IMPROVING FOUNDATION MODELS FOR FEW-SHOT LEARNING VIA MULTITASK FINETUNING

Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Yin Li, Yingyu Liang
University of Wisconsin-Madison
{zhuoyan.xu,jwei53,yin.li}@wisc.edu,{zhmeishi,yliang}@cs.wisc.edu

ABSTRACT

Foundation models have become essential tools for AI. In this paper, we study the problem of adapting foundation models, pre-trained using contrastive learning, to downstream tasks with limited labels. We explore the paradigm of finetuning a foundation model before adapting to a target task, using a set of related tasks with a few labeled samples. We show both theoretically and empirically that with a diverse set of related tasks this finetuning leads to reduced error in the target task, when compared with directly adapting the same pre-trained model, e.g., at least 6% target accuracy improvements on the miniImageNet.

1 INTRODUCTION

Foundation models, pre-trained on broad data, promise to adapt to a wide range of downstream tasks. These models celebrate recent success in both vision and language, and have emerged as an essential tool in AI, with examples including BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), CLIP (Radford et al., 2021), and DALL-E 2 (Ramesh et al., 2022). A foundation model is first pre-trained on large-scale data to learn a representation function, often with self-supervised learning (e.g., contrastive learning), and then finetuned to adapt to downstream tasks with potential novel classes. This paradigm is particularly helpful for tasks with limited labels, as only a simple function (e.g., linear) needs to be learned on top of the representation from the foundation model. Despite the tremendous success, effective adaptation of foundation models, especially to tasks with limited labels, remains an open practical question (Wortsman et al., 2022) that lacks theoretical insight.

Inspired by meta learning (Finn et al., 2017), we explore a paradigm that further finetunes a foundation model with multiple relevant tasks, before adapting the pre-trained model to a target task. The crux of this approach lies in a multitask finetuning stage, during which the model is trained using supervised learning on a set of tasks related to the target task, obtained from a handy source. Each of these tasks might have a small number of labeled samples, and the classes of these samples might not overlap with those in the target task. Our key intuition is that a sufficiently diverse set of relevant tasks can capture similar latent characteristics as the target task, thereby producing better representation and reducing errors in the target task.

We study the effect of this approach theoretically and empirically. In particular, we present a theoretical framework for analyzing pre-training followed by multitask finetuning. Our analysis shows that with limited but diverse labeled data, finetuning can improve the prediction performance on the downstream task. Empirically, we perform controlled experiments in finetuning while varying the task size and sample size. Our results suggest that finetuning increases the prediction accuracy compared to direct adaptation with a small amount of data. Further, our results show that with sufficient number of tasks, increasing the sample size per task does not provide notable improvement. While these results are still preliminary, we consider our analysis and experiments point towards a promising direction for effective adaptation of foundation models.

Related Work. Contrastive learning is one of the most effective self-supervised learning techniques recently, in both language and vision pre-training tasks (Oord et al., 2018; Chen et al., 2020; He et al., 2020; Tian et al., 2020a; Reed et al., 2022). Zoph et al. (2020) study the empirical effect of supervised pre-training and self-supervised training on pre-training representations. Arora et al. (2019) establish theoretical guarantees on downstream classification performance. HaoChen et al.

(2021) provide analysis on spectral contrastive loss. Their analysis assumes the pre-training and target tasks share the same data distribution. Multitask supervised learning has been used to obtain the representation models for downstream target tasks. A line of theoretical work provides the error bound of the target task in terms of sample complexity (Du et al., 2020; Tripuraneni et al., 2021; Shi et al., 2023). Tripuraneni et al. (2020) establish a unified framework of multitask learning and proposed the notion of task diversity for the training data. Their work mainly analyzes the supervisedly pre-trained representations by multitasks. Our work focuses on self-supervised pre-trained representations and proposes to use multitasks for finetuning the pre-trained model. Our approach and analysis guarantee that limited but diverse finetuning data can improve the prediction performance on the target task with novel classes. Multitasks have also been used to finetune supervisedly pre-trained representations, and it is observed that direct training with limited data may lead to overfitting so few-shot learning techniques like meta-learning are often used (Wang et al., 2020; Tian et al., 2020b; Chen et al., 2021; Yang et al., 2022).

2 Theoretical Analysis

Contrastive Learning. Let $\mathcal{X}$ denote the input space (e.g., images) and $\overline{\mathcal{Z}} \subseteq \mathbb{R}^d$ the output space of the foundation model. Let Φ denote the hypothesis class of foundation models $\phi : \mathcal{X} \mapsto \overline{\mathcal{Z}}$. Contrastive learning pre-trains a foundation model via contrastive loss: First sample a point x and then apply some transformation (e.g., flipping, cropping) to obtain x^+; independently sample another point x^-. The population contrastive loss is then

$$\mathcal{L}_{un}(\phi) := \mathbb{E}\left[\ell_u\left(\phi(x)^\top\left(\phi(x^+) - \phi(x^-)\right)\right)\right], \tag{1}$$

where the loss function ℓ_u is non-negative decreasing function. In particular, logistic loss $\ell_u(v) = \log\left(1 + \exp\left(-v\right)\right)$ recovers the typical contrastive loss in related work (Logeswaran & Lee, 2018; Oord et al., 2018; Chen et al., 2020). For training set $\mathcal{S}_{un} := \left\{x_i, x_i^+, x_i^-\right\}_{i=1}^N$ with N samples, the empirical contrastive loss is $\widehat{\mathcal{L}}_{un}(\phi) := \frac{1}{N}\sum_{i=1}^N \left[\ell_u\left(\phi(x_i)^\top\left(\phi(x_i^+) - \phi(x_i^-)\right)\right)\right]$.

For theoretical analysis, we follow the setup of Arora et al. (2019). Assume the data are generated from a set of latent classes $\mathcal{C}$. There is a distribution η over the classes, and each class $z \in \mathcal{C}$ has a distribution $\mathcal{D}(z)$ over inputs x. Then to generate the contrastive data (x, x^+, x^-), sample $(z, z^-) \sim \eta^2$ and $x, x^+ \sim \mathcal{D}(z),\ x^- \sim \mathcal{D}(z^-)$. Let $\mathcal{D}_{\mathrm{con}}(\eta)$ denote this distribution of (x, x^+, x^-).

Downstream Prediction Tasks. A pre-trained foundation model ϕ can be used for downstream prediction tasks by learning linear classifiers on ϕ. Consider binary classification $\mathcal{T} = \{z_1, z_2\}$ where z_1, z_2 are two classes. (The general multiclass setting is in Appendix B.) A linear classifier on ϕ is given by $g(x) = W\phi(x)$ where $W \in \mathbb{R}^{2\times d}$, and the supervised loss on data point (x, z) is

$$\ell(g(x), z) := \ell_u\left((g(x))_z - (g(x))_{z'\neq z}\right). \tag{2}$$

The data in task $\mathcal{T}$ is by uniformly drawing $z \in \{z_1, z_2\}$ (denote as $z \sim \mathcal{T}$) and then drawing $x \sim \mathcal{D}(z)$. The supervised loss of ϕ w.r.t the task $\mathcal{T}$ is then

$$\mathcal{L}_{sup}(\mathcal{T}, \phi) := \min_W \mathop{\mathbb{E}}_{z\sim\mathcal{T}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}(z)} \left[\ell\left(W\phi(x), z\right)\right]. \tag{3}$$

In few-shot learning on novel classes, there are limited labeled data points for learning the linear classifier, and furthermore, the target task $\mathcal{T}_0$ can contain classes different from those in pre-training (i.e., $\mathcal{T}_0 \subseteq \mathcal{C}_0$ where the label classes $\mathcal{C}_0$ may or may not overlap with the pre-training latent classes $\mathcal{C}$). We are interested in obtaining a model ϕ such that $\mathcal{L}_{sup}(\mathcal{T}_0, \phi)$ is small.

Multitask Finetuning. To improve the performance on the target task $\mathcal{T}_0$, we explore using multitask finetuning on a pre-trained model $\hat{\phi}$. Suppose we have M tasks, and each task contains m labeled sample. Let $\mathcal{S} := \{(x_j^i, z_j^i) : i \in [M], j \in [m]\}$ denote the finetuning data. Suppose the pre-training ensures $\widehat{\mathcal{L}}_{un}(\hat{\phi}) \leq \epsilon_0$, we further finetune to get a new model ϕ' by

$$\min_{W_i\in\mathbb{R}^d, \phi\in\Phi} \quad \frac{1}{M}\sum_{i=1}^M \frac{1}{m}\sum_{j=1}^m \ell(W_i \cdot \phi(x_j^i), z_j^i), \quad \text{s.t.} \quad \widehat{\mathcal{L}}_{un}(\phi) \leq \epsilon_0. \tag{4}$$

Main Results. We are interested in comparing the performance of $\hat{\phi}$ (the model from pre-training) and ϕ' (the model from pre-training + multitask finetuning) on a target task $\mathcal{T}_0$, i.e., comparing $\mathcal{L}_{sup}(\mathcal{T}_0, \hat{\phi})$ and $\mathcal{L}_{sup}(\mathcal{T}_0, \phi')$. For the analysis, we assume there is a model ϕ^* that allows low average supervised loss on all tasks $\mathcal{T}$. More precisely, let ζ denote the conditional distribution of $(z_1, z_2) \sim \eta^2$ conditioned on $z_1 \neq z_2$, and define the average supervised loss of a model ϕ as

$$\mathcal{L}_{sup}(\phi) = \mathop{\mathbb{E}}_{\mathcal{T}\sim\zeta}\left[\mathcal{L}_{sup}(\mathcal{T}, \phi)\right]. \tag{5}$$

Suppose $\mathcal{L}_{sup}(\phi^*)$ and $\mathcal{L}_{sup}(\mathcal{T}_0, \phi^*)$ are small. We will also need some mild regularity assumptions:

(R) $\|\phi\|_2 \leq R$ and linear operator $\|W\|_2 \leq B$. We assume loss ℓ_u are bounded by $[0, C]$ and $\ell_u(\cdot)$ is L-Lipschitz and supervised loss $\mathcal{L}_{sup}(\mathcal{T}, \phi)$ is $\tilde{L}$-Lipschitz with respect to ϕ.

Finally, we use a slight generalization of the task diversity notion from Tripuraneni et al. (2020).

Definition 1. *The **averaged representation difference** for two model $\phi, \tilde{\phi}$ on a distribution ζ over tasks is $\bar{d}_\zeta(\phi, \tilde{\phi}) := \mathop{\mathbb{E}}_{\mathcal{T}\sim\zeta}\left[\mathcal{L}_{sup}(\mathcal{T}, \phi) - \mathcal{L}_{sup}(\mathcal{T}, \tilde{\phi})\right] = \mathcal{L}_{sup}(\phi) - \mathcal{L}_{sup}(\tilde{\phi})$. The **worst-case representation difference** between representations $\phi, \tilde{\phi}$ on the family of classes $\mathcal{C}_0$ is $d_{\mathcal{C}_0}(\phi, \tilde{\phi}) := \sup_{\mathcal{T}_0\subseteq\mathcal{C}_0}\left[\mathcal{L}_{sup}(\mathcal{T}_0, \phi) - \mathcal{L}_{sup}(\mathcal{T}_0, \tilde{\phi})\right]$. We say the model class Φ has (ν, ϵ)-diversity for ζ and $\mathcal{C}_0$ if for any $\phi, \tilde{\phi} \in \Phi$, $d_{\mathcal{C}_0}(\phi, \tilde{\phi}) \leq \bar{d}_\zeta(\phi, \tilde{\phi})/\nu + \epsilon$.*

If we only perform pre-training without multitask finetuning, then we have

Theorem 1. *Assume Assumption (**R**) and that Φ has (ν, ϵ)-diversity for ζ and $\mathcal{C}_0$. Suppose $\hat{\phi}$ satisfies $\hat{\mathcal{L}}_{un}(\hat{\phi}) \leq \epsilon_0$. Let $\tau := \Pr_{(z_1,z_2)\sim\eta^2}\{z_1 = z_2\}$. Then for any target task $\mathcal{T}_0 \subset \mathcal{C}_0$,*

$$\mathcal{L}_{sup}(\mathcal{T}_0, \hat{\phi}) - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) \leq \frac{1}{\nu}\left[\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon. \tag{6}$$

We now consider pre-training followed by multitask finetuning. Define the subset of models with contrastive loss smaller than ϵ_0 as $\Phi(\epsilon_0) := \left\{\phi \in \Phi : \widehat{\mathcal{L}}_{un}(\phi) \leq \epsilon_0\right\}$. Recall the Rademacher complexity of Φ on n points is $\mathcal{R}_n(\Phi) := \mathop{\mathbb{E}}_{\{\sigma_i\}_{j=1}^n, \{x_j\}_{j=1}^n}\left[\sup_{\phi\in\Phi}\sum_{j=1}^n \sigma_j \phi(x_j)\right]$.

Theorem 2. *Assume Assumption (**R**) and that Φ has (ν, ϵ)-diversity for ζ and $\mathcal{C}_0$. Suppose for some small constant $\alpha \in (0, 1)$, we solve (4) with empirical loss lower than $\epsilon_1 = \frac{\alpha}{3}\frac{1}{1-\tau}(2\epsilon_0 - \tau)$ and obtain ϕ'. For any $\delta > 0$, if*

$$M \geq \frac{1}{\epsilon_1}\left[4\sqrt{2}\tilde{L}\mathcal{R}_M(\Phi(\epsilon_0)) + \frac{4C^2}{\epsilon_1}\log(\frac{2}{\delta})\right], Mm \geq \frac{1}{\epsilon_1}\left[16LB\mathcal{R}_{Mm}(\Phi(\epsilon_0)) + \frac{4C^2}{\epsilon_1}\log(\frac{2}{\delta})\right],$$

then with probability $1 - \delta$, for any target task $\mathcal{T}_0 \subseteq \mathcal{C}_0$,

$$\mathcal{L}_{sup}(\mathcal{T}_0, \phi') - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) \leq \frac{1}{\nu}\left[\alpha\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon. \tag{7}$$

With sufficient samples in finetuning, we can get a ϕ' such that the prediction performance $\mathcal{L}_{sup}(\phi')$ is significantly better than that of pre-training alone $\mathcal{L}_{sup}(\hat{\phi})$. The task sample complexity is $O(\frac{\mathcal{R}_M(\Phi(\epsilon_0))}{\epsilon_1} + \frac{\log(1/\delta)}{\epsilon_1^2})$. The first term is the Rademacher complexity of reduced representation space $\Phi(\epsilon_0)$ with tasks number M. The second term relates to the generalization bound. The total labeled sample complexity is $O(\frac{\mathcal{R}_{Mm}(\Phi(\epsilon_0))}{\epsilon_1} + \frac{\log(1/\delta)}{\epsilon_1^2})$. Theorem 2 shows finetuning will reduce target task error bound from $\frac{1}{\nu}\left[\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \epsilon^*\right] + \epsilon$ to $\frac{1}{\nu}\left[\alpha\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \epsilon^*\right] + \epsilon$, resulting in $\frac{1}{\nu}\left[(1-\alpha)\frac{1}{1-\tau}(2\epsilon_0 - \tau)\right]$ reduction, where labeled sample complexity is proportional to $\frac{1}{\alpha\epsilon_0}$. Note that multitask training has been used to train representations directly (Tripuraneni et al., 2020; Du et al., 2020) (instead of finetuning representations pre-trained via self-supervised learning). Their

analysis shows training needs $O\left(\frac{\mathcal{R}_{Mm}(\Phi)}{\epsilon_1}+\frac{\log(1/\delta)}{\epsilon_1^2}\right)$ to get the same error. Thus, pre-training narrows down the hypothesis search space, and the sample complexity of labeled data is reduced by $O\left(\frac{1}{\epsilon_1}\left[\mathcal{R}_{Mm}(\Phi)-\mathcal{R}_{Mm}\left(\Phi(\epsilon_0)\right)\right]\right)$ by pre-training. Equivalently, the error bound is reduced by $O\left(\frac{1}{Mm}\left[\mathcal{R}_{Mm}(\Phi)-\mathcal{R}_{Mm}\left(\Phi(\epsilon_0)\right)\right]\right)$. See full proof in Appendix A.2.

3 Experiments and Results

We experiment with CLIP models (Radford et al., 2021) as an exemplary foundation model. The pre-trained ViT-B/32 and ResNet50 were used as image encoders. We perform evaluation and finetuning in form of few-shot tasks. A target task typically contains N classes with K support samples and Q query samples in each class. The goal is to classify query samples into the N classes based on the support samples. We use the nearest-centroid method on top of representations for evaluation. During finetuning, the image encoder is directly optimized on few-shot classification tasks. We conduct experiments on a widely used few-shot image recognition benchmark: miniImageNet (Vinyals et al., 2016). Table 1 compares the performance of the CLIP model between direct adaption to target task and multitask finetuning with limited data. There are 200 finetuning tasks, each containing 50 images. After only 10 epochs, finetuning improves the average accuracy by 6% and 2.8% for ViT-B32 and ResNet50 respectively. We also provide results for vision language models and language models. More experimental details and results can be found in Appendix C.

Backbone	Direct Adaptation	Finetuning
ViT-B32	83.03 ± 0.24	**89.07** ± 0.20
ResNet50	78.36 ± 0.25	**81.19** ± 0.25

Table 1: Effects of multitask finetuning.

Task (M) vs Sample (m). We vary the task size and sample size per task during finetuning. We verify the trend of different numbers of tasks and numbers of images per task. Each task contains 5 classes. For finetuning tasks, $m = 40$ indicates each class contains the 1-shot image and 7-query images. m = 200 indicates each class contains 5-shot and 35-query images. $m = 1000$ indicates each class contains 25-shot and 175-query images. $M = m = 0$ indicates direct evaluation without finetuning. For target tasks, each class contains the 1-shot image and 15 query images.

Task (M) \ Sample (m)	0	40	200	1000
0	83.03 ± 0.24			
200		88.53 ± 0.22	89.50 ± 0.20	**89.93** $\pm$ **0.20**
1000		89.37 ± 0.20	**90.81** $\pm$ **0.19**	90.97 ± 0.19
5000		**89.95** $\pm$ **0.20**	90.94 ± 0.19	91.16 ± 0.18

Table 2: Accuracy with varying number of tasks and samples (ViT-B32 backbone).

Table 2 shows the results on the pre-trained CLIP model using ViT backbone. For direct adaptation without finetuning, the model achieves 83.03% accuracy. Multitask finetuning improves the average accuracy at least by 5.5%. For a fixed number of tasks or samples per task, increasing samples or tasks improves the accuracy. These results suggest that the total number of samples ($M \times m$) will determine the overall performance, supporting our main theorem.

4 Conclusions

In this work, we considered using multitask finetuning to adapt pre-trained foundation models to downstream tasks with limited labels. Our key contribution lies in the theoretical guarantees that finetuning using a diverse set of relevant tasks can improve the performance on the target task, in comparison to direct adaptation. Our analysis was confirmed empirically by our preliminary results. Admittedly, our work is at an early stage. Yet we consider that our analysis and results provide useful insight to the critical problem of effective adaptation of foundation models. We will further develop our work, and explore directions such as (1) concrete and well-motivated problem instances satisfying the task diversity assumptions for instantiating the error guarantees; and (2) better finetuing methods and strategies for constructing and choosing finetuning tasks.

ACKNOWLEDGMENTS

The work is partially supported by Air Force Grant FA9550-18-1-0166, the National Science Foundation (NSF) Grants 2008559-IIS, CCF-2046710, and 2023239-DMS.

REFERENCES

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *36th International Conference on Machine Learning, ICML 2019*, pp. 9904–9923. International Machine Learning Society (IMLS), 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9062–9071, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

Siddhant Garg and Yingyu Liang. Functional regularization for representation learning: A unified theoretical perspective. *Advances in Neural Information Processing Systems*, 33:17187–17199, 2020.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2584–2594, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha. The trade-off between universality and label efficiency of representations from contrastive learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rvsbw2YthH_.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020a.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pp. 266–282. Springer, 2020b.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852–7862, 2020.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022.

Zhanyuan Yang, Jinghua Wang, and Yingying Zhu. Few-shot classification with contrastive learning. In *European Conference on Computer Vision*, pp. 293–309. Springer, 2022.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33: 3833–3845, 2020.

Appendix

A DEFERRED PROOFS

In this section, we provide a formal setting and proof. We first formalize our setting in K classes:

Contrastive Learning. In contrastive learning, we sampled one image x from any latent class z, then apply data augmentation module that randomly transforms such image into another view of the original example denoted x^+. We also sample other K images $\{x_k^-\}_{k=1}^K$ from other latent classes $\{z_k^-\}_{k=1}^K$. We treat (x, x^+) as a positive pair and (x, x_k^-) as negative pairs. We define $\mathcal{D}_{\text{con}}(\eta)$ over sample $(x, x^+, x_1^-, \ldots, x_K^-)$ by following sampling procedure

$$(z, z_1^-, \ldots, z_K^-) \sim \eta^{K+1} \tag{8}$$

$$x \sim \mathcal{D}(z),\ x^+ \sim \mathcal{D}(z),\ x_k^- \sim \mathcal{D}(z_k^-), k = 1, \ldots, K. \tag{9}$$

We consider general contrastive loss $\ell_u\left(\left\{\phi(x)^\top\left(\phi(x^+) - \phi(x_k^-)\right)\right\}_{k=1}^K\right)$, where loss function ℓ_u is non-negative decreasing function. Minimizing the loss is equivalent to maximizing the similarity between positive pairs while minimizing it between negative pairs. In particular, logistic loss $\ell_u(\boldsymbol{v}) = \log\left(1 + \sum_i \exp\left(-\boldsymbol{v}_i\right)\right)$ for $\boldsymbol{v} \in \mathbb{R}^K$ recovers the one used in most empirical works: $-\log\left(\frac{\exp\left\{\phi(x)^\top\phi(x^+)\right\}}{\exp\{\phi(x)^\top\phi(x^+)\} + \sum_{i=1}^K \exp\left\{\phi(x)^\top\phi(x_i^-)\right\}}\right)$. The population contrastive loss is defined as $\mathcal{L}_{un}(\phi) := \mathbb{E}\left[\ell_u\left(\left\{\phi(x)^\top\left(\phi(x^+) - \phi(x_k^-)\right)\right\}_{k=1}^K\right)\right]$. Let $\mathcal{S}_{un} := \left\{x_j, x_j^+, x_{j1}^-, \ldots, x_{jK}^-\right\}_{j=1}^N$ denote our contrastive training set with N samples, sampled from $\mathcal{D}_{\text{con}}(\eta)$, we have empirical contrastive loss $\widehat{\mathcal{L}}_{un}(\phi) := \frac{1}{N}\sum_{i=1}^N\left[\ell_u\left(\left\{\phi(x)^\top\left(\phi(x^+) - \phi(x_k^-)\right)\right\}_{k=1}^K\right)\right]$.

Supervised Tasks. Given a representation function ϕ, we apply a task-specific linear transformation W to the representation to obtain the final prediction. Consider $(K+1)$-way supervised task $\mathcal{T}$ consist a set of distinct classes $(z_1, \ldots, z_{K+1}) \subseteq \mathcal{C}$. We define $\mathcal{D}_{\mathcal{T}}(z)$ as the distribution of randomly drawing $z \in (z_1, \ldots, z_{K+1})$, we denote this process as $z \sim \mathcal{T}$. Let $\mathcal{S}_{\mathcal{T}} := \{x_j, z_j\}_{j=1}^m$ denote our labeled training set with m samples, sampled i.i.d. from $z_j \sim \mathcal{T}$ and $x_j \sim \mathcal{D}(z_j)$. Define $g(\phi(\mathbf{x})) := W\phi(x) \in \mathbb{R}^{K+1}$ as prediction logits, where $W \in \mathbb{R}^{(K+1)\times d}$. The typical supervised logistic loss is $\ell(g \circ \phi(x), z) := \ell_u\left(\{g(\phi(\mathbf{x}))_z - g(\phi(\mathbf{x}))_{z'}\}_{z' \neq z}\right)$. Similar to Arora et al. (2019), define supervised loss w.r.t the task $\mathcal{T}$

$$\mathcal{L}_{sup}(\mathcal{T}, \phi) := \min_{W \in \mathbb{R}^{(K+1)\times d}} \mathop{\mathbb{E}}_{z\sim\mathcal{T}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}(z)}\left[\ell\left(W \cdot \phi(x), z\right)\right]. \tag{10}$$

Define supervised loss with mean classifier as $\mathcal{L}_{sup}^{\mu}(\mathcal{T}, \phi) := \mathop{\mathbb{E}}_{z\sim\mathcal{T}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}(z)}\left[\ell\left(W^\mu \cdot \phi(x), z\right)\right]$ where each row of W^μ is the mean of each class in $\mathcal{T}$, $W_{z_k}^\mu := \mu_{z_k} = \mathop{\mathbb{E}}_{x\sim z_k}(\phi(x)), k = 1, \ldots, (K+1)$. In the target task, suppose we have $K+1$ distinct classes from $\mathcal{C}$ with equal weights. Consider $\mathcal{T}$ follows a general distribution ζ. Define expected supervised loss as $\mathcal{L}_{sup}(\phi) := \mathop{\mathbb{E}}_{\mathcal{T}\sim\zeta}\left[\mathcal{L}_{sup}(\mathcal{T}, \phi)\right]$.

We formalize our assumption (**R**) below.

Assumption 1 (Regularity Conditions)**.** *The following regularity conditions hold:*

- ***(A1)*** *Representation function ϕ satisfies $\|\phi\|_2 \leq R$.*
- ***(A2)*** *Linear operator W satisfies bounded sprectral norm $\|W\|_2 \leq B$.*
- ***(A3)*** *The loss function ℓ_u are bounded by $[0, C]$ and $\ell(\cdot)$ is L-Lipschitz.*
- ***(A4)*** *The supervised loss $\mathcal{L}_{sup}(\mathcal{T}, \phi)$ is $\tilde{L}$-Lipschitz with respect to ϕ for $\forall\mathcal{T}$.*

A.1 PRE-TRAINING ERROR

In this section, we present pre-training error in binary classification (i.e. $K = 1$) and $\mathcal{D}_{\mathcal{T}}(z)$ as uniform. See Theorem 6 for the result for the general condition with multi-class in Appendix B.

We re-state the theorem below.

Theorem 1. *Assume Assumption (**R**) and that Φ has (ν, ϵ)-diversity for ζ and $\mathcal{C}_0$. Suppose $\hat{\phi}$ satisfies $\hat{\mathcal{L}}_{un}(\hat{\phi}) \leq \epsilon_0$. Let $\tau := \Pr_{(z_1,z_2)\sim\eta^2}\{z_1 = z_2\}$. Then for any target task $\mathcal{T}_0 \subset \mathcal{C}_0$,*

$$\mathcal{L}_{sup}(\mathcal{T}_0, \hat{\phi}) - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) \leq \frac{1}{\nu}\left[\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon. \tag{6}$$

The contrastive sample complexity is $O(\frac{\mathcal{R}_N(\Phi)}{\epsilon_0} + \frac{\log(1/\delta)}{\epsilon_0^2})$. The first term is the Rademacher complexity of the entire representation space Φ with sample size N. The second term relates to the generalization bound.

Proof of Theorem 1. Recall in binary classes, $\mathcal{S}_{un} = \left\{x_j, x_j^+, x_j^-\right\}_{j=1}^N$ denote our contrastive training set, sampled from $\mathcal{D}_{\text{con}}(\eta)$. Then by Lemma A.2 in Arora et al. (2019), with **(A1)** and **(A3)**, we have for $\forall\phi \in \Phi$ with probability $1 - \delta$,

$$\mathcal{L}_{un}(\phi) - \widehat{\mathcal{L}}_{un}(\phi) \leq \frac{4LR\mathcal{R}_N(\Phi)}{N} + C\sqrt{\frac{\log\frac{1}{\delta}}{N}}. \tag{11}$$

To have above $\leq \epsilon_0$, we have sample complexity

$$N \geq \frac{1}{\epsilon_0}\left[8LR\mathcal{R}_N(\Phi) + \frac{8C^2}{\epsilon_0}\log(\frac{2}{\delta})\right].$$

Consider in pre-training we have $\hat{\phi}$ such that

$$\hat{\mathcal{L}}_{un}(\hat{\phi}) \leq \epsilon_0.$$

Then with the above sample complexity, we have pre-training $\hat{\phi}$

$$\mathcal{L}_{un}(\hat{\phi}) \leq 2\epsilon_0.$$

Recall with (ν, ϵ)-diversity, for any task $\mathcal{T}$, we have that for $\hat{\phi}$ and ϕ^*,

$$\begin{aligned}
\mathcal{L}_{sup}(\mathcal{T}, \hat{\phi}) &\leq \mathcal{L}_{sup}(\mathcal{T}, \phi^*) + \frac{1}{\nu}\bar{d}_{\zeta}(\hat{\phi}, \phi^*) + \epsilon && (12)\\
&\leq \mathcal{L}_{sup}(\mathcal{T}, \phi^*) + \frac{1}{\nu}\left[\mathcal{L}_{sup}(\hat{\phi}) - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon && (13)\\
&\leq \mathcal{L}_{sup}(\mathcal{T}, \phi^*) + \frac{1}{\nu}\left[\frac{1}{1-\tau}(\mathcal{L}_{un}(\hat{\phi}) - \tau) - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon, && (14)
\end{aligned}$$

where last inequality comes from Lemma 4.3 in Arora et al. (2019): for $\forall\phi \in \Phi$, $\mathcal{L}_{sup}(\phi) \leq L_{sup}^{\mu}(\phi) \leq \frac{1}{1-\tau}(\mathcal{L}_{un}(\phi) - \tau)$.

For such $\hat{\phi}$ and ϕ^*, we have for target task $\mathcal{T}_0$, the bound reduces to

$$\begin{aligned}
\mathcal{L}_{sup}(\mathcal{T}_0, \hat{\phi}) - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) &\leq \frac{1}{\nu}\left[\frac{1}{1-\tau}(\mathcal{L}_{un}(\hat{\phi}) - \tau) - \epsilon^*\right] + \epsilon && (15)\\
&\leq \frac{1}{\nu}\left(\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \epsilon^*\right) + \epsilon. && (16)
\end{aligned}$$

□

A.2 FINETUNING ERROR

In this section, we provide proof of Theorem 2. We consider binary classification for simplicity for now, see Theorem 7 for the result with multi-class in Appendix B.

Following the intuition in (Garg & Liang, 2020), we first re-state the definition of representation space.

Definition 2. *The subset of representation space is*

$$\Phi(\epsilon_0) = \left\{\phi \in \Phi : \hat{\mathcal{L}}_{un}(\phi) \leq \epsilon_0\right\}.$$

Recall $\mathcal{S} = \{(x_j^i, z_j^i) : i \in [M], j \in [m]\}$ as finetuning dataset.

We define two function classes and associated Rademacher complexity.

Definition 3. *Consider function class*

$$\mathcal{G}_\ell = \left\{g_{W,\phi}(x,z) : g_{W,\phi}(x,z) = \ell(W \cdot \phi(x_j^i), z_j^i), \phi \in \Phi(\epsilon_0), \|W\|_2 \leq B\right\}.$$

We define Rademacher complexity as

$$\mathcal{R}_n(\mathcal{G}_\ell) = \mathop{\mathbb{E}}_{\{\sigma_i\}_{j=1}^n, \{x_j, z_j\}_{j=1}^n} \left[\sup_{\ell \in \mathcal{G}_\ell} \sum_{j=1}^n \sigma_j \ell(W \cdot \phi(x_j), z_j)\right].$$

Definition 4. *Consider function class*

$$\mathcal{G}(\epsilon_0) = \{g_\phi : g_\phi(\mathcal{T}) = \mathcal{L}_{sup}(T, \phi), \phi \in \Phi(\epsilon_0)\}.$$

We define Rademacher complexity as

$$\mathcal{R}_M(\mathcal{G}(\epsilon_0)) = \mathop{\mathbb{E}}_{\{\sigma_i\}_{i=1}^M, \{\mathcal{T}_i\}_{i=1}^M} \left[\sup_{\phi \in \Phi(\epsilon_0)} \sum_{i=1}^M \sigma_i \mathcal{L}_{sup}(\mathcal{T}_i, \phi)\right].$$

We re-state the theorem below.

Theorem 2. *Assume Assumption (**R**) and that Φ has (ν, ϵ)-diversity for ζ and $\mathcal{C}_0$. Suppose for some small constant $\alpha \in (0,1)$, we solve (4) with empirical loss lower than $\epsilon_1 = \frac{\alpha}{3}\frac{1}{1-\tau}(2\epsilon_0 - \tau)$ and obtain ϕ'. For any $\delta > 0$, if*

$$M \geq \frac{1}{\epsilon_1}\left[4\sqrt{2}\tilde{L}\mathcal{R}_M(\Phi(\epsilon_0)) + \frac{4C^2}{\epsilon_1}\log(\frac{2}{\delta})\right], Mm \geq \frac{1}{\epsilon_1}\left[16LB\mathcal{R}_{Mm}(\Phi(\epsilon_0)) + \frac{4C^2}{\epsilon_1}\log(\frac{2}{\delta})\right],$$

then with probability $1-\delta$, for any target task $\mathcal{T}_0 \subseteq \mathcal{C}_0$,

$$\mathcal{L}_{sup}(\mathcal{T}_0, \phi') - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) \leq \frac{1}{\nu}\left[\alpha\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon. \tag{7}$$

Proof of Theorem 2. Recall with (ν, ϵ)-diversity, for any task $\mathcal{T}$, we have that for ϕ' and ϕ^*,

$$\mathcal{L}_{sup}(\mathcal{T}, \phi') \leq \mathcal{L}_{sup}(\mathcal{T}, \phi^*) + \frac{1}{\nu}\bar{d}_\zeta(\phi', \phi^*) + \epsilon \tag{17}$$

$$\leq \mathcal{L}_{sup}(\mathcal{T}, \phi^*) + \frac{1}{\nu}\left[\mathcal{L}_{sup}(\phi') - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon \tag{18}$$

$$\leq \frac{1}{\nu}\left[\alpha\frac{1}{1-\tau}(2\epsilon_0 - \tau) - \epsilon^*\right] + \epsilon, \tag{19}$$

□

where the last inequality comes from Lemma 3.

Lemma 3. *Consider the notation and sample complexity in Theorem 2, solving (4) with empirical risk lower than ϵ_1 is sufficient to learn an ϕ' with expected supervised loss $\mathcal{L}_{sup}(\phi') \leq \alpha\frac{1}{1-\tau}(2\epsilon_0 - \tau)$, with probability $1-\delta$.*

Proof. Consider in (4) we have $\widehat{\mathbf{W}} := (\widehat{W}_1, \ldots, \widehat{W}_M)$ and ϕ' such that $\frac{1}{M}\sum_{i=1}^{M}\frac{1}{m}\sum_{j=1}^{m}\ell(\widehat{W}_i \cdot \phi'(x_j^i), z_j^i) \leq \epsilon_1 < \frac{\alpha}{3}\epsilon_0$.

We tried to bound

$$\mathcal{L}_{sup}(\phi') - \frac{1}{m}\sum_{j=1}^{m}\ell(\widehat{W}_i \cdot \phi'(x_j^i), z_j^i).$$

Recall that

$$\mathcal{L}_{sup}(\mathcal{T}_i, \phi) = \min_{W\in\mathbb{R}^{(K+1)\times d}} \mathop{\mathbb{E}}_{z\sim\mathcal{T}_i} \mathop{\mathbb{E}}_{x\sim\mathcal{D}(z)} \left[\ell\left(W \cdot \phi(x), z\right)\right].$$

For $\forall\phi \in \Phi(\epsilon_0)$

$$\mathcal{L}_{sup}(\phi) = \mathbb{E}_{\mathcal{T}\sim\zeta}\left[\mathcal{L}_{sup}(\mathcal{T}, \phi)\right] = \mathbb{E}_{\mathcal{T}\sim\zeta}\left[\min_{W\in\mathbb{R}^{(K+1)\times d}} \mathop{\mathbb{E}}_{z\sim\mathcal{T}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}(z)} \left[\ell\left(W \cdot \phi(\mathbf{x}), z\right)\right]\right].$$

We have for $\forall\phi \in \Phi(\epsilon_0)$, by uniform convergence (see Mohri et al. (2018) Theorem 3.3), we have with probability $1-\delta/2$

$$\mathbb{E}_{\mathcal{T}\sim\zeta}\left[\mathcal{L}_{sup}(\mathcal{T}, \phi)\right] - \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{sup}(\mathcal{T}_i, \phi) \leq \frac{2\mathcal{R}_M(\mathcal{G}(\epsilon_0))}{M} + \sqrt{\frac{\log(2/\delta)}{M}} \tag{20}$$

$$\leq \frac{2\sqrt{2}\tilde{L}\mathcal{R}_M(\Phi(\epsilon_0))}{M} + \sqrt{\frac{\log(2/\delta)}{M}}, \tag{21}$$

where the last inequality comes from **(A4)** and Corollary 4 in Maurer (2016). To have above $\leq \epsilon_1/2$, we have sample complexity

$$M \geq \frac{1}{\epsilon_1}\left[4\sqrt{2}\tilde{L}\mathcal{R}_M(\Phi(\epsilon_0)) + \frac{4C^2}{\epsilon_1}\log(\frac{2}{\delta})\right].$$

Then we consider generalization bound for $\forall\phi$ and $\mathbf{W} := (W_1, \ldots, W_M)$

$$\mathcal{L}_{sup}(\phi, \mathbf{W}) = \frac{1}{M}\sum_{i=1}^{M}\mathop{\mathbb{E}}_{z^i\sim\mathcal{T}_i} \mathop{\mathbb{E}}_{x^i\sim\mathcal{D}(z^i)} \ell\left(W_i \cdot \phi(x^i), z^i\right) \tag{22}$$

$$\hat{\mathcal{L}}_{sup}(\phi, \mathbf{W}) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{m}\sum_{j=1}^{m}\ell(W_i \cdot \phi(x_j^i), z_j^i), \tag{23}$$

where $\mathbf{W} = (W_1, \ldots, W_M)$.

By uniform convergence (see Mohri et al. (2018) Theorem 3.3), we have with probability $1-\delta/2$,

$$\mathcal{L}_{sup}(\phi, \mathbf{W}) - \hat{\mathcal{L}}_{sup}(\phi, \mathbf{W}) \leq \frac{2\mathcal{R}_{Mm}(\mathcal{G}_\ell)}{Mm} + \sqrt{\frac{\log(2/\delta)}{Mm}} \leq \frac{8\sqrt{K}LB\mathcal{R}_{Mm}(\Phi(\epsilon_0))}{Mm} + C\sqrt{\frac{\log(2/\delta)}{Mm}},$$

where the last inequality comes from Lemma 4. Recall in binary classification we have $K = 1$, to have above $\leq \epsilon_1/2$, we have sample complexity

$$Mm \geq \frac{1}{\epsilon_1}\left[16LB\mathcal{R}_{Mm}(\Phi(\epsilon_0)) + \frac{4C^2}{\epsilon_1}\log(\frac{2}{\delta})\right],$$

satisfying $\forall\phi \in \Phi(\epsilon_0)$

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_{sup}(\mathcal{T}_i, \phi) &= \frac{1}{M}\sum_{i=1}^{M}\min_{W\in\mathbb{R}^{(K+1)\times d}} \mathop{\mathbb{E}}_{z\sim\mathcal{T}_i} \mathop{\mathbb{E}}_{x\sim\mathcal{D}(z)} \left[\ell\left(W \cdot \phi(x), z\right)\right] \\
&\leq \frac{1}{M}\sum_{i=1}^{M}\mathop{\mathbb{E}}_{z\sim\mathcal{T}_i} \mathop{\mathbb{E}}_{x\sim\mathcal{D}(z)} \left[\ell\left(\widehat{W}_i \cdot \phi(x), z\right)\right] \\
&= \mathcal{L}_{sup}(\phi, \widehat{\mathbf{W}}) \\
&\leq \hat{\mathcal{L}}_{sup}(\phi, \widehat{\mathbf{W}}) + \epsilon_1/2.
\end{aligned}$$

Then combine above with (20)

$$\begin{aligned}\mathcal{L}_{sup}(\phi) &= \mathop{\mathbb{E}}_{\mathcal{T}\sim\zeta}[\mathcal{L}_{sup}(\mathcal{T},\phi)] \\ &\leq \hat{\mathcal{L}}_{sup}(\phi,\widehat{\mathbf{W}}) + \epsilon_1.\end{aligned}$$

We have

$$\begin{aligned}&\mathcal{L}_{sup}(\phi') - \frac{1}{m}\sum_{j=1}^{m}\ell(\widehat{W}_i\cdot\phi'(x_j^i), z_j^i) \leq \epsilon_1 \\ &\mathcal{L}_{sup}(\phi') \leq 2\epsilon_1 \leq \alpha\frac{1}{1-\tau}(2\epsilon_0-\tau).\end{aligned}$$

□

Lemma 4 (Bounded Rademacher complexity)**.** *By **(A2)** and **(A3)**, we have for $\forall n$*

$$\mathcal{R}_n(\mathcal{G}_\ell) \leq 4\sqrt{K}LB\mathcal{R}_n(\Phi(\epsilon_0)).$$

Proof. We first prove $\ell(g(\phi(x)), z)$ is $\sqrt{2K}LB$-Lipschitz with respect to ϕ for all $\forall z \in \mathcal{C}$. Consider

$$f_z(g(\phi(\mathbf{x}))) = \{g(\phi(\mathbf{x}))_z - g(\phi(\mathbf{x}))_{z'}\}_{z'\neq z},$$

where $f_z : \mathbb{R}^{K+1} \mapsto \mathbb{R}^K$. Note that

$$\begin{aligned}\ell(g\circ\phi(x), z) &= \ell\left(\{g(\phi(\mathbf{x}))_z - g(\phi(\mathbf{x}))_{z'}\}_{z'\neq z}\right) \\ &= \ell(f_z(g(\phi(\mathbf{x})))).\end{aligned}$$

By **(A3)**, we have ℓ is L-Lipschitz. We then prove f_z is $\sqrt{2K}$-Lipschitz. Without loss generality, consider $z = K+1$. We have $f_z(y) = [y_{K+1} - y_i]_{i=1}^K$. We have $\frac{\partial f_j}{y_i} = -\mathbb{1}\{j=i\}, i = 1,\ldots,K$, $\frac{\partial f_j}{y_{K+1}} = 1$. The Jacobian J satisfies $\|J\|_2 \leq \|J\|_F = \sqrt{2K}$.

g is B-Lipschitz by **(A2)**:$\|W\|_2 \leq B$. Then $\ell(g(\phi(x)), z)$ is $\sqrt{2K}LB$-Lipschitz with respect to ϕ for all $\forall z \in \mathcal{C}$. The conclusion follows Corollary 4 in Maurer (2016). □

B MULTI-CLASS CLASSIFICATION

In this section, we provide a general result for multi-classes with general distribution $\mathcal{D}_{\mathcal{T}}(z)$.

Lemma 5 (Theorem 6.1 in Arora et al. (2019))**.** *For multi-classes, we have*

$$\mathcal{L}_{sup}(\phi) \leq \mathcal{L}_{sup}^{\mu}(\phi) \leq \frac{1}{1-\tau_K}\mathcal{L}_{un}(\phi), \tag{24}$$

where $\tau_K = \mathop{\mathbb{E}}_{(z,z_1^-,\ldots,z_K^-)\sim\eta^{K+1}} \mathbb{1}\{z \textit{ does not appear in } (z_1^-,\ldots,z_K^-)\}$.

The proof of Lemma 5 follows the first two steps in the proof of Theorem B.1 of Arora et al. (2019).

Theorem 6 (Pre-training sample complexity in multi-classes)**.** *Consider a pre-training set $\mathcal{S}_{un} = \left\{x_j, x_j^+, x_{j1}^-,\ldots,x_{jK}\right\}_{j=1}^N$, by pre-training we get $\hat{\phi}$ with empirical contrastive loss $\hat{\mathcal{L}}_{un}(\hat{\phi}) \leq \epsilon_0$. For target task $\mathcal{T}_0$, with sample complexity*

$$N \geq \frac{1}{\epsilon_0}\left[8LR\sqrt{K}\mathcal{R}_N(\Phi) + \frac{8C^2}{\epsilon_0}\log(\frac{2}{\delta})\right],$$

it's sufficient to learn an $\hat{\phi}$ with classification error $\mathcal{L}_{sup}(\mathcal{T}_0,\hat{\phi}) - \mathcal{L}_{sup}(\mathcal{T}_0,\phi^*) \leq \frac{1}{\nu}\left[\frac{1}{1-\tau}(2\epsilon_0-\tau) - \epsilon^*\right] + \epsilon$, with probability $1-\delta$.

Proof of Theorem 6. Following similar step of proof of Theorem 1, we have with

$$N \geq \frac{1}{\epsilon_0}\left[8LR\sqrt{K}\mathcal{R}_N(\Phi) + \frac{8C^2}{\epsilon_0}\log(\frac{2}{\delta})\right].$$

We have pre-training $\hat{\phi}$

$$\mathcal{L}_{un}(\hat{\phi}) \leq 2\epsilon_0.$$

With (ν, ϵ)-diversity, for any task $\mathcal{T}$, we have that for $\hat{\phi}$ and ϕ^*,

$$\mathcal{L}_{sup}(\mathcal{T}, \hat{\phi}) \leq \mathcal{L}_{sup}(\mathcal{T}, \phi^*) + \frac{1}{\nu}\bar{d}_\zeta(\hat{\phi}, \phi^*) + \epsilon \tag{25}$$

$$\leq \mathcal{L}_{sup}(\mathcal{T}, \phi^*) + \frac{1}{\nu}\left[\mathcal{L}_{sup}(\hat{\phi}) - \mathcal{L}_{sup}(\phi^*)\right] + \epsilon. \tag{26}$$

Consider Lemma 5, we have for target task $\mathcal{T}_0$

$$\mathcal{L}_{sup}(\mathcal{T}_0, \hat{\phi}) - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) \leq \frac{1}{\nu}\left[\frac{1}{1-\tau_K}\mathcal{L}_{un}(\hat{\phi}) - \epsilon^*\right] + \epsilon \tag{27}$$

$$= \frac{1}{\nu}\left(\frac{2\epsilon_0}{1-\tau_K} - \epsilon^*\right) + \epsilon. \tag{28}$$

□

Below, we provide our main result similar to Theorem 2 for multi-classes setting.

Theorem 7. *For target evaluation task $\mathcal{T}_0$, consider the error bound in pre-training is $\mathcal{L}_{sup}(\mathcal{T}_0, \hat{\phi}) - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) \leq \frac{1}{\nu}\left[\frac{2\epsilon_0}{1-\tau_K} - \epsilon^*\right] + \epsilon$. Consider α as any small constant, for any $\epsilon_1 < \frac{\alpha}{3}\frac{2\epsilon_0}{1-\tau_K}$, consider a multitask finetuning set $\mathcal{S} = \{(x_j^i, z_j^i) : i \in [M], j \in [m]\}$, with M number of tasks, and m number of samples in each task. Then, with sample complexity*

$$M \geq \frac{1}{\epsilon_1}\left[4\sqrt{2}\tilde{L}\mathcal{R}_M(\Phi(\epsilon_0)) + \frac{4C^2}{\epsilon_1}\log(\frac{2}{\delta})\right]$$

$$Mm \geq \frac{1}{\epsilon_1}\left[16LB\sqrt{K}\mathcal{R}_{Mm}(\Phi(\epsilon_0)) + \frac{4C^2}{\epsilon_1}\log(\frac{2}{\delta})\right].$$

Solving (4) with empirical risk lower than ϵ_1 is sufficient to learn an ϕ' with classification error $\mathcal{L}_{sup}(\mathcal{T}_0, \phi') - \mathcal{L}_{sup}(\mathcal{T}_0, \phi^*) \leq \frac{1}{\nu}(\alpha\frac{2\epsilon_0}{1-\tau_K} - \epsilon^*) + \epsilon$, with probability $1-\delta$.

The proof follows the same steps in the proof of Theorem 2.

C Experimental Results

C.1 Vision tasks

C.1.1 Datasets

The miniImageNet dataset contains 100 classes sampled from ILSVRC-2012 (Russakovsky et al., 2015), then are randomly split into 64, 16, and 20 classes as training, validation, and testing set respectively following the protocol in (Chen et al., 2021).

The tieredImageNet dataset contains 608 classes from 34 super-categoriessampled from ILSVRC-2012. They are then split into 20, 6, 8 supercategories, resulting in 351, 97, 160 classes as training, validation, testing set respectively.

C.1.2 Experimental Setup

We use the SGD optimizer with momentum 0.9. The learning rate is fixed as 10^{-5}. The weight decay is 5.0×10^{-6}. In each few-shot task, we sample shot images and query images sum to m. We apply sampling for evaluating the performance. For the novel class split in a dataset, the sampling of testing few-shot tasks follows a deterministic order. We sample 1500 tasks and show the accuracy confidence interval below.

C.1.3 More Results

Task (M) vs Sample (m). We vary the task size and sample size per task during finetuning. We verify the trend of different numbers of tasks and numbers of images per task. Each task contains 5 classes. For finetuning tasks, $m = 50$ indicates each class contains the 1-shot image and 9-query images. $m = 100$ indicates each class contains 2-shot and 18-query images. $m = 200$ indicates each class contains 4-shot and 36-query images. $M = m = 0$ indicates direct evaluation without finetuning. For target tasks, each class contains the 1-shot image and 15 query images.

Task (M) \ Sample (m)	0	50	100	200
0	83.03 ± 0.24			
200		89.07 ± 0.20	89.95± 0.19	**90.09 ± 0.19**
400		89.31 ± 0.19	**90.11± 0.19**	90.70 ± 0.18
800		**89.71 ± 0.19**	90.27± 0.19	90.80 ± 0.18

Table 3: Accuracy with varying number of tasks and samples (ViT-B32 backbone).

Table 3 shows the results on the pre-trained CLIP model using ViT backbone. For direct adaptation without finetuning, the model achieves 83.03% accuracy. Multitask finetuning improves the average accuracy at least by 6%. For a fixed number of tasks or samples per task, increasing samples or tasks improves the accuracy. These results suggest that the total number of samples ($M \times m$) will determine the overall performance, supporting our main theorem.

Task Diversity. Task diversity is crucial for the foundation model to perform well on novel classes in target tasks. Task diversity can be measured by class diversity in finetuning stage. We vary the number of classes model access to in finetuning stage. The number of classes varies from all classes, i.e., 64 classes, to 8 classes. Each task contains 5 classes. For finetuning tasks, each class contains 1 shot image and 10 query images. For target tasks, each class contains the 1-shot image and 15 query images.

# limited classes	64	32	16	8	0
Accuracy	90.02 ± 0.15	88.54 ± 1.11	87.94 ± 0.22	87.07 ± 0.20	83.03 ± 0.24

Table 4: Class diversity on ViT-B32 backbone on miniImageNet.

Table 4 shows the accuracy of ViT-B32 on different numbers of classes in finetuning stage, where class 0 indicates direct evaluation without finetuning. Finetuning improves the average accuracy by 4%. As class diversity increases, performance increases.

Few Shot Effect. We perform experiments on the few-shot effects of finetuning tasks. We evaluate whether increasing few-shot images in finetuning task will provide significant improvement. Each task contains 5 classes. For finetuning tasks, each class contains 10 query images. For target tasks, each class contains 1 shot image and 15 query images.

# shot images	20	10	5	1	0
Accuracy	91.03 ± 0.18	90.93 ± 0.18	90.54 ± 0.18	90.02 ± 0.15	83.03 ± 0.24

Table 5: Few shot effect on ViT-B32 backbone on miniImageNet.

Table 5 shows the accuracy of ViT-B32 on different numbers of few-shot images in finetuning tasks. Increasing the few shot images, which will increase the sample number in each task, improves the performance. This corresponds to our sample complexity statement.

C.1.4 TASK DIVERSITY FOR OMNIGLOT

For task diversity, we also use dataset Omniglot (Lake et al., 2015). The Omniglot dataset is designed for developing more human-like learning algorithms. It contains 1623 different handwritten characters from 50 different alphabets. The 1623 classes are divided into 964, 301, and 358 classes as training, validation, and testing set respectively. We sample multitask in finetuning stage from training data and the target task from testing data.

# limited classes	964	482	241	50	10	0
Accuracy	95.35 ± 0.14	95.08 ± 0.14	94.29 ± 0.15	88.48 ± 0.20	80.26 ± 0.24	74.69 ± 0.26

Table 6: Class diversity on ViT-B32 backbone on Omniglot.

Table 6 shows the accuracy of ViT-B32 on different numbers of classes in finetuning stage, where class 0 indicates direct evaluation without finetuning. Finetuning improves the average accuracy by 5.5%. As class diversity increases, performance increases.

C.2 VISION LANGUAGE TASKS

C.2.1 IMPROVING ZERO-SHOT PERFORMANCE

We also examine how well CLIP models perform on miniImageNet in a zero-shot manner, using the protocol established for our vision tasks. Each task consists of 10 classes, with 15 query images per class. We use text features along with class information as the centroid to classify query images among the 10 classes. The text template utilized in this experiment was adapted from the CLIP documentation.

a photo of a {}
itap of a {}.
a bad photo of the {}.
a origami {}.
a photo of the large {}.
a {} in a video game.
art of the {}.
a photo of the small {}.

Table 7: Templates adapted from CLIP.

To obtain the centroid feature vector, we forward the text through the CLIP text encoder and calculate the average.

Backbone	Zero-shot	Multitask finetune
Accuracy	94.43 ± 0.05	95.03 ± 0.05

Table 8: Multitask finetune on zero-shot performance with ViT-B32 backbone on miniImageNet.

Table 8 demonstrates that CLIP already exhibits a high level of zero-shot performance. This is due to the model classifying images based on text information rather than relying on another image from the same class, which enables the model to utilize more accurate information to classify among query images. It is noteworthy that our multitask finetune paradigm still improves the zero-shot performance of the model.

	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)

Table 9: Our main results using RoBERTa-large. †: Result in (Gao et al., 2020); We report mean (and standard deviation) performance over 5 different splits of few-shot examples as in (Gao et al., 2020)

D LANGUAGE MODEL RESULTS

We also tested our pipeline for multitask finetuning on masked language models following the procedure described in Gao et al. (2020).

D.1 DATASETS

The text datasets contains 8 single-sentence and 7 sentence-pair English tasks, including 8 tasks from the GLUE benchmark (Wang et al., 2018), SNLI (Bowman et al., 2015), and 6 other popular sentence classification tasks (SST-5, MR, CR, MPQA, Subj, TREC). The goal is to predict the label based on a single sentence or a sentence-pair. For single sentences, we predict their semantics, whether they are positive or negative, while for sentence-pairs, we predict the relationship between them. We use $K = 16$ (per class) for few-shot experiments.

D.2 MULTITASK FINETUNING

To perform multitask finetuning, we select few-shot examples from other tasks as training finetuning examples for a specific target task (e.g., QNLI). We then multitask finetune the model using these selected finetuning examples, followed by prompt-based finetuning as described in Gao et al. (2020).

D.2.1 TASK SELECTION

We first forward text examples through the BERT backbone to obtain text features for each data point in the dataset. We then compute the first principal component and get one feature vector per dataset. We further perform training task selection based on the relative distance among the feature vectors extracted from each task. Our multitask finetuning protocol provided improvements for most of the datasets, as shown in Table 9.