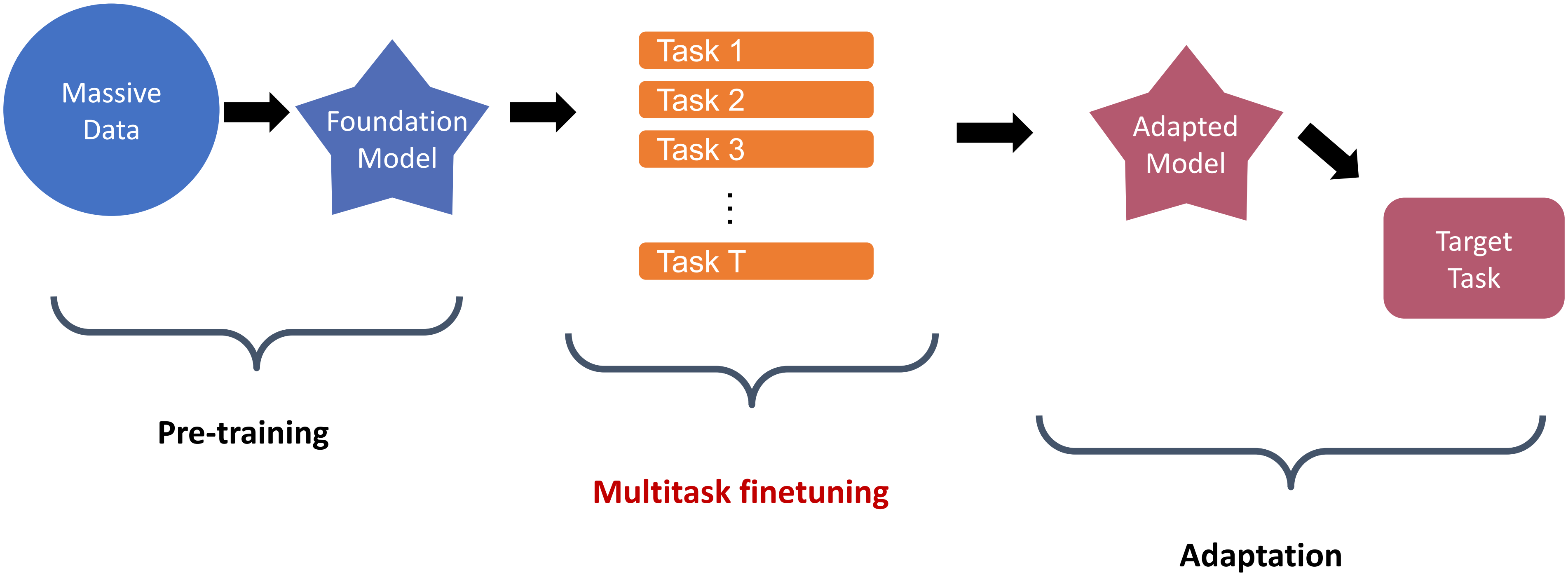
Figures from: https://www.youtube.com/watch?v=U6uFOIURcD0&ab_channel=ShusenWang, 2020

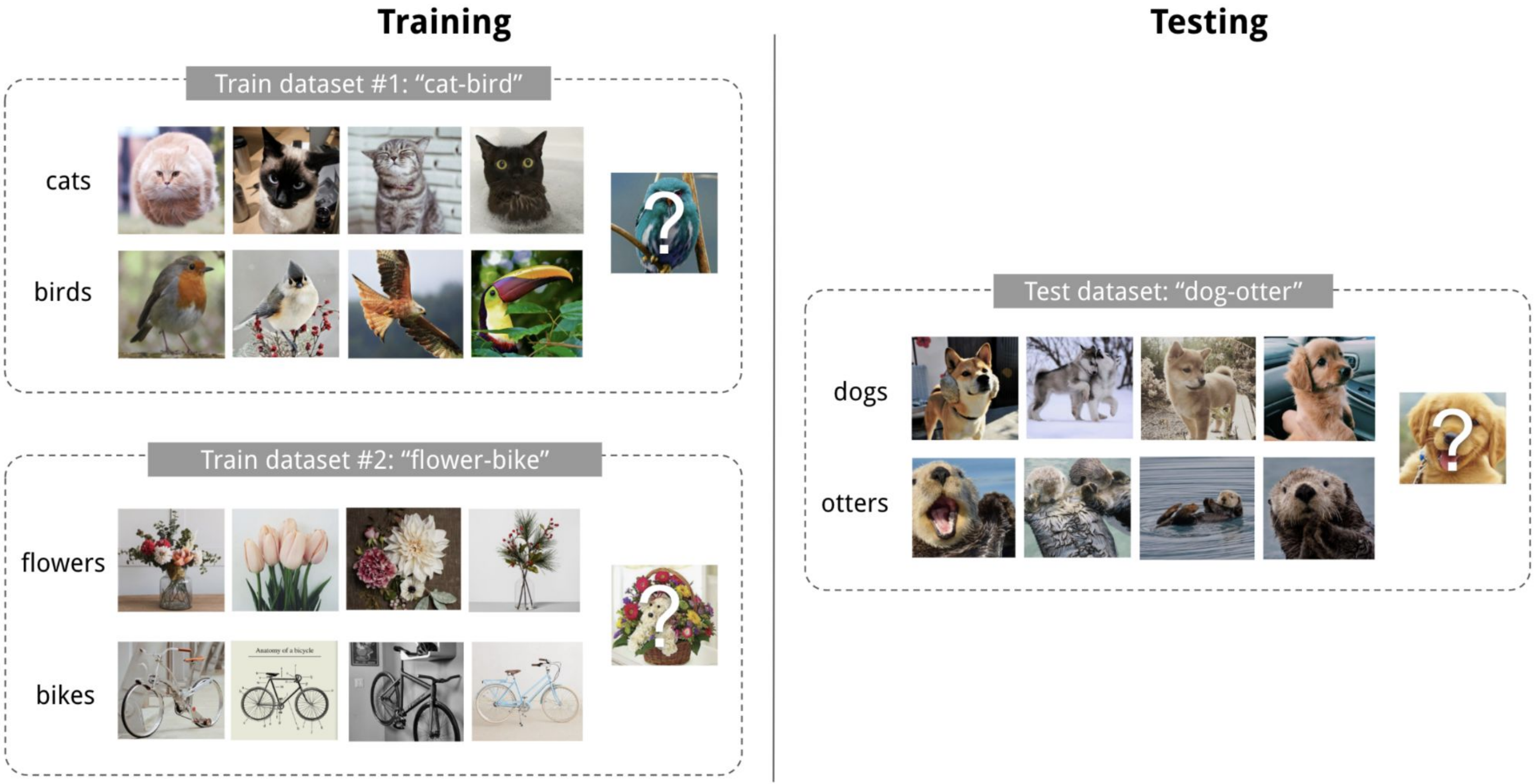
Paradigm: Pre-training + Adaptation



Q: Can we improve this?

Pre-training + Finetuning + Adaptation





An example of 4-shot 2-class image classification

Figures from: *Meta-Learning: Learning to Learn Fast*, *2018.*

Problem Setup - Hidden representation data model

- Class $y \in \mathcal{C}$ over distribution $y \sim \eta$
- Task $\mathcal{T} = (y_1, \dots, y_K) \subseteq \mathcal{C}$, 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 latent class

$$y \rightarrow x \rightarrow \phi(x) \rightarrow g(x) \rightarrow \text{loss}$$

Dog

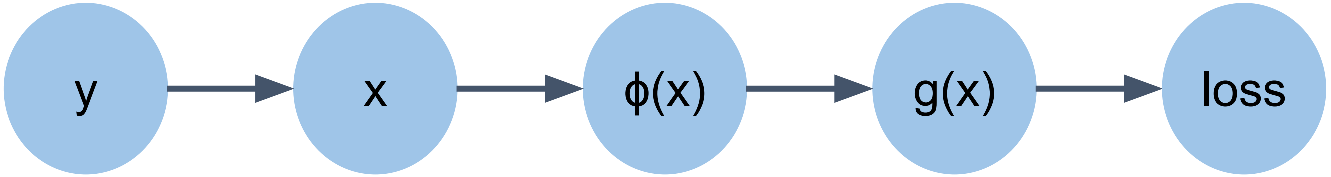
$$\begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_d \end{bmatrix} \qquad \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_K \end{bmatrix}$$

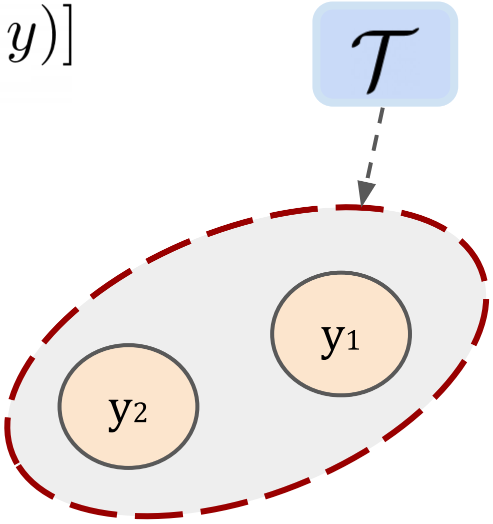
$$\ell(g(x), y) = -\log\left\{\frac{\exp\left(g(\mathbf{x})_y\right)}{\sum_{k=1}^{K} \exp\left(g(\mathbf{x})_k\right)}\right\}$$

Problem Setup - Objective for a downstream task

- Class $y \in \mathcal{C}$ over distribution $y \sim \eta$
- Task $\mathcal{T} = \{y_1, y_2\} \subseteq \mathcal{C}$, instance $x \sim \mathcal{D}(y)$
- $g(x) = W\phi(x)$ as prediction logits of latent class
- supervised loss w.r.t a task:

$$\mathcal{L}_{\text{sup}}(\mathcal{T}, \phi) := \min_{W} \mathop{\mathbb{E}}_{y \sim \mathcal{T}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}(y)} [\ell(W\phi(x), y)]$$

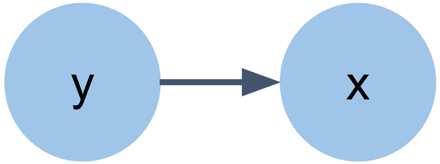




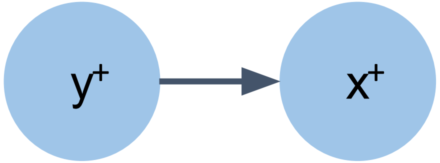
Pretraining - Contrastive learning

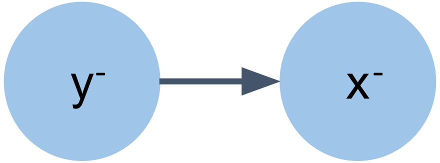
- $(y, y^-) \sim \eta^2, x, x^+ \sim \mathcal{D}(y), x^- \sim \mathcal{D}(y^-), \tau := \Pr_{(y,y^-)\sim\eta^2}\{y = y^-\}$
- Contrastive loss:

$$\mathbb{E}\left[-\log\left(\frac{e^{\phi(x)^\top \phi(x^+)}}{e^{\phi(x)^\top \phi(x^+)} + e^{\phi(x)^\top \phi(x^-)}}\right)\right]$$



positive pair





negative pair

Data Model

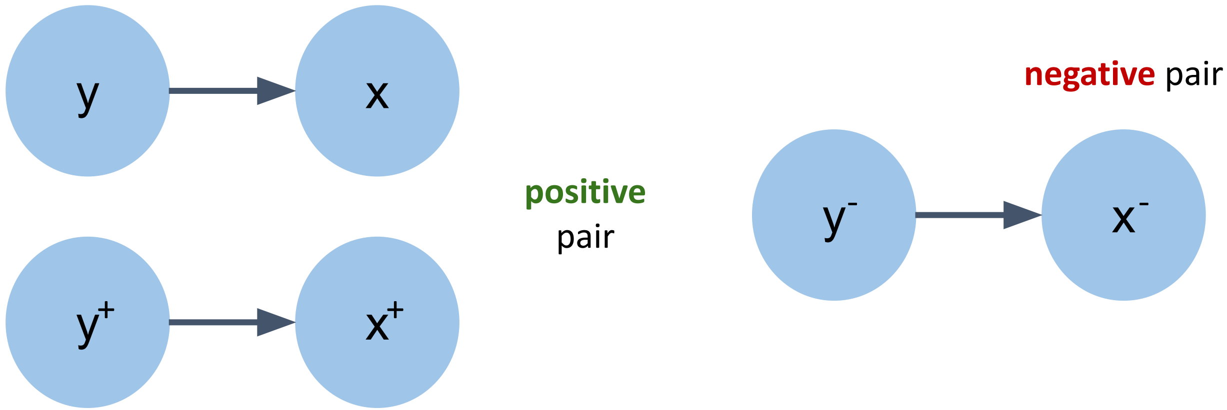
Figures from: *Expanding Small-Scale Datasets with Guided Imagination, 2023*

Pretraining - Contrastive learning

- $(y, y^-) \sim \eta^2, \; x, x^+ \sim \mathcal{D}(z), \; x^- \sim \mathcal{D}(z^-)$
- Contrastive loss: $\mathcal{L}_{con-pre}(\phi) := \mathbb{E}\left[\ell_u\left(\phi(x)^\top\left(\phi\left(x^+\right) - \phi\left(x^-\right)\right)\right)\right]$

$$\widehat{\mathcal{L}}_{con-pre}(\phi) := \frac{1}{N}\sum_{i=1}^{N}\left[\ell_u\left(\phi(x_i)^\top\left(\phi\left(x_i^+\right) - \phi\left(x_i^-\right)\right)\right)\right]$$

- In particular: $\ell_u(v) = \log(1 + \exp(-v))$ will recover the contrastive loss in previous slide

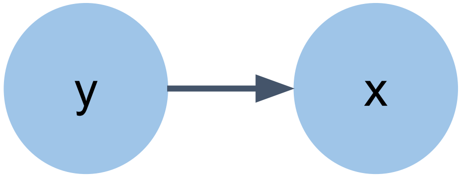


Pretraining - Supervised learning

- $y \sim \eta$, $x \sim \mathcal{D}(y)$
- supervised loss: $\ell(g(x), y) = \ell_u\left((g(x))_y - (g(x))_{y' \neq y, y' \in \mathcal{C}}\right)$

$$\mathcal{L}_{sup-pre}(\phi) = \min_W \mathbb{E}_{x,y}[\ell(W\phi(x), y)]$$

- In particular: $\ell_u(v) = \log(1 + \exp(-v))$ will recover the logistic loss

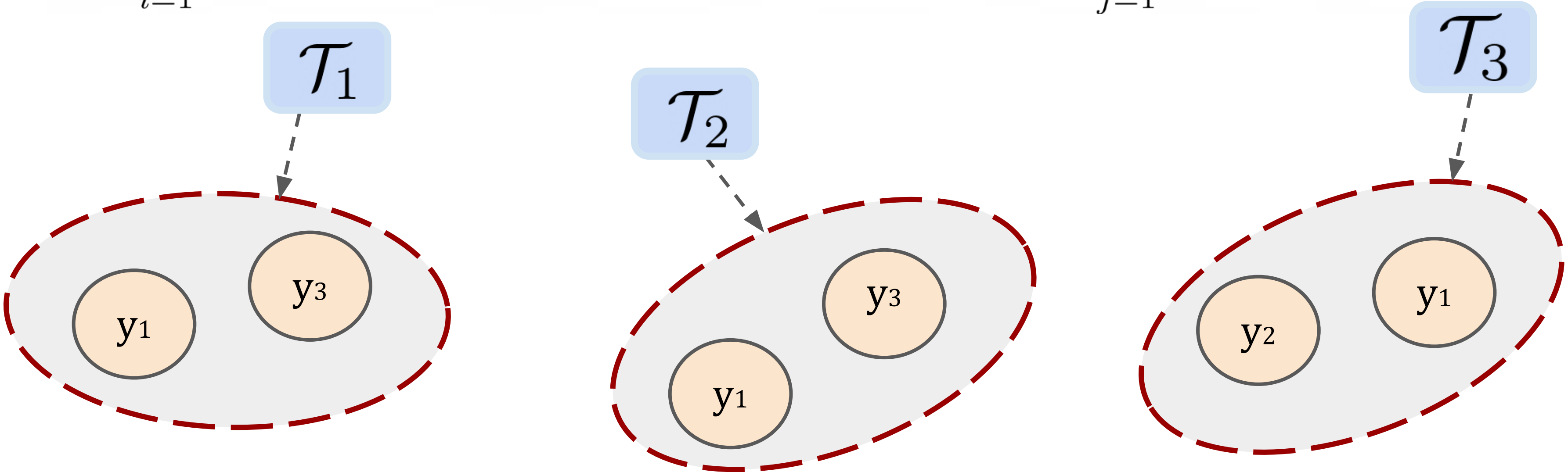


To simplify notation, we will use $\mathcal{L}_{pre}(\phi)$, we denote pretrained model as $\hat{\phi}$

Problem Setup - Multitask Finetuning

- Suppose we construct M tasks $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_M\}$
- Suppose each task with m sample $\mathcal{S}_i := \{(x_j^i, y_j^i) : j \in [m]\}$
- Given pretrained $\hat{\phi}$. We further multitask finetune it by objective:

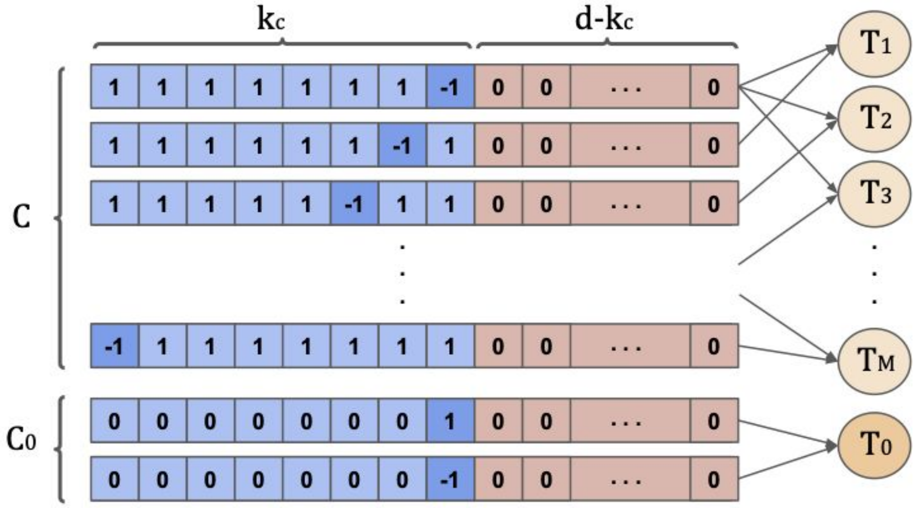
$$\min_{\phi \in \Phi} \frac{1}{M} \sum_{i=1}^{M} \widehat{\mathcal{L}}_{\text{sup}} \left(\mathcal{T}_i, \phi\right), \quad \text{where } \widehat{\mathcal{L}}_{\text{sup}} \left(\mathcal{T}_i, \phi\right) := \min_{W_i \in \mathbb{R}^d} \frac{1}{m} \sum_{j=1}^{m} \ell\left(W_i^\top \phi\left(x_j^i\right),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



Diversity and Consistency

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

- Suppose target task is $\mathcal{T}_0$



Main Result

- Suppose target task is $\mathcal{T}_0$
- Let $\phi^* \in \Phi$ denote the model with the lowest target task loss $\mathcal{L}_{sup}(\mathcal{T}_0, \phi^*)$
- We want to bound $\mathcal{E}(\phi) = \mathcal{L}_{\text{sup}}(\mathcal{T}_0, \phi) - \mathcal{L}_{\text{sup}}(\mathcal{T}_0, \phi^*)$
- Pretraining loss as $\hat{\mathcal{L}}_{\text{pre}}(\hat{\phi})$

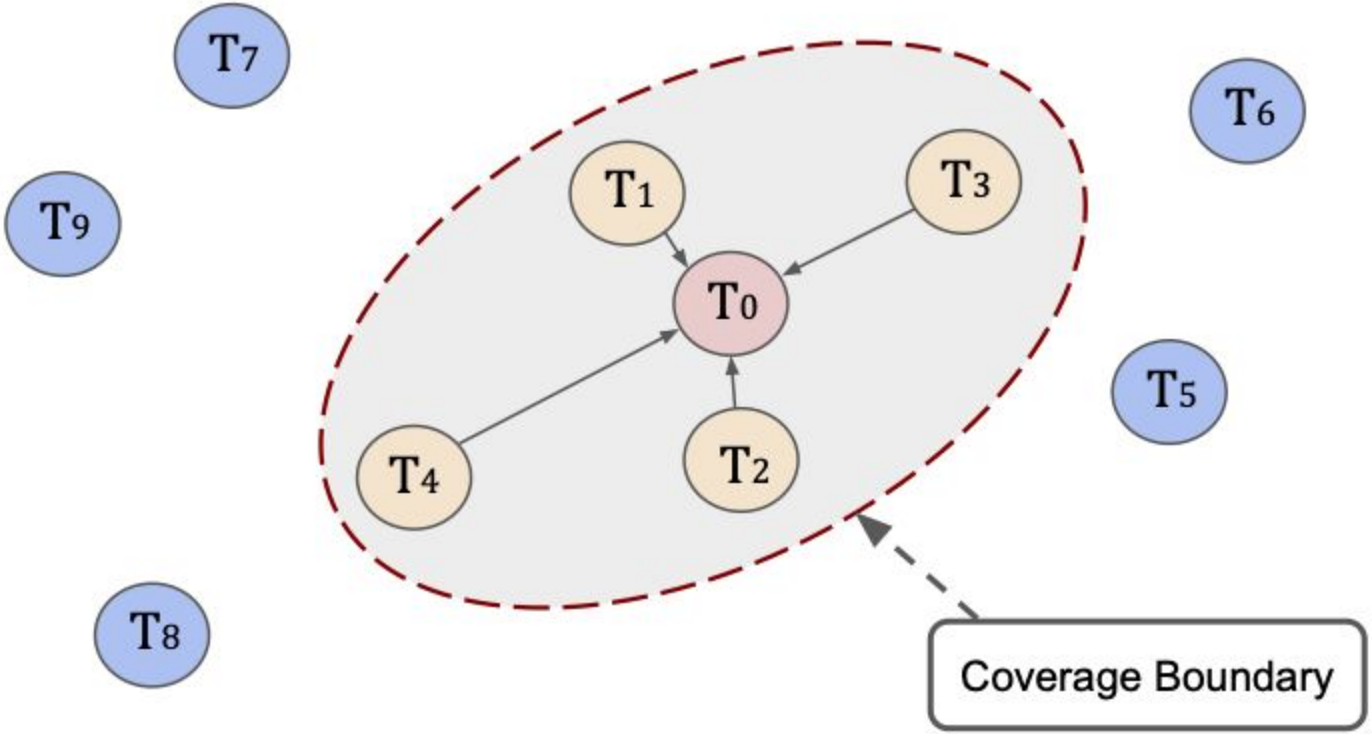
Theorem (Multitask finetuning loss (Informal))
Suppose in pretraining we have empirical pretraining loss $\hat{\mathcal{L}}_{\text{pre}}(\hat{\phi}) \leq \epsilon_0$
The error will be $\mathcal{E}(\hat{\phi}) \leq \mathcal{O}(\epsilon_0)$. After sufficient multitask finetuning and get ϕ', the error will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

Remark

- Comparing to pretraining + adaptation (baseline), the multitask fineutning procedure reduce error on target task by $(1-\alpha)\frac{2\epsilon_0}{1-\tau}$ with required sample complexity $\Omega\left(\frac{1}{\alpha\epsilon_0}\right)$

- Ideally, data from the finetuning tasks should satisfy two requirements:
  - **Consistency**: finetuning tasks similar the target task,
  - **Diversity**: finetuning tasks are sufficiently diverse to cover a wide range of patterns that may be encountered in the target task.

Practical solution: Task selection



Practical solution: Task selection

Algorithm 1 Consistency-Diversity Task Selection

Input: Target task $\mathcal{T}_0$, candidate finetuning tasks: $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_M\}$, model ϕ, threshold p.

1: Compute $\phi(\mathcal{T}_i)$ and $\mu_{\mathcal{T}_i}$ for $i = 0, 1, \ldots, M$.

2: Sort $\mathcal{T}_i$'s in descending order of similarity $(\mathcal{T}_0, \mathcal{T}_i)$. Denote the sorted list as $\{\mathcal{T}_1', \mathcal{T}_2', \ldots, \mathcal{T}_M'\}$.

3: $L \leftarrow \{\mathcal{T}_1'\}$

4: **for** $i = 2, \ldots, M$ **do**

5: If coverage$(L \cup \mathcal{T}_i'; \mathcal{T}_0) \geq (1 + p) \cdot$ coverage$(L; \mathcal{T}_0)$, then $L \leftarrow L \cup \mathcal{T}_i'$; otherwise, break.

6: **end for**

Output: selected data L for multitask finetuning.

Experiments: Few-shot Vision tasks

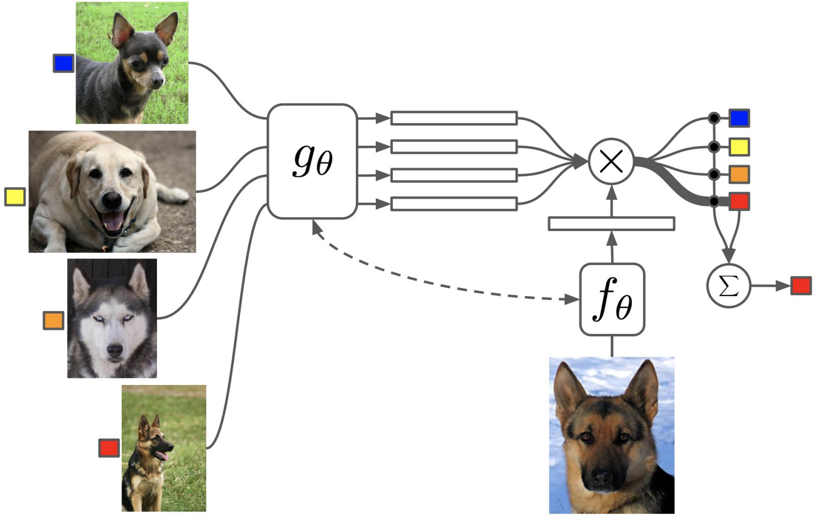
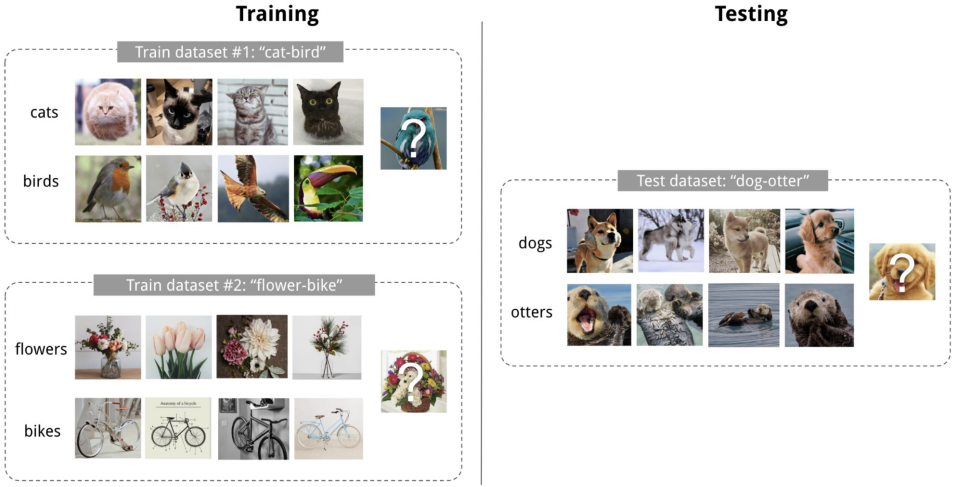
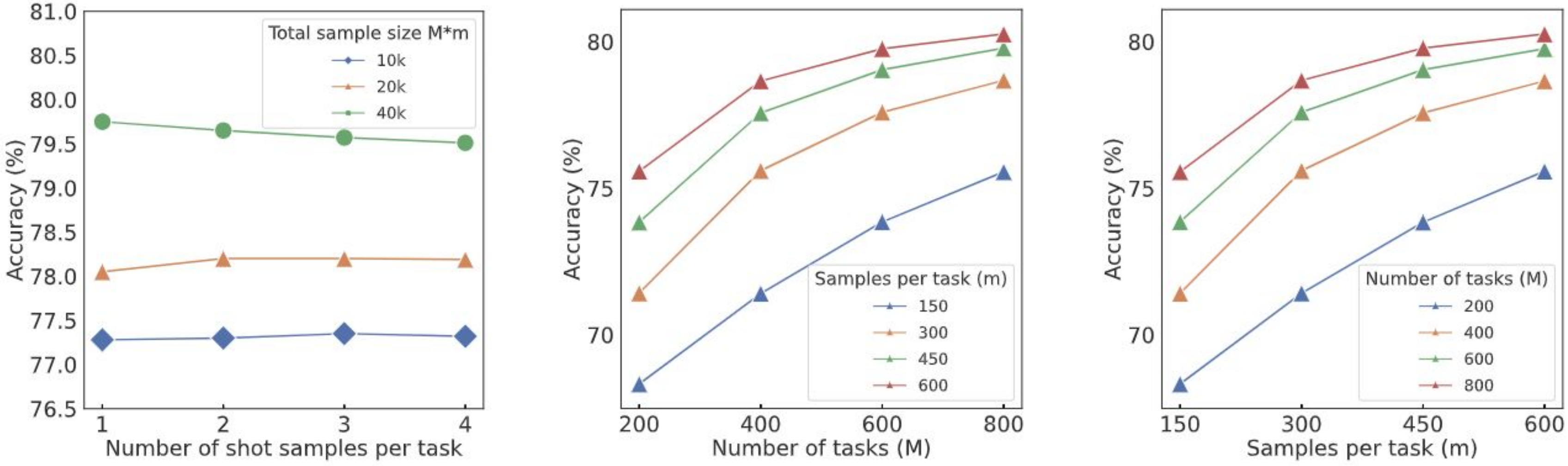


Figure 1: Matching Networks architecture



Experiments: Verification of Theoretical Analysis



(a) # shots during finetuning. (b) # tasks during finetuning. (c) # samples during finetuning.

Figure 3: Results on ViT-B backbone pretrained by MoCo v3. (a) Accuracy v.s. number of shots per finetuning task. Different curves correspond to different total numbers of samples Mm. (b) Accuracy v.s. the number of tasks M. Different curves correspond to different numbers of samples per task m. (c) Accuracy v.s. number of samples per task m. Different curves correspond to different numbers of tasks M.

Experiments: Task selection algorithm

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.

Experiments: Effectiveness of Multitask Finetuning

			miniImageNet		tieredImageNet		DomainNet	
pretrained	**backbone**	**method**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**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)

Table 2: **Results of few-shot image classification.** We report average classification accuracy (%) with 95% confidence intervals on test splits. Adaptation: Direction adaptation without finetuning; Standard FT: Standard finetuning; Ours: Our multitask finetuning; 1-/5-shot: number of labeled images per class in the target task.

Experiments: Few-shot Language task

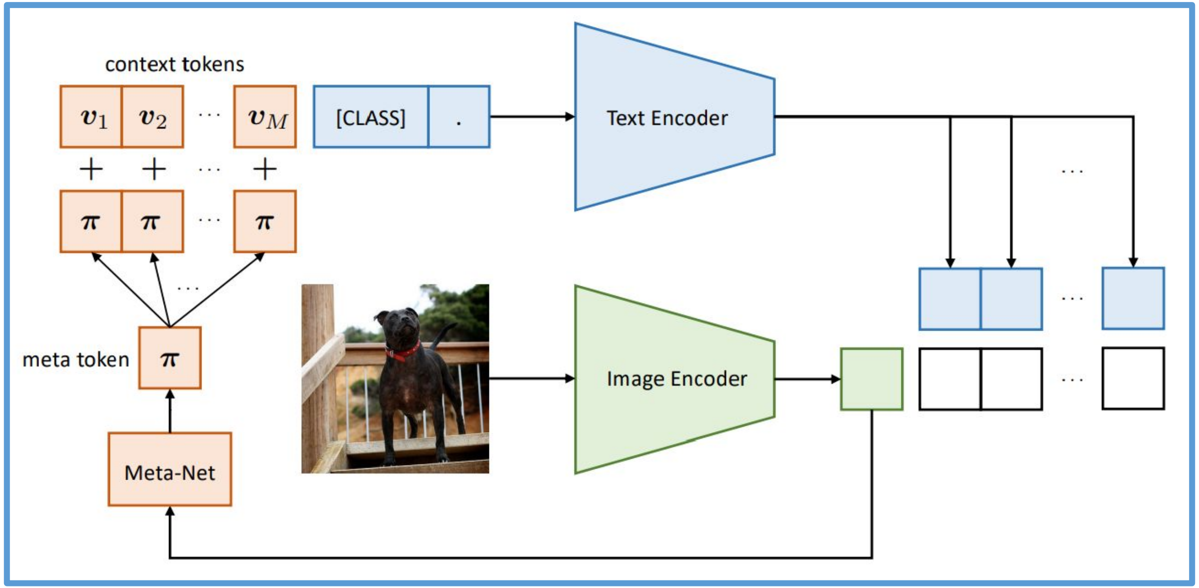
	SST-2 (acc)	SST-5 (acc)	MR (acc)	CR (acc)	MPQA (acc)	Subj (acc)	TREC (acc)	CoLA (Matt.)
Prompt-based zero-shot	83.6	35.0	80.8	79.5	67.6	51.4	32.0	2.0
Multitask FT zero-shot	**92.9**	37.2	86.5	88.8	73.9	55.3	36.8	-0.065
+ task selection	92.5	34.2	87.1	88.7	71.8	72.0	36.8	0.001
Prompt-based FT†	92.7 (0.9)	47.4 (2.5)	87.0 (1.2)	90.3 (1.0)	84.7 (2.2)	**91.2** (1.1)	84.8 (5.1)	**9.3** (7.3)
Multitask Prompt-based FT	92.0 (1.2)	**48.5** (1.2)	86.9 (2.2)	90.5 (1.3)	**86.0** (1.6)	89.9 (2.9)	83.6 (4.4)	5.1 (3.8)
+ task selection	92.6 (0.5)	47.1 (2.3)	**87.2** (1.6)	**91.6** (0.9)	85.2 (1.0)	90.7 (1.6)	**87.6** (3.5)	3.8 (3.2)
	MNLI (acc)	**MNLI-mm** (acc)	**SNLI** (acc)	**QNLI** (acc)	**RTE** (acc)	**MRPC** (F1)	**QQP** (F1)	
Prompt-based zero-shot	50.8	51.7	49.5	50.8	51.3	61.9	49.7	
Multitask FT zero-shot	63.2	65.7	61.8	65.8	74.0	81.6	63.4	
+ task selection	62.4	64.5	65.5	61.6	64.3	75.4	57.6	
Prompt-based FT†	68.3 (2.3)	70.5 (1.9)	77.2 (3.7)	64.5 (4.2)	69.1 (3.6)	74.5 (5.3)	65.5 (5.3)	
Multitask Prompt-based FT	70.9 (1.5)	73.4 (1.4)	**78.7** (2.0)	71.7 (2.2)	**74.0** (2.5)	**79.5** (4.8)	67.9 (1.6)	
+ task selection	**73.5** (1.6)	**75.8** (1.5)	77.4 (1.6)	**72.0** (1.6)	70.0 (1.6)	76.0 (6.8)	**69.8** (1.7)	

Table 18: **Results of few-shot learning with NLP benchmarks.** All results are obtained using RoBERTa-large. We report the mean (and standard deviation) of metrics over 5 different splits. †: Result in Gao et al. (2021a) in our paper; FT: finetuning; task selection: select multitask data from customized datasets.

[Gao et al.] Gao, Fisch, and Chen. Making pre-trained language models better few-shot learners. ACL'2020.

Future Work

- Does this multitask finetuning approach also work on multimodal tasks?
- Does our task selection algorithm apply?



CoCoOp

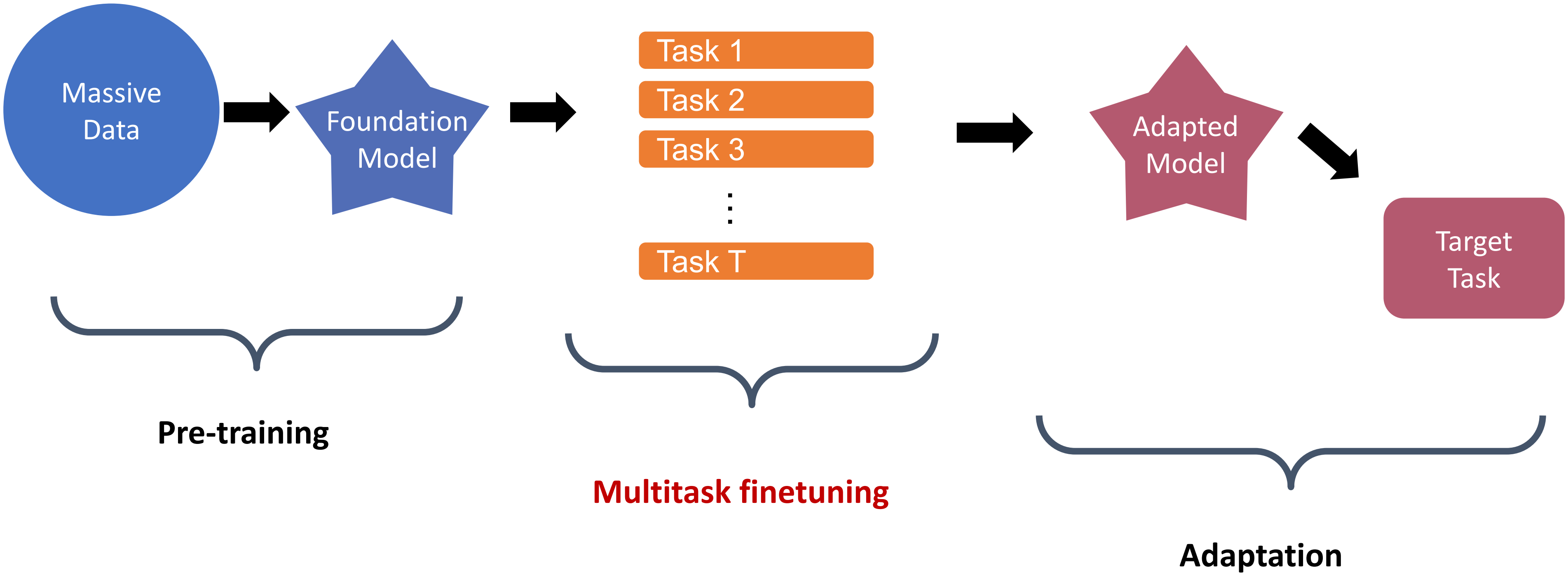
Figures from: *Conditional Prompt Learning for Vision-Language Models, 2022.*

Future Work

- Currently, generative models are a hot topic in research, attracting both theorists and practitioners. Does this framework apply to generative models as well?
    - Our theoretical framework mainly based on discriminative tasks. Can we derive similar conclusion for generative tasks? (In-context learning)

- Recent empirical achievements highlight the effectiveness of generative models in both natural language processing (e.g., GPT, Llama) and multimodal areas (e.g., Llava, GPT4-V). Is it possible to develop a task selection algorithm that better tailors these foundational models to a range of downstream tasks?

Take Home Message



Thanks!

Appendix

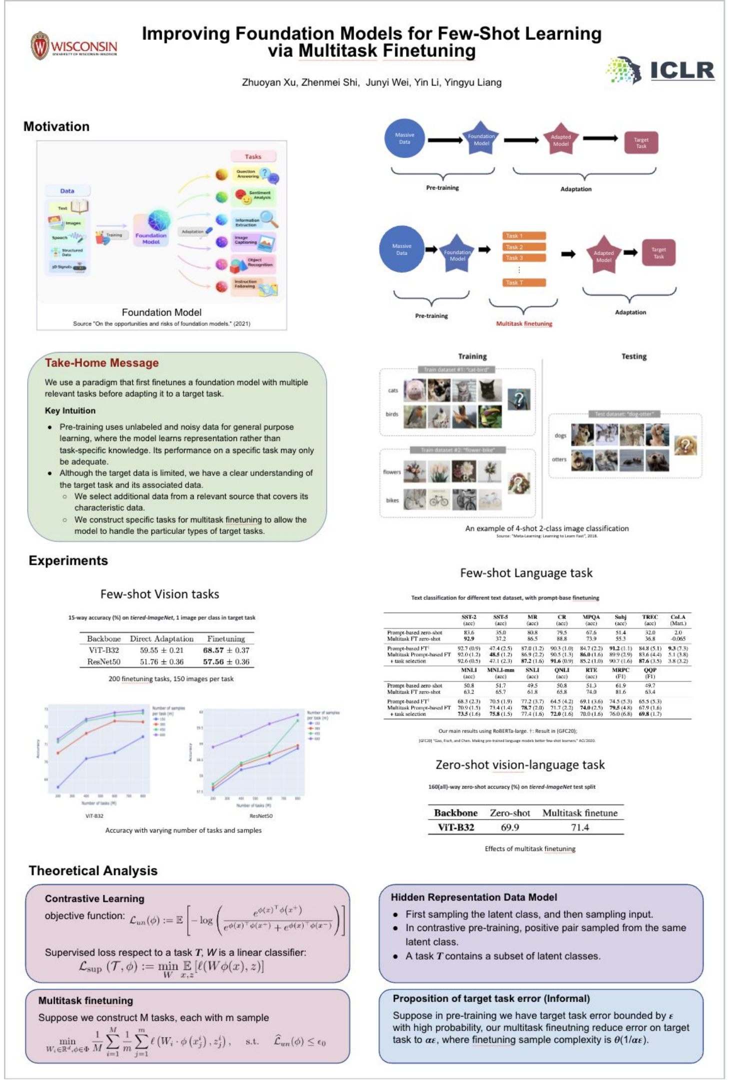
Our paper

Our slides

Appendix

Our Workshop Poster: [link]

Our Workshop Paper: [link]



Main Result

- Suppose target task is $\mathcal{T}_0$
- Let $\phi^* \in \Phi$ denote the model with the lowest target task loss $\mathcal{L}_{sup}(\mathcal{T}_0, \phi^*)$
- We want to bound $\mathcal{E}(\phi) = \mathcal{L}_{\text{sup}}(\mathcal{T}_0, \phi) - \mathcal{L}_{\text{sup}}(\mathcal{T}_0, \phi^*)$
- Pretraining loss as $\hat{\mathcal{L}}_{\text{pre}}(\hat{\phi})$

Theorem 1 (Contrastive pre-training loss (Informal))
Suppose in pre-training we have $\hat{\mathcal{L}}_{\text{pre}}(\hat{\phi}) \leq \epsilon_0$, and $\tau := \Pr_{(y_1,y_2)\sim\eta^2}\{y_1 = y_2\}$
then:

$$\mathcal{L}_{\text{sup}}\left(\mathcal{T}_0, \hat{\phi}\right) - \mathcal{L}_{\text{sup}}(\mathcal{T}_0, \phi^*) \leq \mathcal{O}\left(\frac{2\epsilon_0}{1-\tau}\right)$$

Main Result

- Suppose target task is $\mathcal{T}_0$
- We want to bound $\mathcal{L}_{\text{sup}}\left(\mathcal{T}_0, \phi\right)-\mathcal{L}_{\text{sup}}\left(\mathcal{T}_0, \phi^*\right)$

Theorem 2 (Multitask finetuning loss (Informal))

Suppose we solve multitask finetuning optimization with empirical loss smaller than $\epsilon_1=\frac{\alpha}{3} \frac{2 \epsilon_0}{1-\tau}$ and obtain ϕ'. If $\tilde{\epsilon}=\widehat{\mathcal{L}}_{pre}\left(\phi'\right)$:

$$M \geq \Omega\left(\frac{1}{\epsilon_1}\left[\mathcal{R}_M(\Phi(\tilde{\epsilon}))+\frac{1}{\epsilon_1} \log \left(\frac{1}{\delta}\right)\right]\right), \quad M m \geq \Omega\left(\frac{1}{\epsilon_1}\left[\mathcal{R}_{M m}(\Phi(\tilde{\epsilon}))+\frac{1}{\epsilon_1} \log \left(\frac{1}{\delta}\right)\right]\right)$$

Then with prob $1-\delta$,

$$\mathcal{L}_{\text{sup}}\left(\mathcal{T}_0, \phi'\right)-\mathcal{L}_{\text{sup}}\left(\mathcal{T}_0, \phi^*\right) \leq \mathcal{O}\left(\alpha \frac{2 \epsilon_0}{1-\tau}\right)$$

Experiments: zero-shot vision language task

160(all)-way zero-shot accuracy (%) on *tiered-ImageNet* test split

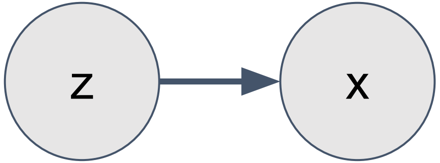
Backbone	Zero-shot	Multitask finetune
ViT-B32	69.9	71.4

Effects of multitask finetuning

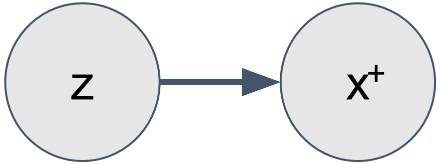
Problem Setup - Contrastive pre-training

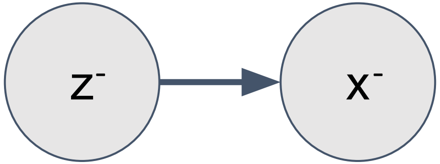
- $(z, z^-) \sim \eta^2, \; x, x^+ \sim \mathcal{D}(z), \; x^- \sim \mathcal{D}(z^-)$
- Contrastive loss:

$$\mathbb{E}\left[-\log\left(\frac{e^{\phi(x)^\top \phi(x^+)}}{e^{\phi(x)^\top \phi(x^+)} + e^{\phi(x)^\top \phi(x^-)}}\right)\right]$$



positive pair





negative pair

Data Model

Figures from: *Expanding Small-Scale Datasets with Guided Imagination, 2023*

Main Result

- Suppose target task is $\mathcal{T}_0$
- We want to bound $\mathcal{L}_{sup}(\mathcal{T}_0, \phi)$
- let ζ denote the conditional distribution of $(z_1, z_2) \sim \eta^2$ conditioned on $z_1 \neq z_2$

Definition 1 (Averaged representation difference)

$$\bar{d}_\zeta(\phi, \tilde{\phi}) := \mathop{\mathbb{E}}_{\mathcal{T} \sim \zeta}\left[\mathcal{L}_{sup}(\mathcal{T}, \phi) - \mathcal{L}_{sup}(\mathcal{T}, \tilde{\phi})\right] = \mathcal{L}_{sup}(\phi) - \mathcal{L}_{sup}(\tilde{\phi})$$

Definition 2 (worst-case representation difference)

$$d_{\mathcal{C}_0}(\phi, \tilde{\phi}) := \sup_{\mathcal{T}_0 \subseteq \mathcal{C}_0}\left[\mathcal{L}_{\text{sup}}\left(\mathcal{T}_0, \phi\right) - \mathcal{L}_{\text{sup}}\left(\mathcal{T}_0, \tilde{\phi}\right)\right]$$

(ν, ϵ)-diversity: For any $\phi, \tilde{\phi} \in \Phi$, $d_{\mathcal{C}_0}(\phi, \tilde{\phi}) \leq \bar{d}_\zeta(\phi, \tilde{\phi})/\nu + \epsilon$

Main Result

- Suppose target task is $\mathcal{T}_0$
- let ζ denote the conditional distribution of $(z_1, z_2) \sim \eta^2$ conditioned on $z_1 \neq z_2$
- (ν, ϵ) -diversity: For any $\phi, \tilde{\phi} \in \Phi, d_{\mathcal{C}_0}(\phi, \tilde{\phi}) \leq \bar{d}_\zeta(\phi, \tilde{\phi})/\nu + \epsilon$
- Suppose there is ϕ^* such that supervised loss are small across all tasks

Theorem 1 (Contrastive pre-training loss(baseline))
Suppose in pre-training we have $\hat{\mathcal{L}}_{un}(\hat{\phi}) \leq \epsilon_0$, then:

$$\mathcal{L}_{sup}(\mathcal{T}_0, \hat{\phi}) - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) \leq \frac{1}{\nu}\left[\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon$$

Main Result

- Suppose target task is $\mathcal{T}_0$
- let ζ denote the conditional distribution of $(z_1, z_2) \sim \eta^2$ conditioned on $z_1 \neq z_2$
- (ν, ϵ) -diversity: For any $\phi, \tilde{\phi} \in \Phi, d_{\mathcal{C}_0}(\phi, \tilde{\phi}) \leq \bar{d}_\zeta(\phi, \tilde{\phi})/\nu + \epsilon$

Theorem 2 (Multitask finetuning loss(Ours))

Suppose we solve multitask finetuning optimization with empirical loss smaller than $\epsilon_1 = \frac{\alpha}{3}\frac{1}{1-\tau}(2\epsilon_0 - \tau)$ and got ϕ'. If:

$$M \geq \Omega\left(\frac{1}{\epsilon_1}\left[\mathcal{R}_M\left(\Phi\left(\epsilon_0\right)\right) + \frac{1}{\epsilon_1}\log\left(\frac{1}{\delta}\right)\right]\right), \quad Mm \geq \Omega\left(\frac{1}{\epsilon_1}\left[\mathcal{R}_{Mm}\left(\Phi\left(\epsilon_0\right)\right) + \frac{1}{\epsilon_1}\log\left(\frac{1}{\delta}\right)\right]\right)$$

Then with prob $1 - \delta$,

$$\mathcal{L}_{sup}(\mathcal{T}_0, \phi') - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) \leq \frac{1}{\nu}\left[\alpha\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon$$

Remark

- Comparing to pre-training + adaptation(baseline), our multitask fineutning reduce error on target task by $\frac{1}{\nu}\left[(1-\alpha)\frac{1}{1-\tau}(2\epsilon_0-\tau)\right]$ where finetuning sample complexity is $\Theta\left(\frac{1}{\alpha\epsilon_0}\right)$

- Comparing to traditional supervised learning, self-supervised pre-training reduce error by $O\left(\frac{1}{Mm}\left[\mathcal{R}_{Mm}(\Phi)-\mathcal{R}_{Mm}\left(\Phi(\epsilon_0)\right)\right]\right)$

Experiments: Few-shot Vision tasks

5-way accuracy (%) on *mini-ImageNet*, 1/10/20 image per class in target task

ViT-B32

Accuracy with varying number shot images