Improving Foundation Models for Few-Shot Learning via Multitask Finetuning

Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Yin Li, Yingyu Liang
UW-Madison

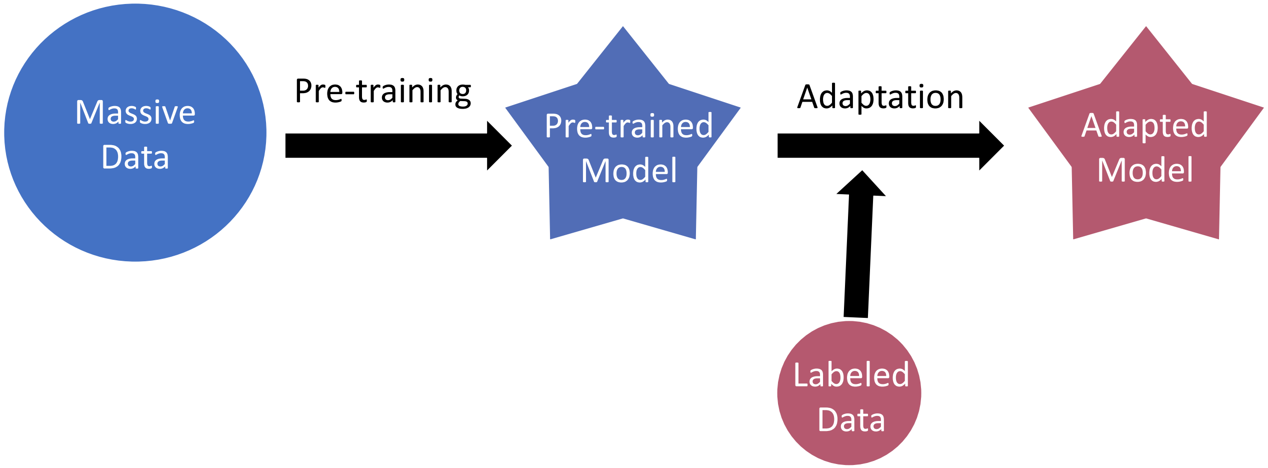
IFDS

New Paradigm: Pretraining + Adaptation

Paradigm shift: supervised learning ⟹ pre-training + adaptation

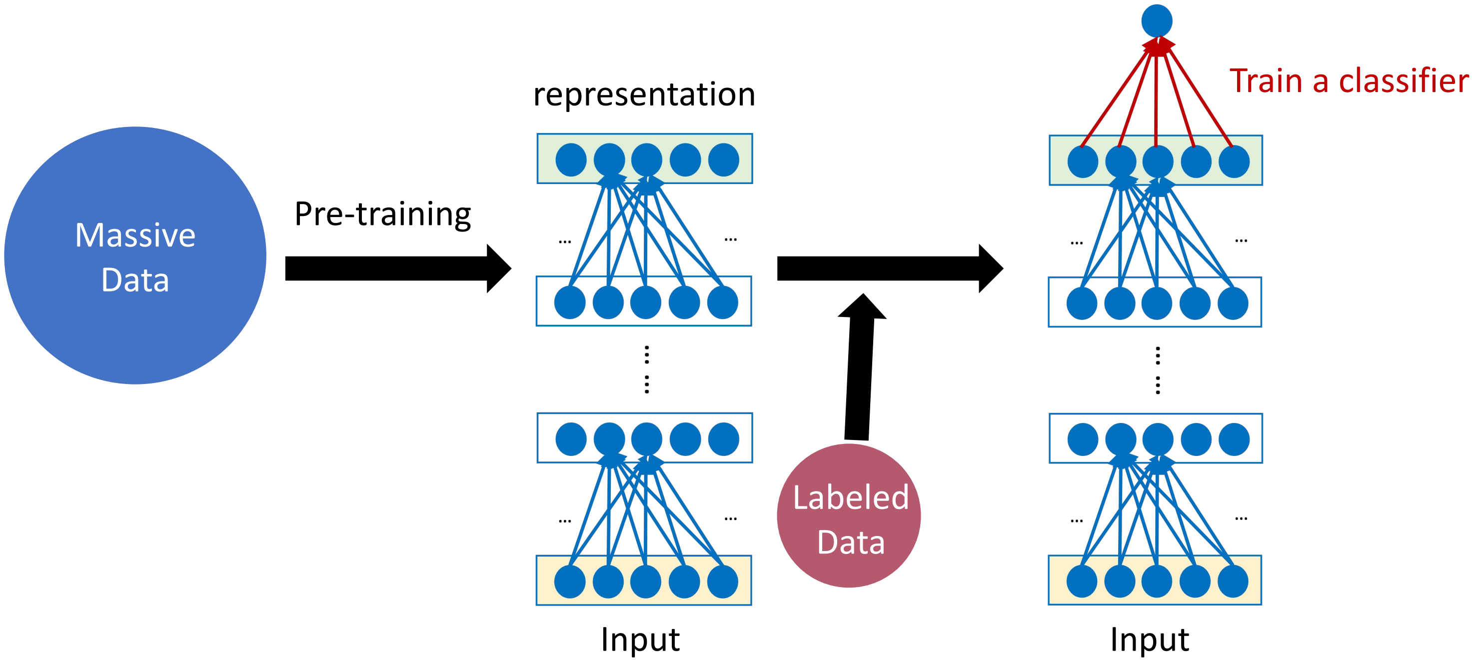
New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ⟹ pre-training + adaptation



New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ⟹ pre-training + adaptation



New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning ⟹ pre-training + adaptation

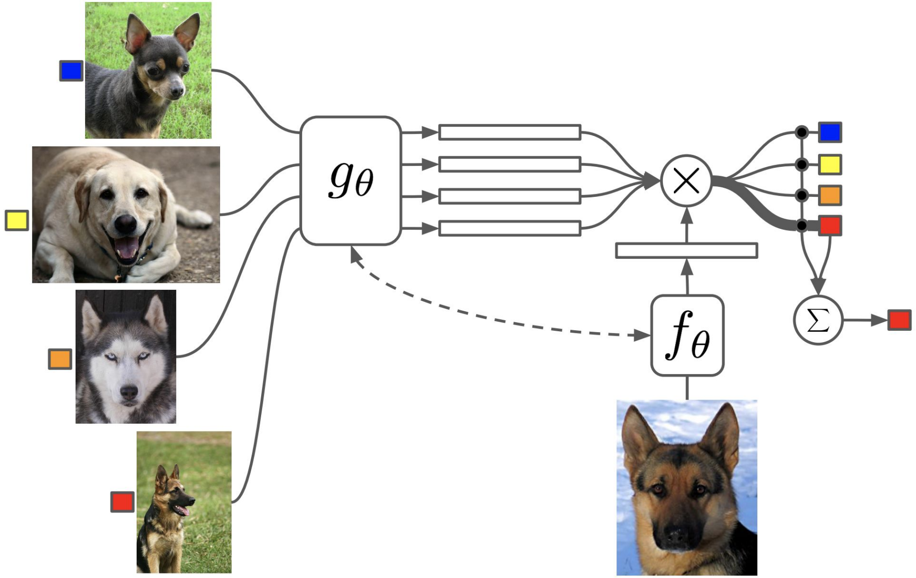


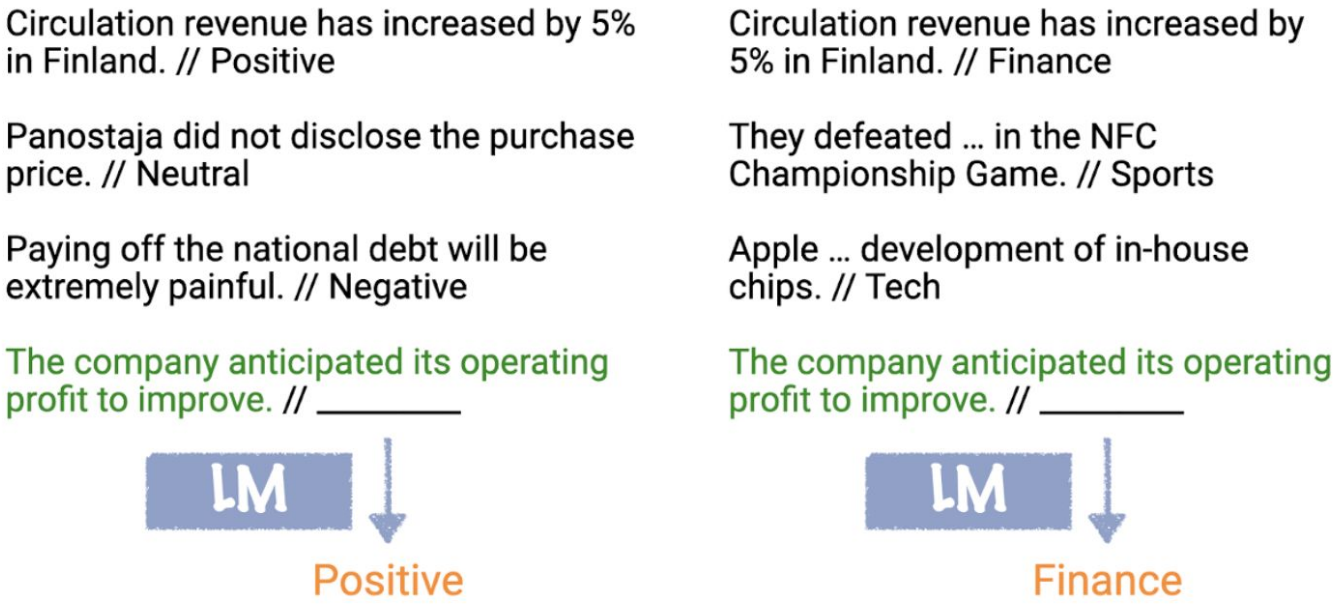
Figure 1: Matching Networks architecture

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

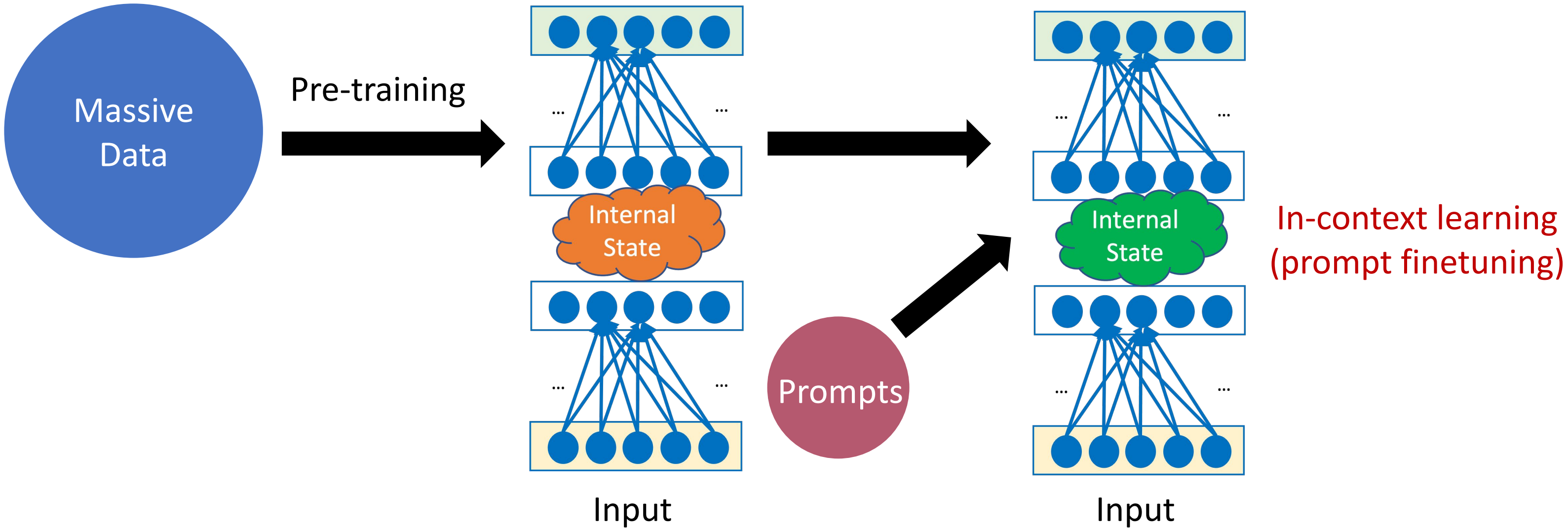


Adaptation of a pre-trained language decoder

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

New Paradigm: Pre-trained Representations

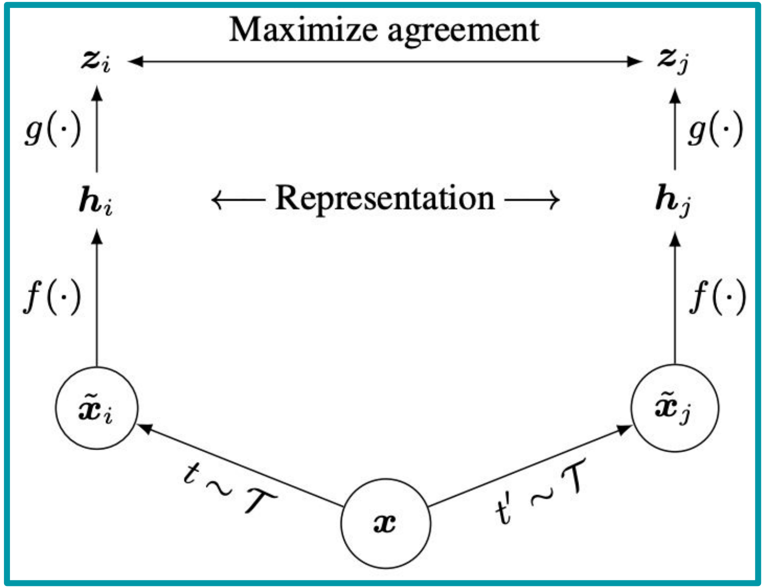
Paradigm shift: supervised learning ⟹ pre-training + adaptation



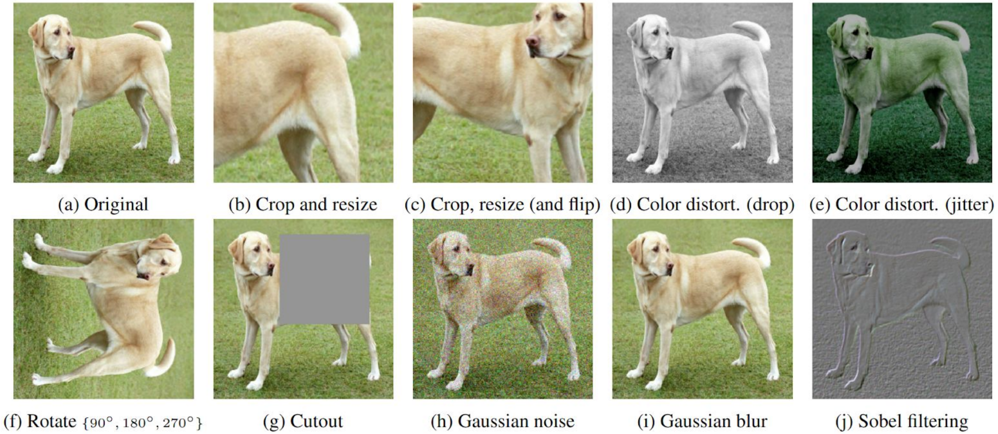
What does pre-training look like?

- Supervised learning
- Self-supervised learning:
  - Next sentence prediction (BERT)
  - Masked language prediction (BERT, RoBERTa)
  - Auto-regressive language modeling (GPT series)
  - 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)

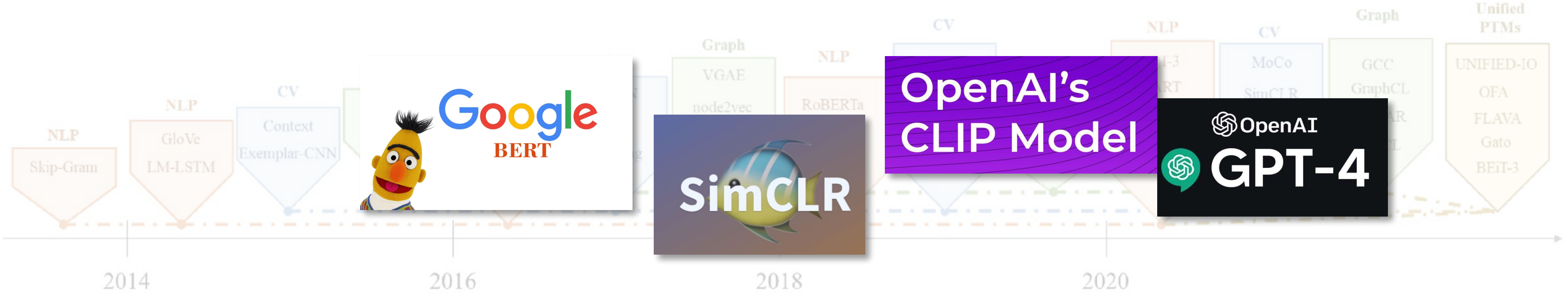
No need labels

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*

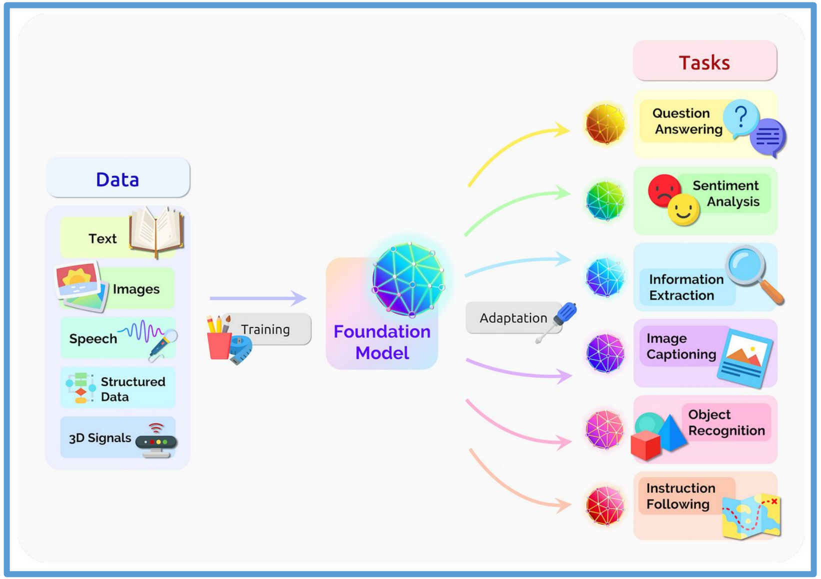
Intro - Foundation Model



The history and evolution of foundation models

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

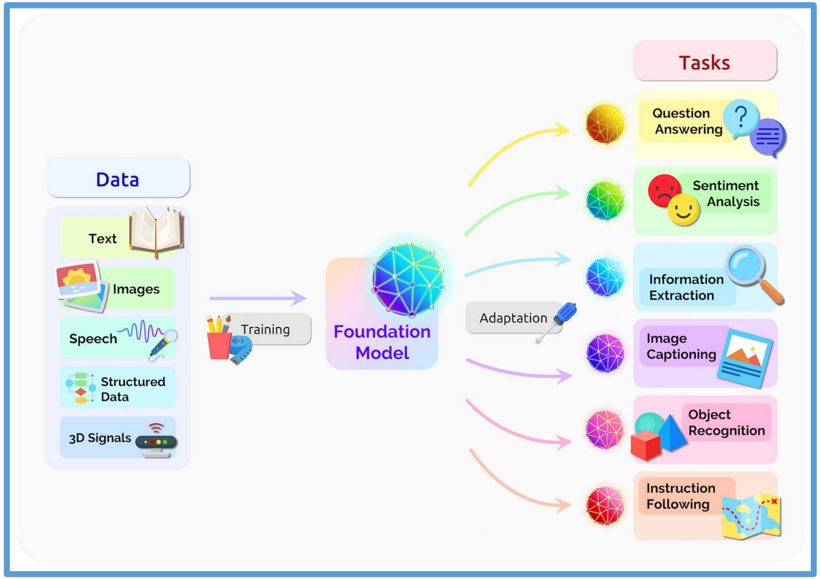
Intro - Foundation Model



Universality

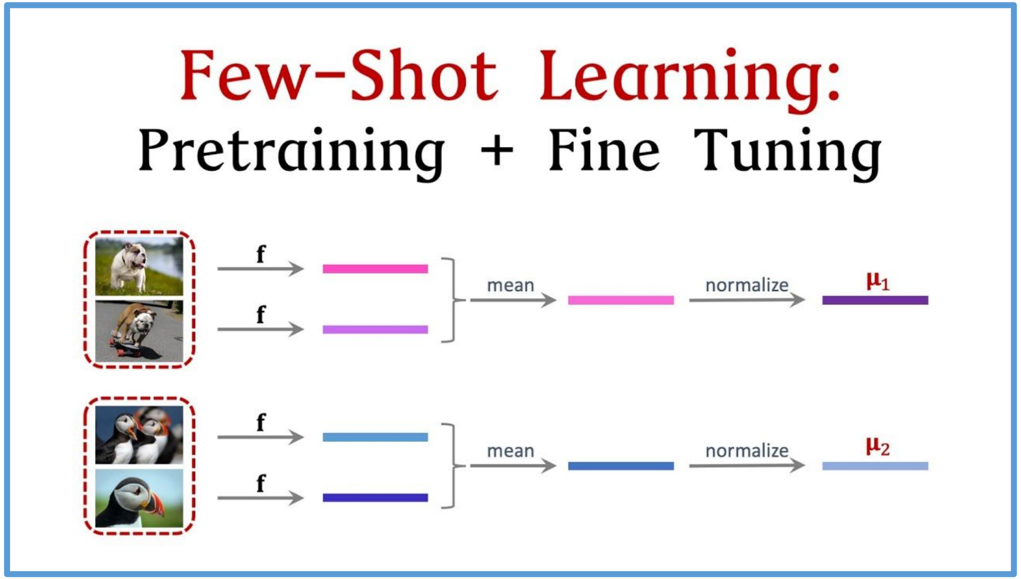
Figures from: *On the opportunities and risks of foundation models, 2021.*

Intro - Foundation Model



Universality

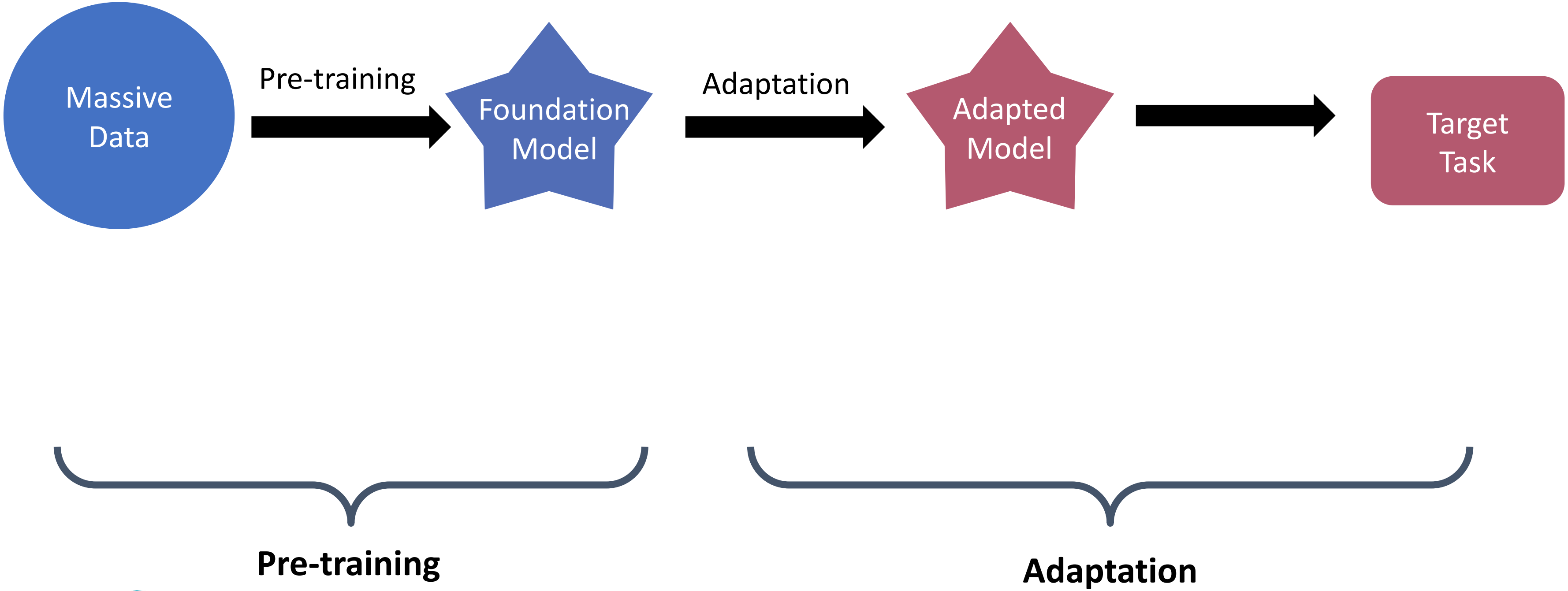
Figures from: *On the opportunities and risks of foundation models, 2021.*



Label Efficiency

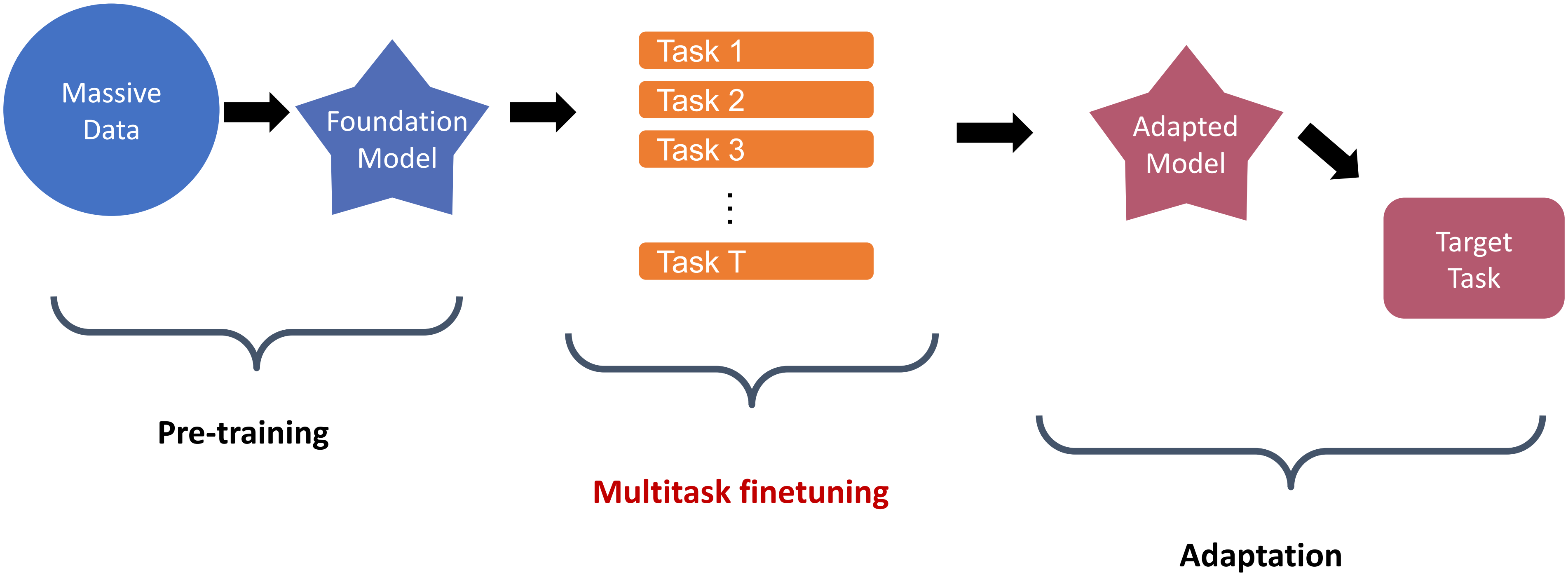
Figures from: https://www.youtube.com/watch?v=U6uFOIURcD0&ab_channel=ShusenWang, *2020*

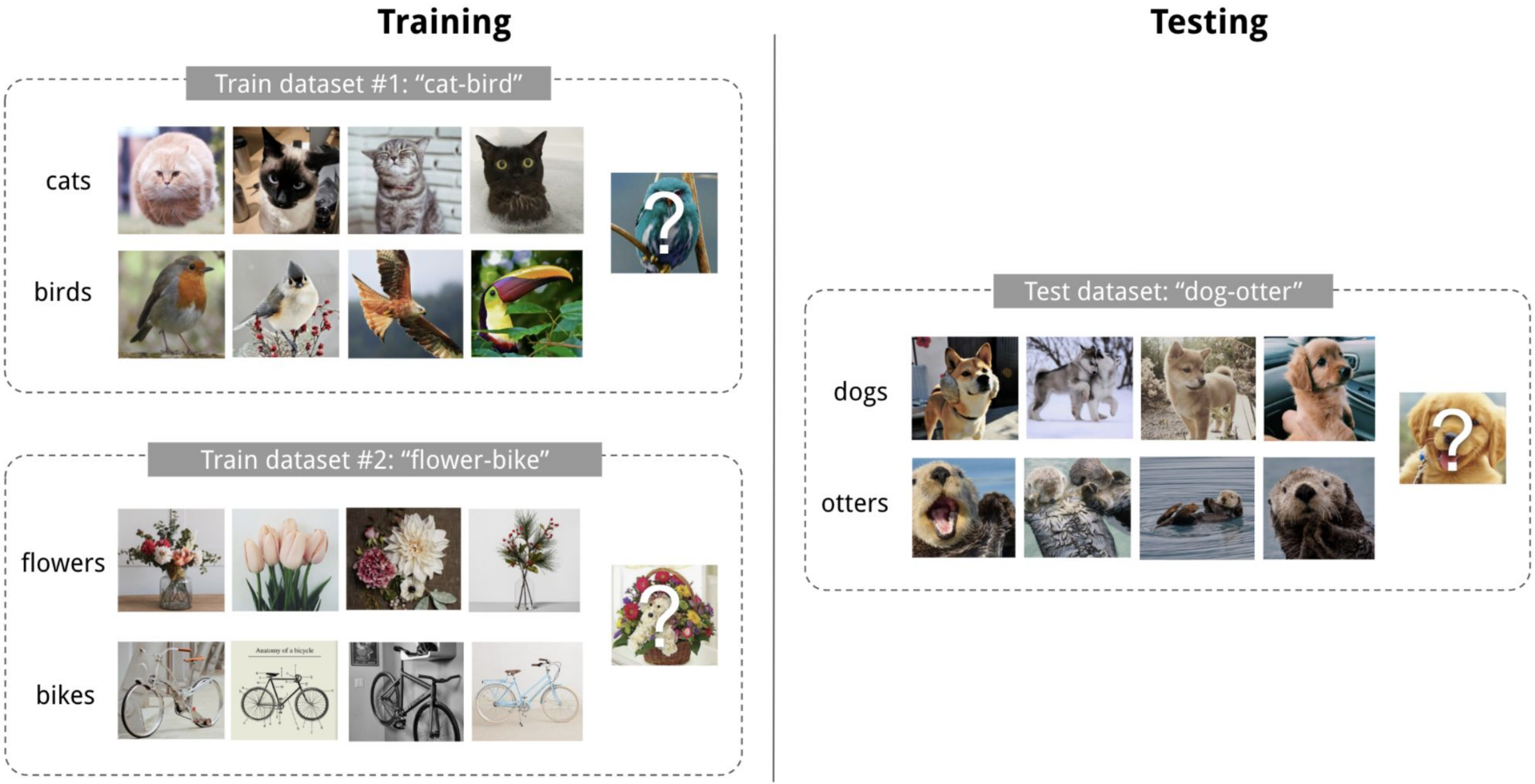
Paradigm: Pre-training + Adaptation



Q: Can we improve this?

Pre-training + Finetuning + Adaptation



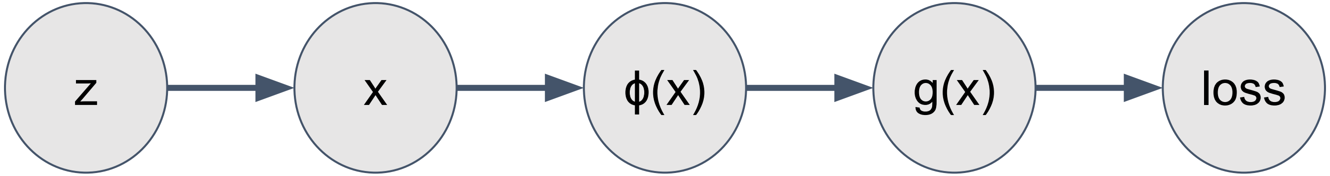


An example of 4-shot 2-class image classification

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

Problem Setup - Hidden representation data model

- Latent class $z \in \mathcal{C}$ over distribution $z \sim \eta$
- Task $\mathcal{T} = (z_1, \ldots, z_{K+1}) \subseteq \mathcal{C}$, instance $x \sim \mathcal{D}(z)$
- $\phi \in \Phi$ hypothesis class of representation functions, e.g, ResNet, ViT
- $g(x) = W\phi(x)$ as prediction logits of latent class



Dog

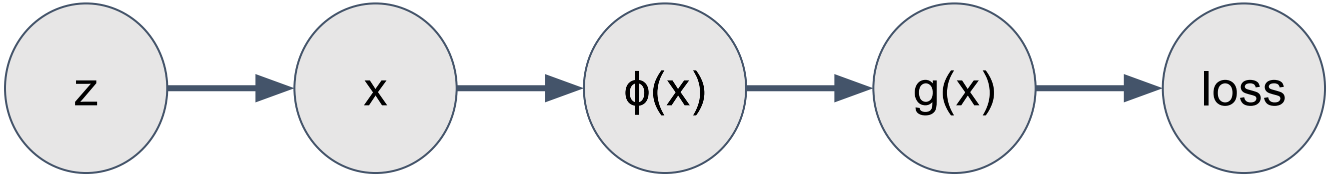
$$\begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_d \end{bmatrix} \qquad \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_{K+1} \end{bmatrix}$$

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

Problem Setup - Objective for a downstream task?

- Latent class $z \in \mathcal{C}$ over distribution $z \sim \eta$
- Task $\mathcal{T} = \{z_1, z_2\} \subseteq \mathcal{C}$, instance $x \sim \mathcal{D}(z)$
- $\phi \in \Phi$ hypothesis class of representation functions, e.g, ResNet, ViT
- $g(x) = W\phi(x)$ as prediction logits of latent class
- supervised loss w.r.t a task:

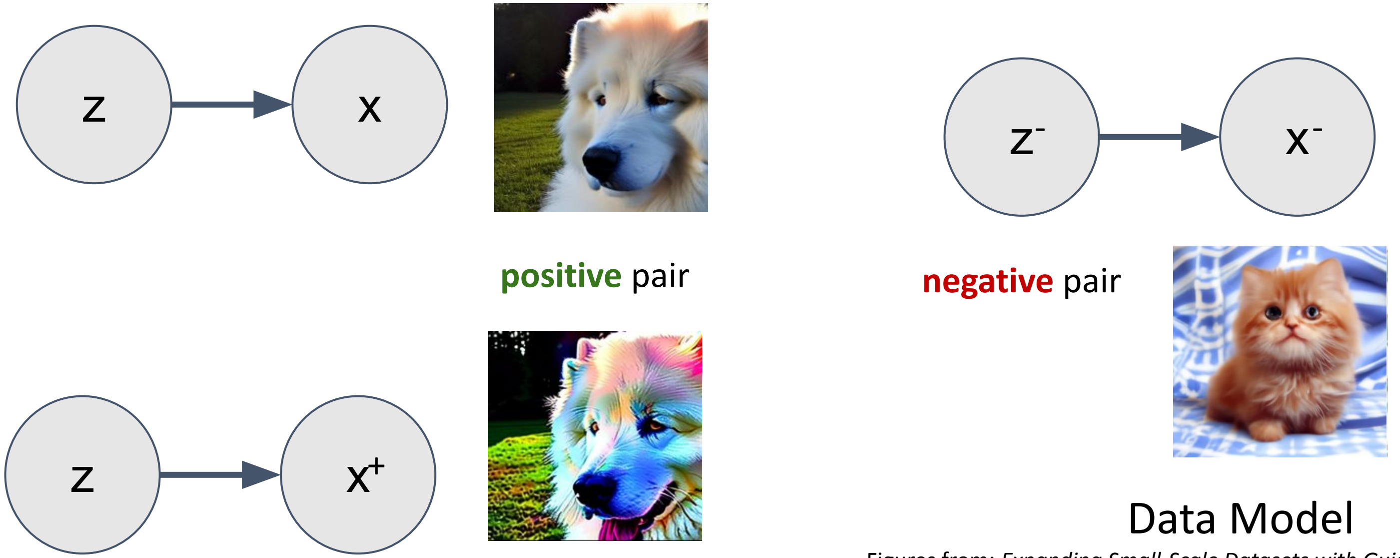
$$\mathcal{L}_{sup}(\mathcal{T}, \phi) := \min_{W} \mathop{\mathbb{E}}_{z \sim \mathcal{T}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}(z)} \left[\ell\left(W\phi(x), z\right)\right]$$



Problem Setup - Contrastive pre-training

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

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



Data Model

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

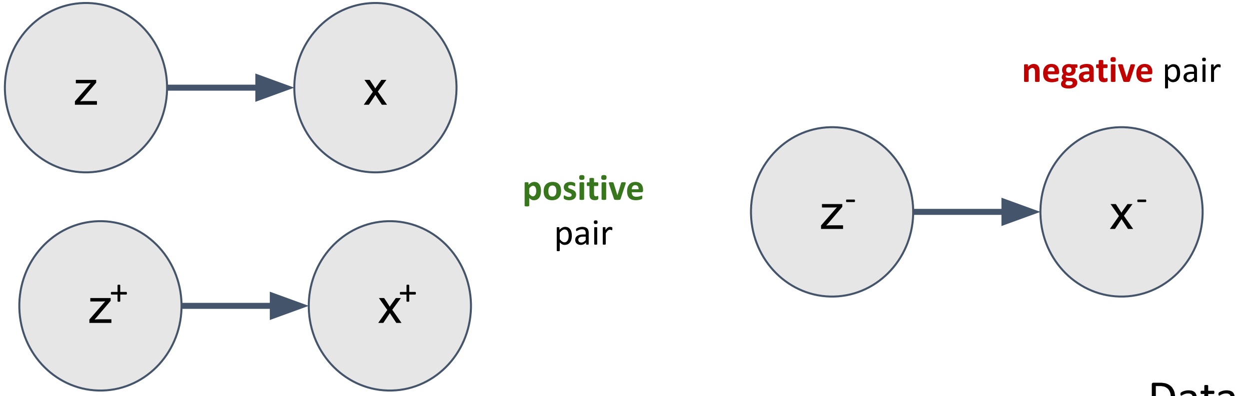
Problem Setup - Contrastive pre-training

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

$$\mathcal{L}_{un}(\phi) := \mathbb{E}\left[\ell_u\left(\phi(x)^\top\left(\phi(x^+) - \phi(x^-)\right)\right)\right]$$

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

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



Data Model

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

Problem Setup - Multitask Finetuning

- Suppose in pre-training we have $\widehat{\mathcal{L}}_{un}(\hat{\phi}) \leq \epsilon_0$
- Suppose we construct M tasks, each with m sample
- We further multitask finetune to get a new ϕ' by:

$$\min_{W_i \in \mathbb{R}^d, \phi \in \Phi} \quad \frac{1}{M} \sum_{i=1}^{M} \frac{1}{m} \sum_{j=1}^{m} \ell(W_i \cdot \phi(x_j^i), z_j^i), \quad \text{s.t.} \quad \widehat{\mathcal{L}}_{un}(\phi) \leq \epsilon_0$$

Intuition: Comparing to direct training, this reduce hypothesis space from Φ to $\Phi(\epsilon_0) = \left\{\phi \in \Phi : \hat{\mathcal{L}}_{un}(\phi) \leq \epsilon_0\right\}$

Main Result

- Suppose target task is $\mathcal{T}_0$
- Suppose there is ϕ^* such that supervised loss are small across all tasks
- We want to bound $\mathcal{L}_{sup}(\mathcal{T}_0, \phi) - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*)$

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

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

Main Result

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

Theorem 2 (Multitask finetuning loss(Ours))
Suppose we solve multitask finetuning optimization with empirical loss smaller than $\epsilon_1 = 2\alpha\epsilon_0$ and got ϕ'. If:

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

Then with prob $1 - \delta$,

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

Remark

- Comparing to pre-training + adaptation(baseline), our multitask fineutning reduce error on target task by $2(1-\alpha)\epsilon_0$
  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

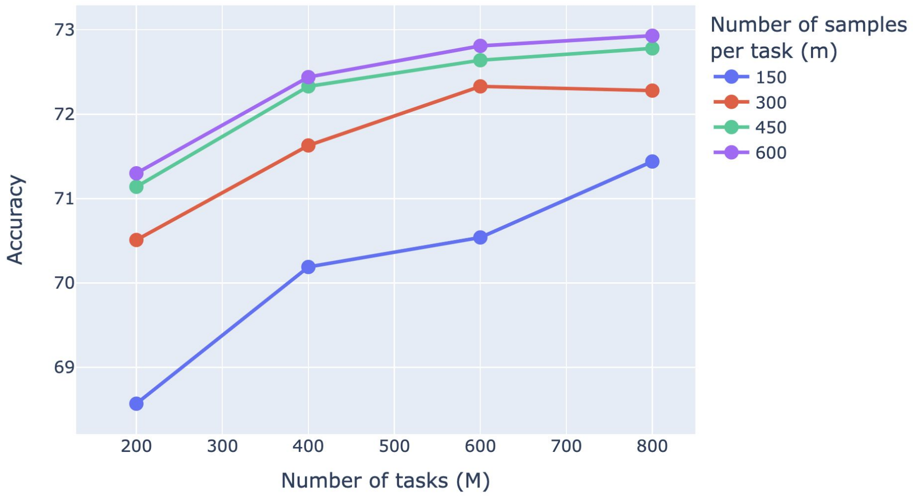
15-way accuracy (%) on *tiered-ImageNet*, 1 image per class in target task

Backbone	Direct Adaptation	Finetuning
ViT-B32	59.55 ± 0.21	$\mathbf{68.57} \pm 0.37$
ResNet50	51.76 ± 0.36	$\mathbf{57.56} \pm 0.36$

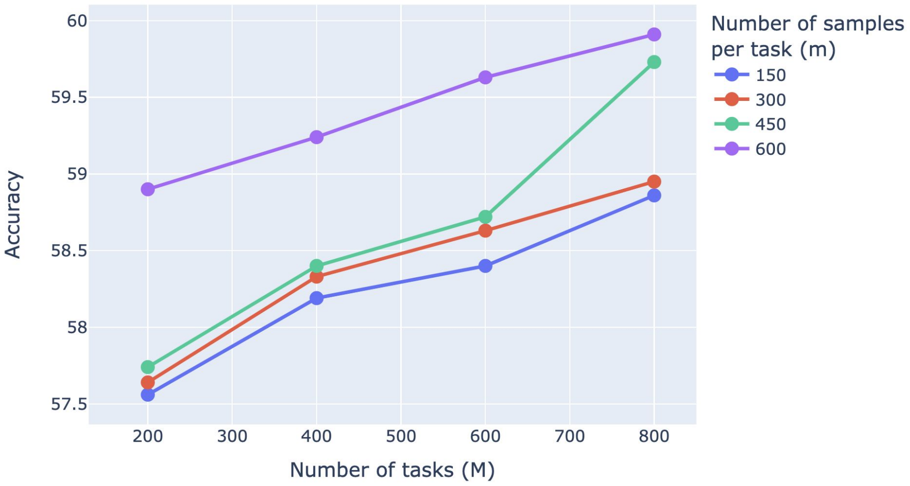
Effects of multitask finetuning

Experiments: Few-shot Vision tasks

15-way accuracy (%) on *tiered-ImageNet*, 1 image per class in target task



ViT-B32



ResNet50

Accuracy with varying number of tasks and samples

Experiments: Few-shot Language task

Text classification for different text dataset, with prompt-base finetuning

	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
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)

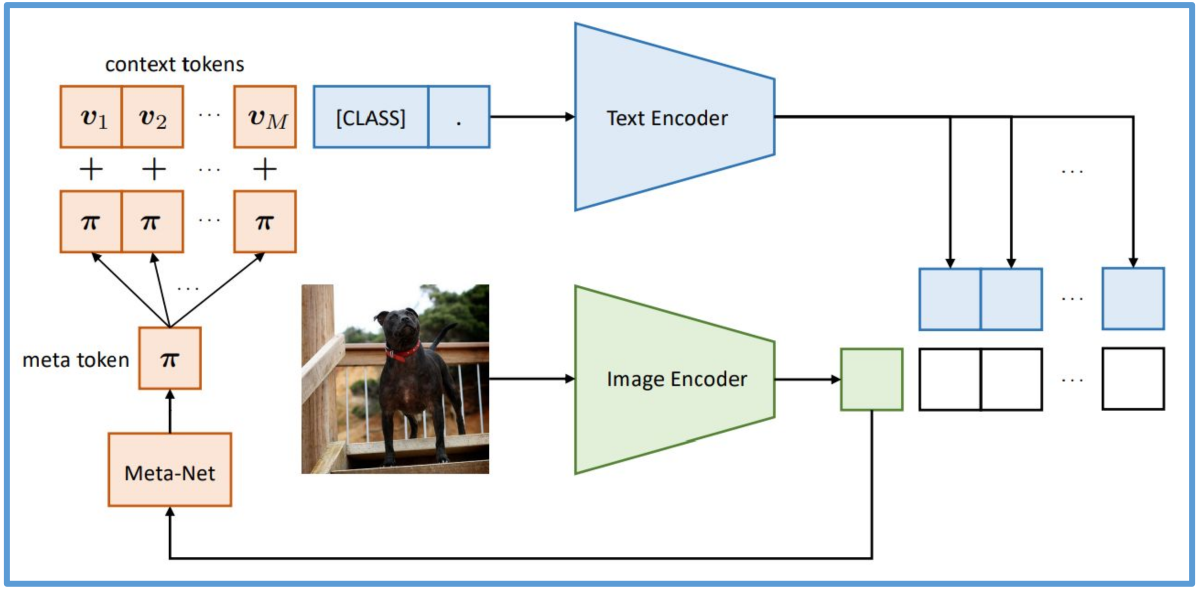
	MNLI (acc)	MNLI-mm (acc)	SNLI (acc)	QNLI (acc)	RTE (acc)	MRPC (F1)	QQP (F1)
Prompt-based zero-shot	50.8	51.7	49.5	50.8	51.3	61.9	49.7
Multitask FT zero-shot	63.2	65.7	61.8	65.8	74.0	81.6	63.4
Prompt-based FT†	68.3 (2.3)	70.5 (1.9)	77.2 (3.7)	64.5 (4.2)	69.1 (3.6)	74.5 (5.3)	65.5 (5.3)
Multitask Prompt-based 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)

Our main results using RoBERTa-large. †: Result in (GFC20);

[GFC20] Gao, Fisch, and Chen. Making pre-trained language models better few-shot learners. ACL’2020.

Experiments: zero-shot vision language task

Conditional context optimization for CLIP model



CoCoOp

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

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

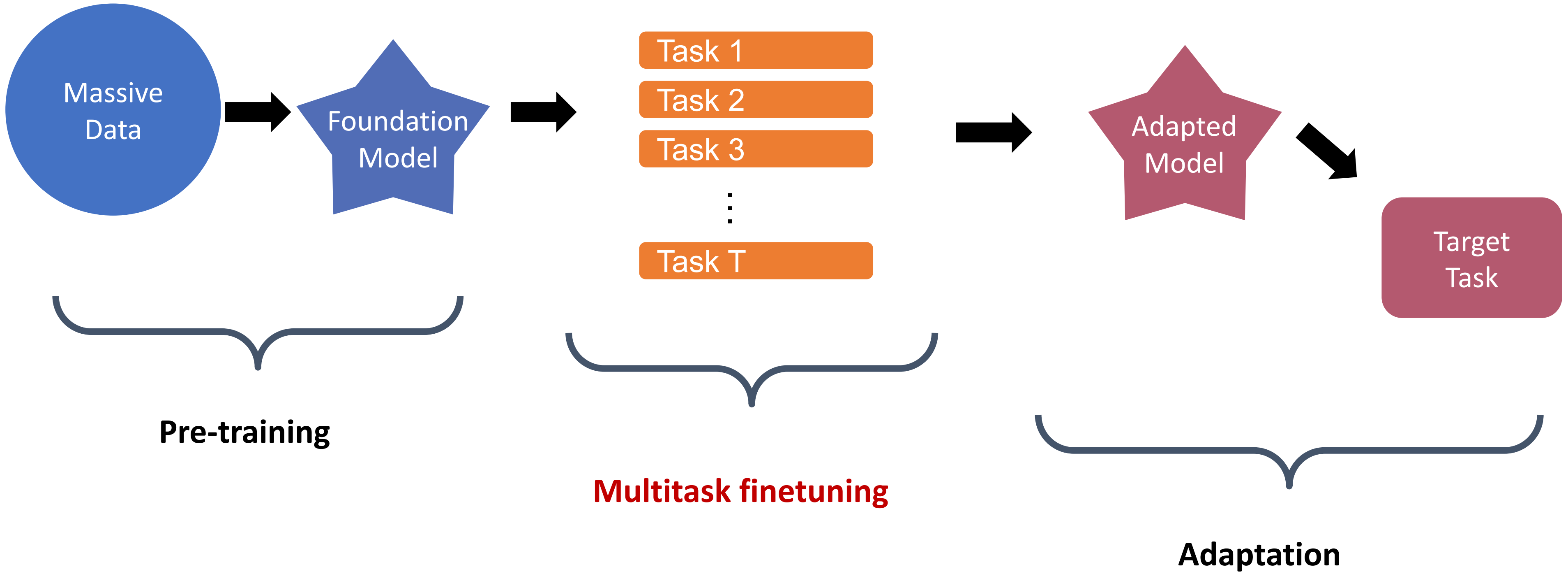
Effects of multitask finetuning

Future Work

- Theoretically: How would we quantify the relationship of data between multitask and target task? Concrete and well-motivated problem instances satisfying the task diversity assumptions for instantiating the error guarantee.

- Empirically: Does task diversity provide any insights on data selection in multitask finetuning? Can we design better strategies for constructing and choosing finetuning task?

Take Home Message

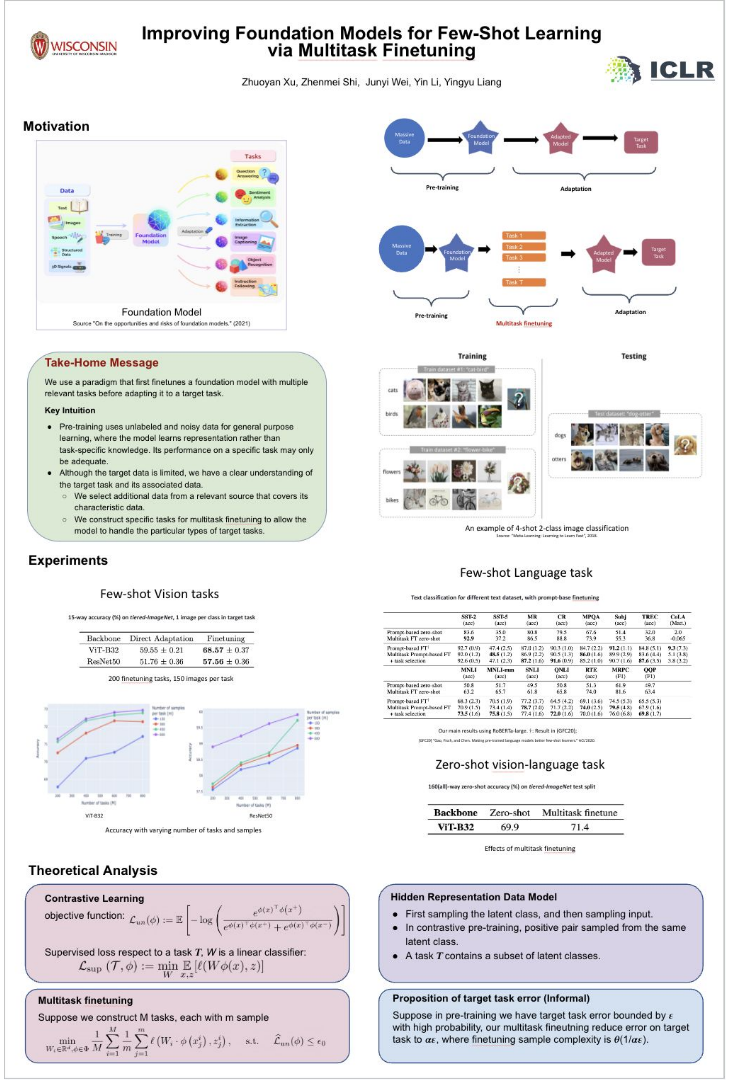


Thanks!

Appendix

Our Workshop Poster: [link]

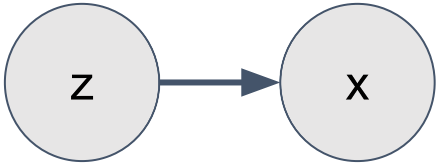
Our Workshop Paper: [link]



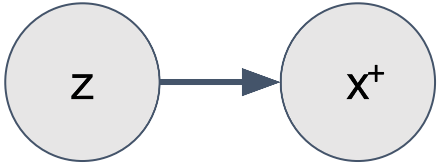
Problem Setup - Contrastive pre-training

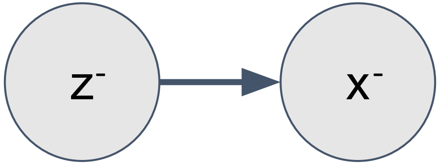
- $(z, z^-) \sim \eta^2, \; x, x^+ \sim \mathcal{D}(z), \; x^- \sim \mathcal{D}(z^-)$
- Contrastive loss:

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



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)

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

Definition 2 (worst-case representation difference)

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

(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