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 $\rightarrow$ pre-training + adaptation
New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning $\rightarrow$ pre-training + adaptation
New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning $\rightarrow$ pre-training + adaptation
New Paradigm: Pre-trained Representations

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

Circulation revenue has increased by 5% in Finland. // Positive
Panostaja did not disclose the purchase price. // Neutral
Paying off the national debt will be extremely painful. // Negative
The company anticipated its operating profit to improve. // ________

Circulation revenue has increased by 5% in Finland. // Finance
They defeated ... in the NFC Championship Game. // Sports
Apple ... development of in-house chips. // Tech
The company anticipated its operating profit to improve. // ________

Adaptation of a pre-trained language decoder

New Paradigm: Pre-trained Representations

Paradigm shift: supervised learning → pre-training + adaptation

[Diagram showing the process of pre-training and in-context learning with massive data and prompts]
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

SimCLR - (Image, Image)
No need labels

Image Data Augmentation

The history and evolution of foundation models

Intro - Foundation Model

Universality

Intro - Foundation Model

Universality

Label Efficiency
Figures from: [https://www.youtube.com/watch?v=U6uFOIURcD0&ab_channel=ShusenWang](https://www.youtube.com/watch?v=U6uFOIURcD0&ab_channel=ShusenWang), 2020
Paradigm: Pre-training + Adaptation

Massive Data → Pre-training → Foundation Model → Adaptation → Adapted Model → Target Task

Q: Can we improve this?
Pre-training + Finetuning + Adaptation

Massive Data → Foundation Model → Task 1, Task 2, Task 3, ..., Task T → Adapted Model

Pre-training → Multitask finetuning → Adaptation

Target Task
An example of 4-shot 2-class image classification

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

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

$$\mathcal{L}_{sup}(\mathcal{T}, \phi) := \min_{W} \mathbb{E}_{z \sim \mathcal{T}} \mathbb{E}_{x \sim \mathcal{D}(z)} [\ell (W\phi(x), z)]$$
Problem Setup - Contrastive pre-training

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

- Contrastive loss:

\[
\mathbb{E} \left[ - \log \left( \frac{e^{\phi(x)^\top \phi(x^+)} \eta^2}{e^{\phi(x)^\top \phi(x^+)} + e^{\phi(x)^\top \phi(x^-)}} \right) \right]
\]

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

- In particular:
  \[
  \ell_u(v) = \log(1 + \exp(-v)) \text{ will recover the loss in previous slide}
  \]

Data Model

Figures from: Expanding Small-Scale Datasets with Guided Imagination, 2023
Problem Setup - Multitask Finetuning

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

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

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

- Suppose target task is $\mathcal{T}_0$
- Suppose there is $\phi^*$ 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:

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

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

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 $\phi'$. 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 $1 - \delta$,

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

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

  where fine-tuning sample complexity is $\Theta\left(\frac{1}{\alpha\epsilon_0}\right)$

- Comparing to traditional supervised learning, self-supervised pre-training reduces error by $O\left(\frac{1}{Mm} [R_{Mm}(\Phi) - R_{Mm}(\Phi(\epsilon_0))]\right)$
Experiments: Few-shot Vision tasks

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

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Direct Adaptation</th>
<th>Finetuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B32</td>
<td>59.55 ± 0.21</td>
<td>68.57 ± 0.37</td>
</tr>
<tr>
<td>ResNet50</td>
<td>51.76 ± 0.36</td>
<td>57.56 ± 0.36</td>
</tr>
</tbody>
</table>

Effects of multitask finetuning
Experiments: Few-shot Vision tasks

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

Accuracy with varying number of tasks and samples

ViT-B32

ResNet50
Experiments: Few-shot Language task

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

<table>
<thead>
<tr>
<th></th>
<th>SST-2 (acc)</th>
<th>SST-5 (acc)</th>
<th>MR (acc)</th>
<th>CR (acc)</th>
<th>MPQA (acc)</th>
<th>Subj (acc)</th>
<th>TREC (acc)</th>
<th>CoLA (Matt.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompt-based zero-shot</td>
<td>83.6</td>
<td>35.0</td>
<td>80.8</td>
<td>79.5</td>
<td>67.6</td>
<td>51.4</td>
<td>32.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Multitask FT zero-shot</td>
<td><strong>92.9</strong></td>
<td>37.2</td>
<td>86.5</td>
<td>88.8</td>
<td>73.9</td>
<td>55.3</td>
<td>36.8</td>
<td>-0.065</td>
</tr>
<tr>
<td>Prompt-based FT†</td>
<td>92.7 (0.9)</td>
<td>47.4 (2.5)</td>
<td>87.0 (1.2)</td>
<td>90.3 (1.0)</td>
<td>84.7 (2.2)</td>
<td><strong>91.2</strong> (1.1)</td>
<td>84.8 (5.1)</td>
<td><strong>9.3</strong> (7.3)</td>
</tr>
<tr>
<td>Multitask Prompt-based FT + task selection</td>
<td>92.0 (1.2)</td>
<td><strong>48.5</strong> (1.2)</td>
<td>86.9 (2.2)</td>
<td>90.5 (1.3)</td>
<td><strong>86.0</strong> (1.6)</td>
<td>89.9 (2.9)</td>
<td>83.6 (4.4)</td>
<td>5.1 (3.8)</td>
</tr>
<tr>
<td></td>
<td>MNLI (acc)</td>
<td>MNLI-mm (acc)</td>
<td>SNLI (acc)</td>
<td>QNLI (acc)</td>
<td>RTE (acc)</td>
<td>MRPC (F1)</td>
<td>QQP (F1)</td>
<td></td>
</tr>
<tr>
<td>Prompt-based zero-shot</td>
<td>50.8</td>
<td>51.7</td>
<td>49.5</td>
<td>50.8</td>
<td>51.3</td>
<td>61.9</td>
<td>49.7</td>
<td></td>
</tr>
<tr>
<td>Multitask FT zero-shot</td>
<td>63.2</td>
<td>65.7</td>
<td>61.8</td>
<td>65.8</td>
<td>74.0</td>
<td>81.6</td>
<td>63.4</td>
<td></td>
</tr>
<tr>
<td>Prompt-based FT†</td>
<td>68.3 (2.3)</td>
<td>70.5 (1.9)</td>
<td>77.2 (3.7)</td>
<td>64.5 (4.2)</td>
<td>69.1 (3.6)</td>
<td>74.5 (5.3)</td>
<td>65.5 (5.3)</td>
<td></td>
</tr>
<tr>
<td>Multitask Prompt-based FT + task selection</td>
<td>70.9 (1.5)</td>
<td>73.4 (1.4)</td>
<td><strong>78.7</strong> (2.0)</td>
<td>71.7 (2.2)</td>
<td><strong>74.0</strong> (2.5)</td>
<td><strong>79.5</strong> (4.8)</td>
<td>67.9 (1.6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>73.5 (1.6)</td>
<td><strong>75.8</strong> (1.5)</td>
<td>77.4 (1.6)</td>
<td><strong>72.0</strong> (1.6)</td>
<td>70.0 (1.6)</td>
<td>76.0 (6.8)</td>
<td><strong>69.8</strong> (1.7)</td>
<td></td>
</tr>
</tbody>
</table>

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

Experiments: zero-shot vision language task

Conditional context optimization for CLIP model

CoCoOp

Experiments: zero-shot vision language task

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

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Zero-shot</th>
<th>Multitask finetune</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B32</td>
<td>69.9</td>
<td>71.4</td>
</tr>
</tbody>
</table>

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

Massive Data → Foundation Model → Pre-training → Multitask finetuning → Adapted Model → Adaptation → Target Task

Task 1 → Task 2 → Task 3 → … → Task T

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:

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

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 $\mathcal{L}_{\text{sup}}(\mathcal{T}_0, \phi)$
- Let $\zeta$ denote the conditional distribution of $(z_1, z_2) \sim \eta^2$ conditioned on $z_1 \neq z_2$

Definition 1 (Averaged representation difference)
$$\bar{d}_\zeta(\phi, \tilde{\phi}) := \mathbb{E}_{\mathcal{T} \sim \zeta} \left[ \mathcal{L}_{\text{sup}}(\mathcal{T}, \phi) - \mathcal{L}_{\text{sup}}(\mathcal{T}, \tilde{\phi}) \right] = \mathcal{L}_{\text{sup}}(\phi) - \mathcal{L}_{\text{sup}}(\tilde{\phi})$$

Definition 2 (worst-case representation difference)
$$d_{C_0}(\phi, \tilde{\phi}) := \sup_{\mathcal{T}_0 \subseteq C_0} \left[ \mathcal{L}_{\text{sup}}(\mathcal{T}_0, \phi) - \mathcal{L}_{\text{sup}}(\mathcal{T}_0, \tilde{\phi}) \right]$$

$(\nu, \epsilon)$-diversity: For any $\phi, \tilde{\phi} \in \Phi$, $d_{C_0}(\phi, \tilde{\phi}) \leq \bar{d}_\zeta(\phi, \tilde{\phi})/\nu + \epsilon$
Main Result

- Suppose target task is $\mathcal{T}_0$
- Let $\zeta$ denote the conditional distribution of $(z_1, z_2) \sim \eta^2$ conditioned on $z_1 \neq z_2$
- $(\nu, \epsilon)$-diversity: For any $\phi, \tilde{\phi} \in \Phi$, $d_{C_0}(\phi, \tilde{\phi}) \leq d_\zeta(\phi, \tilde{\phi}) / \nu + \epsilon$
- Suppose there is $\phi^*$ such that supervised loss are small across all tasks

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

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

- Suppose target task is $\mathcal{T}_0$
- let $\zeta$ denote the conditional distribution of $(z_1, z_2) \sim \eta^2$ conditioned on $z_1 \neq z_2$
- $(\nu, \epsilon)$-diversity: For any $\phi, \tilde{\phi} \in \Phi$, $d_{C_0}(\phi, \tilde{\phi}) \leq \tilde{d}_{\zeta}(\phi, \tilde{\phi})/\nu + \epsilon$

**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 $\phi'$. If:

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

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 fine-tuning reduce error on target task by
  \[ \frac{1}{\nu} \left[ (1 - \alpha) \frac{1}{1 - \tau} (2\epsilon_0 - \tau) \right] \]
  where fine-tuning sample complexity is \( \Theta \left( \frac{1}{\alpha \epsilon_0} \right) \)

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