The Trade-off between Label Efficiency and Universality of Representations from Contrastive Learning

Zhenmei Shi [*1] **Jiefeng Chen** [*1] **Kunyang Li** [1] **Jayaram Raghuram** [1] **Xi Wu** [2] **Yingyu Liang** [1] **Somesh Jha** [1]

Abstract

The pre-train representation learning paradigm is a recent popular approach to address distribution shift and limitations in training data. This approach first pre-trains a representation function using large unlabeled datasets from multiple tasks by self-supervised (e.g., contrastive) learning, and then learns a simple classifier on the representation using small labeled datasets from the downstream target tasks. The representation should have two key properties: *label efficiency* (i.e., ability to learn an accurate classifier with a small amount of labeled data) and *universality* (i.e., usefulness across a wide range of downstream tasks). In this paper, we focus on contrastive learning and systematically study the trade-off between label efficiency and universality both theoretically and empirically. We empirically show that this trade-off exists in different models and datasets. Theoretically, we propose a data model with a hidden representation and provide analysis in a simplified linear setting. Our analysis shows that compared to pre-training on the target task, pre-training on diverse tasks leads to a larger sample complexity for learning the optimal classifier, and thus has worse prediction performance.

1. Introduction

The pre-train representation learning paradigm is a recent successful approach to utilize large-scale unlabeled data to address the challenges of labeled data scarcity and distribution shift. Different from the traditional supervised learning approach using a large set of labeled data, representation learning first pre-trains a representation function using large-scale diverse unlabeled datasets by self-supervised learning (e.g., contrastive learning), and then learns a predictor on the representation using a small labeled dataset for a downstream target task. The pre-trained model is sometimes referred to as a *foundation model* (Bommasani et al., 2021), and has achieved good performance in many applications, e.g., BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020) and CLIP (Radford et al., 2021). While the foundation models exist for different applications and with different learning methods, the following two properties are key to their success: (1) *label efficiency*: with the pre-trained representation, only a small amount of labeled data is needed to build accurate predictors for downstream target tasks; (2) *universality*: the pre-trained representation can be used across various downstream tasks.

This work points out and studies a potential trade-off between label efficiency and universality, though ideally one would like to have these two key properties simultaneously. We focus on *contrastive learning* which is one exemplary pre-training method, while we believe similar observations and insights apply to other methods.

Empirically, we observe that such a trade-off indeed exists when the representation is pre-trained via contrastive learning on a large and diverse unlabeled dataset consisting of different tasks (or distributions). Since pre-training on diverse tasks is widely used in practice, such a trade-off deserves to be fully understood. More precisely, we perform controlled experiments comparing two cases: (1) specific representation pre-trained on a specific unlabeled dataset similar to that of the target task; (2) universal representation pre-trained on the union of diverse unlabeled datasets. The diverse data include the specific unlabeled dataset and some other datasets that can be quite different from that of the target task, which mimics the practical scenario that the foundation model is pre-trained on diverse data to be widely applicable for various downstream tasks. We observe that when the specific dataset is large, universal representation leads to worse prediction on the target, i.e., adding the diverse datasets in pre-training can harm the label efficiency though enhancing the universality. We also observe that if the specific dataset is small, then adding diverse datasets

[*] Equal contribution [1] Department of Computer Sciences, University of Wisconsin–Madison {zhmeishi, jiefeng, kli253, jayaramr, yliang, jha}@cs.wisc.edu [2] Google LLC wuxi@google.com. Correspondence to: Zhenmei Shi <zhmeishi@cs.wisc.edu>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

can help. These suggest the following explanation: *diverse unlabeled datasets have both positive and negative impacts* for prediction on a specific target task. On one hand, they share some useful semantic features with the target task, that can help learn these features in the representation and improve the prediction. On the other hand, they can have many other features not so useful for the target, and encoding such features into the representation essentially down-weights the useful ones and thus hurts prediction. When the specific dataset is large, the positive impact is marginal and overwhelmed by the negative impact, leading to the trade-off.

Theoretically, we provide analysis formalizing the intuition that the diverse unlabeled datasets have both positive and negative impacts by helping the representation learn various semantic features. We propose a *hidden representation model* for the data, which first generates a hidden representation containing various features, and then uses it to generate the label and the input. We then provide analysis in a simplified setting with linear data and representation functions. We show that contrastive learning can learn hidden features invariant to the transformations, and thus allows an accurate predictor if the label depends only on such invariant features. On the other hand, when pre-trained on diverse unlabeled data, it encodes all invariant features from different tasks and essentially emphasizes those common features but down-weights those specific to the target task. This then leads to a larger sample complexity for prediction on the target task and thus a worse generalization performance.

Related Work. We only discuss the most related ones here and include more in Appendix A. Cole et al. point out the "diversity-difficulty trade-off": pre-training on pooled datasets leads to worse performance on the in-domain task compared to pre-training on the in-domain dataset only. Similarly, Bommasani et al. call for further research on the issue of specialization vs. diversity in foundation model training. Neither of them provide a systematic investigation or theoretical analysis on the trade-off. Our work aims to provide empirical and theoretical analysis on how the pre-training data leads to the trade-off. We also note that existing theoretical analysis (e.g., (Arora et al., 2019; HaoChen et al., 2021)) typically assumes that the pre-training data distribution is the same as the target distribution, while the difference between the two is critical for the trade-off focused in this work. Thus, our work proposes a new analysis approach.

2. Theoretical Analysis

In this section, we will analyze the sample complexity of learning the predictor on top of the pre-trained representation. Here, the representation is pre-trained on diverse data (modeled by the union of unlabeled data from several tasks) via contrastive learning, to obtain representations that are potentially useful for a wide range of diverse tasks; while the predictor is learned for a specific target task using labeled data from that task. Existing analyses on contrastive learning are not applicable since they typically assume the same distribution for the pre-training and the prediction stages.

To model the intuition that the pre-training can learn semantic features from the unlabeled data such that a subset of them can be useful for prediction on even different data distributions, we propose a *hidden representation model* for the data, which first generates a hidden representation z and then uses z to generate the label y and the input x. While we are not able to perform analysis for the most general case, we consider a simplified linear setting where the data model and the representation functions are linear.

In this simplified setting, our analysis shows that contrastive learning can learn hidden features invariant to the transformation, and thus allows an accurate predictor if the label depends only on such invariant features. On the other hand, when pre-trained on a union of unlabeled data from different tasks, it encodes all invariant features from different tasks and essentially emphasizes those that are shared among the tasks, but down-weights those that are specific to a single task. Compared to pre-training only on unlabeled data from the target task, this then leads to a larger sample complexity for prediction on the target task (formalized by a larger Rademacher complexity of the predictor hypothesis class). Equivalently, this gives a worse generalization performance. Therefore, we formally show a trade-off between universality and the label efficiency.

Contrastive Learning. Let $\mathcal{X} = \mathbb{R}^d$ denote the input space, $\mathcal{Y}$ the label space, $\mathcal{Z} = \mathbb{R}^d$ the hidden representation space, $\overline{\mathcal{Z}} = \mathbb{R}^k$ the output space of the learned representation function. Let Φ denote the hypothesis class of representation functions $\phi : \mathcal{X} \mapsto \overline{\mathcal{Z}}$, and $\mathcal{F}_\phi$ the hypothesis class of predictors on the representation ϕ. Each task t is a data distribution $\mathcal{D}_t$ over $\mathcal{X} \times \mathcal{Y}$. In pre-training, using transformations on unlabeled data from the tasks, we have a distribution over positive pairs (x, x^+) and negative examples x^-, where x, x^- are two independent examples, while x^+ is obtained by applying some random transformations on x (e.g., cropping or color jitter for images). The contrastive loss is

$$-\log \frac{e^{\phi(x)^\top \phi(x^+)}}{e^{\phi(x)^\top \phi(x^+)} + e^{\phi(x)^\top \phi(x^-)}} = \ell\left(\phi(x)^\top (\phi(x^+) - \phi(x^-))\right)$$

where $\ell(\cdot)$ is the logistic loss $\ell(z) = \log(1 + \exp(-z))$. In practice multiple independent negative examples are used, and thus we consider the simplified contrastive loss:

$$\min_{\phi \in \Phi} \mathbb{E}_{(x,x^+)} \left[\ell\left(\phi(x)^\top (\phi(x^+) - \mathbb{E}_{x^-} \phi(x^-))\right)\right] \quad (1)$$

to pre-train a representation ϕ. Minimizing the contrastive loss maximizes the representation similarity between positive pairs x and x^+, while minimizes the representation similarity between negative pairs x and x^-. Then we learn a predictor f on top of ϕ using labeled data from a specific target task $\mathcal{D}_t$:

$$\min_{f \in \mathcal{F}_\phi} \mathbb{E}_{(x,y)\sim\mathcal{D}_t} \left[\ell_c(f(\phi(x)), y)\right] \tag{2}$$

where ℓ_c is a prediction loss (e.g. cross-entropy). Usually, f is a linear classifier, and for $\phi \in \Phi$:

$$\mathcal{F}_\phi = \{f(\phi) = u^\top \phi : u \in \mathbb{R}^k, \|u\|_2 \le B\} \tag{3}$$

for some norm bound B specified later.

Hidden Representation Data Model. Suppose the data is generated by first sampling a hidden representation $z \in \mathcal{Z}$ from some distribution, and then generating the input $x \in \mathcal{X}$ and the label $y \in \mathcal{Y}$ from the hidden representation z. We assume that the dimensions of z are partitioned into two subsets: *spurious features* U that are affected by the transformations, and *invariant features* $R = [d] \setminus U$. Intuitively, the transformation should be chosen such that useful semantic features will not be affected, and our goal is to recover the invariant features R. Therefore, we assume that $k = |R|$, and in each task t the label y will only depend on a subset of invariant features $R_t \subseteq R$ of size $|R_t| = r$. Formally, for $\mathcal{D}_t$, we assume $z_{R_t} \sim \mathcal{N}(0, I)$, $z_{R\setminus R_t} = 0$, $z_U \sim \mathcal{N}(0, I)$, and $y = (u_t^*)^\top z_{R_t}$ for some ground-truth parameter u_t^*. Then the positive pairs (x, x^+) are generated as follows:

$$z_{R_t} \sim \mathcal{N}(0, I),\ z_{R\setminus R_t} = 0,\ z_U \sim \mathcal{N}(0, I),\ z_U^+ \sim \mathcal{N}(0, I),$$
$$z = [z_R; z_U],\ z^+ = [z_R; z_U^+], \tag{4}$$

and x, x^+, and x^- are generated from different conditional distributions $\mathcal{D}_z$, $\mathcal{D}_{z^+}$, and $\mathcal{D}_{z^-}$ respectively.

A Simplified Setting. The general case of data distributions and pre-training is challenging to analyze. Here we consider a simplified setting with linear models and binary classification for the downstream task with $\mathcal{Y} = \{-1, +1\}$. Formally,
(1) x is generated from z linearly: $x = Mz$ where $M \in \mathbb{R}^{d\times d}$ is an orthonormal dictionary.
(2) The representations are linear functions with weight matrices of bounded Spectral and Frobenius norms:

$$\Phi = \{\phi(x) = Wx : W \in \mathbb{R}^{k\times d}, \|W\| \le 1, \|W\|_F \le \sqrt{r}\}.$$

Under this simplified setting, we show that pre-training on diverse tasks needs a larger Rademacher complexity to learn the optimal classifier compared to pre-training on the target task. We first assume that the pre-training is on an equal mixture of two tasks $\mathcal{D}_1$ and $\mathcal{D}_2$, with $R_1 = \{1, 2, \ldots, r\}$ and $R_2 = \{1, \ldots, s\} \cup \{r+1, r+2, \ldots, 2r-s\}$ where $s = |R_1 \cap R_2|$; the prediction is on the first task $\mathcal{D}_1$. We include the generalized analysis for multiple tasks in Appendix B. We first analyze the optimal representation.

Proposition 2.1. *There exist* $\min(1/2, s/r) \le \alpha \le 1$, $\beta = \min\left(1, \frac{r-\alpha s}{2(r-s)}\right) \in [1/2, 1]$ *such that* $\phi^*(x) = W^* x$ *is an optimal representation for the contrastive loss (1) with any* W^* *of the form:* $W^* = [OA^*, \mathbf{0}]M^{-1}$ *where* $O \in \mathbb{R}^{k\times k}$ *is any orthonormal matrix,* A^* *is diagonal with* $A^*_{jj} = \sqrt{\alpha}$ *if* $j \in R_1 \cap R_2$*, and* $A^*_{jj} = \sqrt{\beta}$ *otherwise, and the matrix of zeros* $\mathbf{0}$ *has size* $k \times (d-k)$*.*

That is, ϕ^* is a rotation of the weighted features, where the common features in $R_1 \cap R_2$ are weighted by $\sqrt{\alpha}$ and those task-specific features are weighted by $\sqrt{\beta}$.

Given the representation ϕ^*, we would like to ensure there exists a predictor in $\mathcal{F}_{\phi^*}$ matching the ground-truth label. Note that $f(\phi) = u^\top \phi$ with $u = O_{1:k,1:r}(A^*_{1:r,1:r})^{-1} u_1^* + O_{1:k,r+1:k}\, v$ for any $v \in \mathbb{R}^{k-r}$ satisfies $f(\phi^*(x)) = y = (u_1^*)^\top z_{R_1}$, and $u^* = O_{1:k,1:r}(A^*_{1:r,1:r})^{-1} u_1^*$ is the least-norm optimal solution. So the predictor class should be

$$\mathcal{F}_{\phi^*} = \{f(\phi^*) = u^\top \phi^* : u \in \mathbb{R}^k, \|u\|_2 \le \|u^*\|_2\}. \tag{5}$$

Proposition 2.2. *Let* $v_1 = \sum_{j=1}^s (u_{1j}^*)^2$ *and* $v_2 = \sum_{j=s+1}^r (u_{1j}^*)^2$*. Then the Rademacher complexity of* $\mathcal{F}_{\phi^*}$ *in Eqn. (5) satisfies* $\left|\mathcal{R}_m(\mathcal{F}_{\phi^*}) - \tilde{\mathcal{R}}_m(\mathcal{F}_{\phi^*})\right| \le O\left(\sqrt{\frac{1}{m}\left(\frac{1}{\alpha}v_1 + \frac{1}{\beta}v_2\right)}\right)$ *where the estimate* $\tilde{\mathcal{R}}_m(\mathcal{F}_{\phi^*})$ *is*

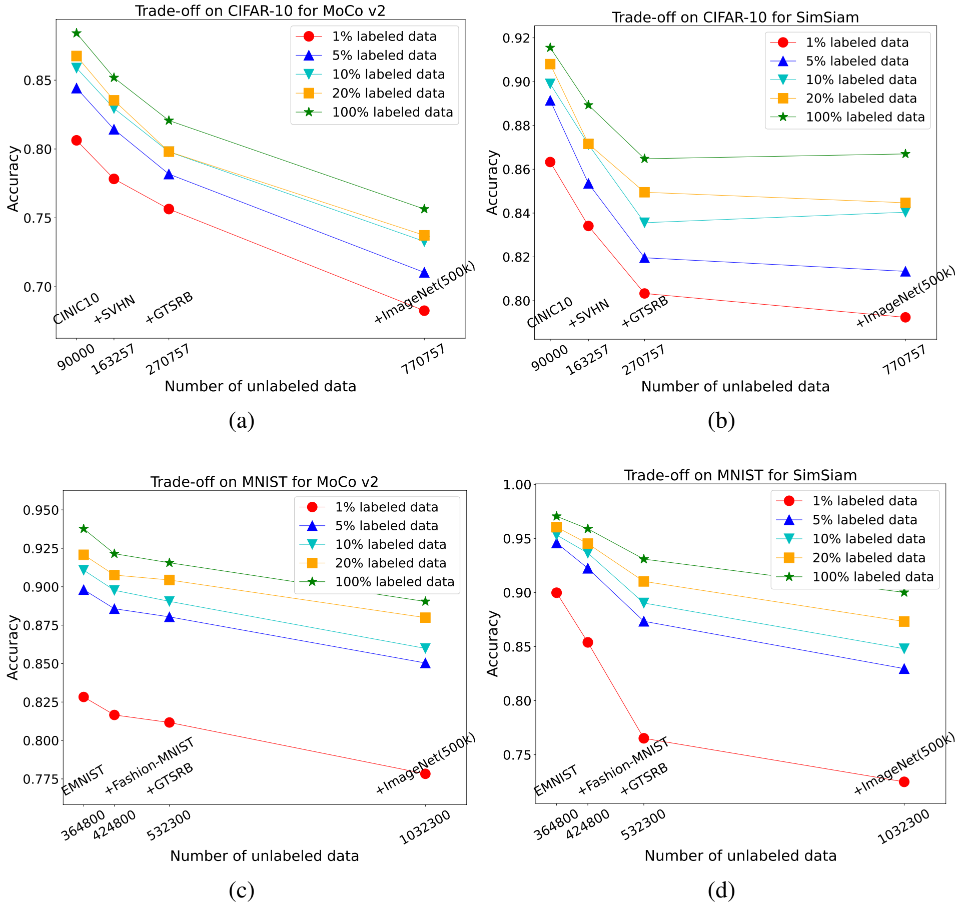
$$\tilde{\mathcal{R}}_m(\mathcal{F}_{\phi^*}) = \sqrt{\frac{1}{m}\left(\frac{1}{\alpha}v_1 + \frac{1}{\beta}v_2\right)(s\alpha + (r-s)\beta)}. \tag{6}$$

So ignoring the small-order term, the Rademacher complexity is roughly $\tilde{\mathcal{R}}_m(\mathcal{F}_{\phi^*})$. We now discuss the implication (details in Appendix B.3). When we pre-train on one task (equivalent to $r = s$), the complexity is roughly $\sqrt{\frac{r}{m}}\|u_1^*\|_2$. Consider pre-training on two tasks with $r = 2s$ and $v_1 = v_2 = \frac{\|u_1^*\|_2^2}{2}$. We can show that the optimal is $\alpha = 1, \beta = \frac{1}{2}$, and the complexity is roughly $\sqrt{\frac{9r}{8m}}\|u_1^*\|_2$. Therefore, the complexity of pre-training on two tasks is larger than just pre-training on the target task. This quantifies the trade-off between universality and label efficiency.

3. Experiments

We conduct experiments via contrastive learning to answer the following questions. (**Q1**) Does the trade-off between universality and label efficiency exist? (**Q2**) What conditions lead to the trade-off, and (**Q3**) how can we use the pretrain-finetune learning paradigm effectively?

We summarize the answers as follows: (**A1**) The trade-off widely exists in different models and datasets. (**A2**) When the task-relevant dataset is large enough, the task-irrelevant datasets will lead to the trade-off. (**A3**) Given knowledge of



(a)

(b)

(c)

(d)

Figure 1. Trade-off on downstream tasks CIFAR-10 and MNIST for MoCo v2 and SimSiam.

the downstream task, pre-training on a task-related dataset is better than pre-training on an unrelated dataset.

3.1. Experimental Setup

Model. We evaluate two popular contrastive learning frameworks, MoCo v2 (He et al., 2020) and SimSiam (Chen & He, 2021). MoCo v2 can be regarded as SimCLR (Chen et al., 2020) equipped with a memory bank, while SimSiam can be regarded as a modification from BYOL (Grill et al., 2020) similar to Barlow Twins (Zbontar et al., 2021), which does not need negative pairs.

Dataset. We consider two sets of data. In the first set, our downstream task is CIFAR-10, and the pre-training datasets may include CIFAR-10, CINIC-10, SVHN, GTSRB, and ImageNet32. CINIC-10 has classes identical to CIFAR-10 and is the most target-relevant, while the others are different. In the second set, our downstream task is MNIST, and the pre-training datasets may include EMNIST-Digits&Letters, Fashion-MNIST, GTSRB, and ImageNet32. Here, EMNIST-Digits&Letters is the most target-relevant.

Evaluation & Methods. We pre-train a ResNet18 network (He et al., 2016) as a feature extractor using SGD for 800 epochs with a cosine learning-rate schedule and a base learning rate of 0.06. Then we fix the pre-trained feature extractor, and train a linear classifier called Linear Probing (LP) on $1\%, 5\%, 10\%, 20\%, 100\%$ of the labeled data from the downstream task. For LP we use SGD for 100 epochs and a cosine learning rate schedule with a base learning rate of 5.0. We finally report the test accuracy on the downstream task.

3.2. Experimental Details

In Figs. 1(a) and (b), we report results for MoCo v2 and SimSiam (respectively) on CIFAR-10 as the downstream task. The size and diversity of unlabeled data for pre-training is increased on the x-axis by incrementally adding datasets in the following order: CINIC-10, SVHN, GTSRB, and ImageNet(500k). Then, we do LP on CIFAR-10 using different proportions of labeled samples. When the pre-training dataset is combined with more diverse data, the test accuracy for the specific downstream task decreases.

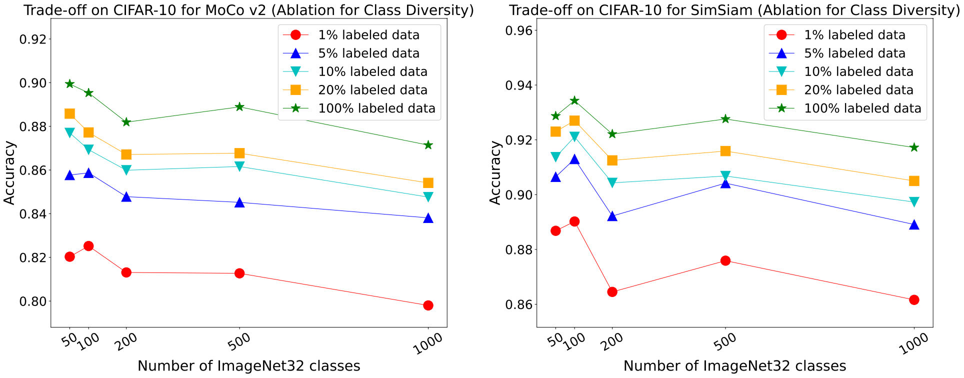


Figure 2. Varying the number of classes of ImageNet32 from 50 to 1000 under a fixed size of pre-training data.

As more diverse unlabeled data are included, more labeled data from the target task is needed to achieve a comparably-good prediction accuracy. This validates our hypothesis about the trade-off between universality and label efficiency. In Figs. 1(c) and (d), we report results for MoCo v2 and SimSiam (respectively) on MNIST as the downstream task. The size and diversity of unlabeled data for pre-training is increased on the x-axis by incrementally adding datasets in the following order: EMNIST-Digits&Letters, Fashon-MNIST, GTSRB, ImageNet(500k). This is followed by LP on MNIST with different proportion of labeled data. We observe the same trend in test accuracy as in Figs. 1(a) and (b). The handwritten dataset (EMNIST) is the most target-relevant, and it helps pre-training features suitable for the handwritten recognition task on MNIST. However, when we mix Fashion-MNIST, GTSRB, and ImageNet32 in the pre-training, the test accuracy on MNIST drops significantly. This supports our claim that pre-training on target-relevant data will learn more target-relevant features and get a better performance on the target task, while introducing diverse pre-training data will allow learning diverse features but can down-weight those for a specific task. The above analysis establishes that the trade-off widely exists in different models and datasets, answering **Q1**.

3.3. Ablation Study

We report results from three ablation studies: (1) varying the class number of ImageNet32, (2) varying the percentage of target-relevant pre-training data, and (3) replacing CINIC-10 with CIFAR-10 in the pre-training dataset.

Varying the Class Number of ImageNet32. To further support **A1**, we show that the trade-off between universality and label efficiency also exists under a fixed dataset size. In Fig. 2, we pre-train MoCo v2 and SimSiam on CIFAR10 + ImageNet(200k) and keep the same setting as Fig. 1 except that we vary the class number of ImageNet(200k). In previous experiments, we randomly pick 500,000 images from ImageNet32 without considering labels. Here, we fix the number of classes to 50, 100, 200, 500, 1000. Then we randomly sample 200,000 images from the subset of classes. The downstream task is CIFAR-10. In Fig. 2, we observe that with a fixed pre-training datasets size, e.g., 250,000, when the data is more diverse, the pre-training will learn more irrelevant features, and the performance will drop on the downstream task. This supports our analysis as well.

Due to the page limit, we provide details of (2) and (3) in Appendix C.2. We answer **Q2** from (2) that the data from diverse tasks may have a positive effect when the data from similar (relevant) tasks is not sufficiently large. Combining (3) with previous results, we answer **Q3**. If we choose a good task-relevant pre-training dataset, we can directly get similar performance as pre-training on the downstream task. However, the performance will drop if we introduce task-irrelevant data in the pre-training dataset.

4. Discussion and Future Work

In this work, we have shown that the trade-off between label efficiency and universality of representations widely exists in contrastive learning. For future work, there are many open questions we will continue to study. (1) What features does the model learn from specific pre-training and diverse pre-training datasets? (2) What properties do these features have? (3) Can we solve the trade-off in a better way in order to gain both properties at the same time?

Acknowledgements

The work is partially supported by Air Force Grant FA9550-18-1-0166, the National Science Foundation (NSF) Grants 2008559-IIS, CCF-2046710, 2023239-DMS, CCF-FMitF-1836978, SaTC-Frontiers-1804648, and CCF-1652140, and ARO grant number W911NF-17-1-0405. Jiefeng Chen and Somesh Jha are partially supported by the DARPA-GARD problem under agreement number 885000.

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Chen, X. and He, K. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15750–15758. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Cole, E., Yang, X., Wilber, K., Aodha, O. M., and Belongie, S. J. When does contrastive visual representation learning work? *CoRR*, abs/2105.05837, 2021. URL https://arxiv.org/abs/2105.05837.

Darlow, L. N., Crowley, E. J., Antoniou, A., and Storkey, A. J. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1422–1430. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.167. URL https://doi.org/10.1109/ICCV.2015.167.

Ericsson, L., Gouk, H., and Hospedales, T. M. How well do self-supervised models transfer? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 5414–5423. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Ericsson_How_Well_Do_Self-Supervised_Models_Transfer_CVPR_2021_paper.html.

Gansbeke, W. V., Vandenhende, S., Georgoulis, S., and Gool, L. V. Revisiting contrastive methods for unsupervised learning of visual representations. *CoRR*, abs/2106.05967, 2021. URL https://arxiv.org/abs/2106.05967.

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

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pp. 491–507. Springer, 2020. doi: 10.1007/978-3-030-58558-7_29. URL https://doi.org/10.1007/978-3-030-58558-7_29.

Kotar, K., Ilharco, G., Schmidt, L., Ehsani, K., and Mottaghi, R. Contrasting contrastive self-supervised representation learning pipelines. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9929–9939. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00980. URL https://doi.org/10.1109/ICCV48922.2021.00980.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Liu, H., HaoChen, J. Z., Gaidon, A., and Ma, T. Self-supervised learning is more robust to dataset imbalance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Newell, A. and Deng, J. How useful is self-supervised pretraining for visual tasks? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 7343–7352. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00737. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Newell_How_Useful_Is_Self-Supervised_Pretraining_for_Visual_Tasks_CVPR_2020_paper.html.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pp. 69–84. Springer, 2016. doi: 10.1007/978-3-319-46466-4_5. URL https://doi.org/10.1007/978-3-319-46466-4_5.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. Understanding contrastive learning requires incorporating inductive biases. *arXiv preprint arXiv:2202.14037*, 2022.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. doi: 10.1016/j.neunet.2012.02.016. URL https://doi.org/10.1016/j.neunet.2012.02.016.

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021.

Tsai, Y.-H. H., Wu, Y., Salakhutdinov, R., and Morency, L.-P. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*, 2020.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Wen, Z. and Li, Y. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pp. 11112–11122. PMLR, 2021.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yang, X., He, X., Liang, Y., Yang, Y., Zhang, S., and Xie, P. Transfer learning or self-supervised learning? A tale of two pretraining paradigms. *CoRR*, abs/2007.04234, 2020. URL https://arxiv.org/abs/2007.04234.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL http://proceedings.mlr.press/v139/zbontar21a.html.

Appendix

A. Related Work

Pretrain-finetune Learning Paradigm. The pretrain-finetune learning paradigm, where a model (or a representation) is pre-trained on a large dataset (e.g., ImageNet) and is then fine-tuned to various downstream tasks, has been widely used in practice (Devlin et al., 2019; Kolesnikov et al., 2020; Brown et al., 2020). There are mainly two kinds of pre-training approaches: one is the supervised pre-training (Kolesnikov et al., 2020), where we pre-train representations on large labeled datasets; the other is the self-supervised pre-training (Newell & Deng, 2020), where we pre-train representations on large and diverse unlabeled datasets. The self-supervised pre-training learning paradigm is sometimes referred to as the foundation models (Bommasani et al., 2021). Recently, it has been demonstrated that self-supervised pre-training can learn effective representations that even outperform the representations learned by supervised pre-training when evaluating them on downstream tasks (Ericsson et al., 2021). Also, some practical examples like BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020) and CLIP (Radford et al., 2021) have demonstrated the effectiveness of self-supervised pre-training in learning universal representations that can be used for a wide range of downstream tasks. In our work, we study the properties of self-supervised pre-training due to its superiority over supervised pre-training.

Self-supervised Representation Learning. Early self-supervised representation learning methods typically focus on solving hand-designed "pretext tasks" (Doersch et al., 2015; Gidaris et al., 2018; Noroozi & Favaro, 2016). Recent works have explored contrastive learning-based approaches where the pretext task is to distinguish matching and non-matching pairs of augmented input images (van den Oord et al., 2018). Common examples include SimCLR (Chen et al., 2020), MoCo (He et al., 2020), SimSiam (Chen & He, 2021), BYOL (Grill et al., 2020) and Barlow Twins (Zbontar et al., 2021).

Analysis of Self-supervised Pre-training Paradigm. Existing works have studied the effect of the pre-training datasets on the performance of the self-supervised pre-training (Arora et al., 2019; Tosh et al., 2021; Garg & Liang, 2020; HaoChen et al., 2021; Tsai et al., 2020; Wen & Li, 2021; Saunshi et al., 2022; Wang & Isola, 2020; Liu et al., 2021; Kotar et al., 2021; Gansbeke et al., 2021; Yang et al., 2020). However, existing analysis typically assumes that the pre-training data distribution is the same as the target distribution, while the difference between the two is the critical reason for the trade-off focused in this work. Thus, our work proposes new analysis approaches. Recently, Cole et al. have tried to identify conditions where self-supervised contrastive representation learning methods can produce "good" visual representations and point out the "diversity-difficulty trade-off" phenomenon, which is most relevant to our work. However, they only empirically show the trade-off, but do not provide a systematic study and analysis to explain why it happens. Bommasani et al. call for further research on the issue of specialization vs. diversity in foundation model training data, but do not provide a thorough study as well. Our work attempts to provide a better understanding of the trade-off between universality and label-efficiency.

B. Proof and More Analysis

We note that the contrastive loss can be written using logistic loss:

$$-\log \frac{e^{\phi(x)^\top \phi(x^+)}}{e^{\phi(x)^\top \phi(x^+)} + e^{\phi(x)^\top \phi(x^-)}} = \ell\left(\phi(x)^\top [\phi(x^+) - \phi(x^-)]\right) \tag{7}$$

where $\ell(\cdot)$ is the logistic loss $\ell(z) = \log(1 + \exp(-z))$. This will be useful for the analysis.

B.1. Proof of Propositions 2.1

Proof of Proposition 2.1. For each $\mathcal{D}_t$,

$$\begin{aligned}
\mathbb{E}_{(x,x^+)}\left[\ell\left(\phi(x)^\top[\phi(x^+) - \mathbb{E}_{x^-}\phi(x^-)]\right)\right] &= \mathbb{E}_{(z,z^+)}\left[\ell\left((WMz)^\top(WMz^+ - \mathbb{E}_{z^-}[WMz^-])\right)\right] && (8)\\
&= \mathbb{E}_{(z,z^+)}\left[\ell\left(z^\top(M^\top W^\top WM)(z^+ - \mathbb{E}_{z^-}[z^-])\right)\right] && (9)\\
&\geq \mathbb{E}_{z_R}\left[\ell\left((\mathbb{E}_{z_U}[z])^\top M^\top W^\top WM(\mathbb{E}_{z_U^+}[z^+] - \mathbb{E}_{z^-}[z^-])\right)\right] && (10)\\
&= \mathbb{E}_{z_R}\left[\ell\left([z_R; \mathbf{0}]^\top M^\top W^\top WM([z_R; \mathbf{0}] - 0)\right)\right] && (11)\\
&= \mathbb{E}_{z_R}\left[\ell\left(\|WM[z_R; \mathbf{0}]\|^2\right)\right] && (12)
\end{aligned}$$

where the inequality comes from the convexity of $\ell(z)$ and Jensen's inequality. Then on the mixture, we have

$$\mathbb{E}_{(x,x^+)}\left[\ell\left(\phi(x)^\top[\phi(x^+)-\mathbb{E}_{x^-}\phi(x^-)]\right)\right] = \sum_{t=1,2}\frac{1}{2}\mathbb{E}_{(x,x^+)\sim\mathcal{D}_t}\left[\ell\left(\phi(x)^\top[\phi(x^+)-\mathbb{E}_{x^-}\phi(x^-)]\right)\right] \tag{13}$$

$$\geq \sum_{t=1,2}\frac{1}{2}\mathbb{E}_{\mathcal{D}_t}\left[\ell\left(\|WM[z_{R_t};\mathbf{0}]\|^2\right)\right]. \tag{14}$$

Let $WM=[A_R,A_U]$ where $A_R\in\mathbb{R}^{k\times k}, A_U\in\mathbb{R}^{k\times(d-k)}$. By rotational invariance of $z_{R_1\cap R_2}, z_{R_1\setminus R_2}$, and $z_{R_2\setminus R_1}$, without loss of generality, we can assume $A_R=OA$ where A is a diagonal matrix with diagonal entries a_{jj}'s and O is any orthonormal matrix. Then

$$\sum_{t=1,2}\frac{1}{2}\mathbb{E}_{\mathcal{D}_t}\left[\ell\left(\|WM[z_{R_t};\mathbf{0}]\|^2\right)\right] = \sum_{t=1,2}\frac{1}{2}\mathbb{E}_{\{z_j\sim\mathcal{N}(0,1)\}}\left[\ell\left(\sum_{j\in R_t}a_{jj}^2z_j^2\right)\right] := g(\{a_{jj}\}). \tag{15}$$

Now consider the minimum of the function $g(\{a_{jj}\})$ on the right hand side, under the constraints that $|a_{jj}|\leq 1$ and $\sum_j a_{jj}^2\leq r$. We have the following claim for this optimization.

Lemma B.1. *There exists* $\min(1/2,s/r)\leq\alpha\leq 1$, $\beta=\min\left(1,\frac{r-\alpha s}{2(r-s)}\right)\in[1/2,1]$, *such that the minimum of the above optimization is achieved when* $a_{jj}^2=\alpha$ *for any* $j\in R_1\cap R_2$, *and* $a_{jj}^2=\beta$ *for any* $j\notin R_1\cap R_2$.

Proof. We first prove that to achieve the minimum, we can set:

(1) $a_{\ell\ell}^2=a_{\ell'\ell'}^2$ for any $\ell\neq\ell'\in R_1\cap R_2$;

(2) $a_{\ell\ell}^2=a_{\ell'\ell'}^2$ for any $\ell\neq\ell'\in R_1\setminus R_2$;

(3) $a_{\ell\ell}^2=a_{\ell'\ell'}^2$ for any $\ell\neq\ell'\in R_2\setminus R_1$.

For (1): By symmetry of z_j's and the convexity of $\ell(\cdot)$,

$$\mathbb{E}\left[\ell\left(\sum_{j\in R_1}a_{jj}^2z_j^2\right)\right] = \frac{1}{2}\mathbb{E}\left[\ell\left(\sum_{j\in R_1,j\neq\ell,j\neq\ell'}a_{jj}^2z_j^2+a_{\ell\ell}^2z_\ell^2+a_{\ell'\ell'}^2z_{\ell'}^2\right)\right] \tag{16}$$

$$+\frac{1}{2}\mathbb{E}\left[\ell\left(\sum_{j\in R_1,j\neq\ell,j\neq\ell'}a_{jj}^2z_j^2+a_{\ell\ell}^2z_{\ell'}^2+a_{\ell'\ell'}^2z_\ell^2\right)\right] \tag{17}$$

$$\geq\mathbb{E}\left[\ell\left(\sum_{j\in R_1,j\neq\ell,j\neq\ell'}a_{jj}^2z_j^2+\frac{a_{\ell\ell}^2+a_{\ell'\ell'}^2}{2}z_{\ell'}^2+\frac{a_{\ell\ell}^2+a_{\ell'\ell'}^2}{2}z_\ell^2\right)\right]. \tag{18}$$

A similar inequality holds for R_2. Then

$$g(\{a_{jj}\})\geq\sum_{t=1,2}\frac{1}{2}\mathbb{E}\left[\ell\left(\sum_{j\in R_t,j\neq\ell,j\neq\ell'}a_{jj}^2z_j^2+\frac{a_{\ell\ell}^2+a_{\ell'\ell'}^2}{2}z_{\ell'}^2+\frac{a_{\ell\ell}^2+a_{\ell'\ell'}^2}{2}z_\ell^2\right)\right]. \tag{19}$$

Therefore, the minimum is achieved when $a_{\ell\ell}^2=a_{\ell'\ell'}^2$.

For (2): The same inequality in Eqn. (16) holds for any $\ell\neq\ell'\in R_1\setminus R_2$, which then implies the statement (2).

For (3): The proof is similar to that for the statement (3).

These statements mean that the minimum is achieved when $a_{jj}^2=\alpha$ for $j\in R_1\cap R_2$, $a_{jj}^2=\alpha_1$ for $j\in R_1\setminus R_2$, and $a_{jj}^2=\alpha_2$ for $j\in R_2\setminus R_1$, for some values $\alpha,\alpha_1,\alpha_2\geq 0$. Let $Z=\sum_{j\in R_1\cap R_2}z_j^2, Z_1=\sum_{j\in R_1\setminus R_2}z_j^2, Z_2=\sum_{j\in R_2\setminus R_1}z_j^2$. By symmetry of z_j's, Z_1 and Z_2 follow the same distribution. Then

$$\begin{aligned} g(\{a_{jj}\}) &= \frac{1}{2}\mathbb{E}\left[\ell\left(\alpha Z + \alpha_1 Z_1\right)\right] + \frac{1}{2}\mathbb{E}\left[\ell\left(\alpha Z + \alpha_2 Z_2\right)\right] && (20) \\ &= \frac{1}{2}\mathbb{E}\left[\ell\left(\alpha Z + \alpha_1 Z_1\right)\right] + \frac{1}{2}\mathbb{E}\left[\ell\left(\alpha Z + \alpha_2 Z_1\right)\right] && (21) \\ &\geq \mathbb{E}\left[\ell\left(\alpha Z + \frac{\alpha_1 + \alpha_2}{2} Z_1\right)\right]. && (22) \end{aligned}$$

So the minimum is achieved when $\alpha_1 = \alpha_2 := \beta$, leading to

$$g(\{a_{jj}\}) = \mathbb{E}\left[\ell\left(\alpha Z + \beta Z_1\right)\right]. \quad (23)$$

Given the constraint $\alpha s + 2\beta(r - s) = \sum_j a_{jj}^2 \leq r$, $0 \leq \alpha, \beta \leq 1$, and that $\ell(\cdot)$ is monotonically non-increasing, we have $\alpha \in [0, 1], \beta = \min\left(1, \frac{r - \alpha s}{2(r-s)}\right)$.

Furthermore, we can show that $\alpha \geq \min(1/2, s/r)$. Suppose $\alpha < \min(1/2, s/r)$ for contradiction.

First consider the case when $r \leq 2s$. Then $1/2 \leq s/r$ and thus $\alpha < 1/2$. Note that

$$g(\{a_{jj}\}) = \mathbb{E}\left[\ell\left(\alpha \sum_{j=1}^{s} z_j^2 + \beta \sum_{j=s+1}^{r} z_j^2\right)\right]$$

and $\alpha \sum_{j=1}^{s} z_j^2 + \beta \sum_{j=s+1}^{r} z_j^2$ is stochastically dominated by $\sum_{j=1}^{s} z_j^2 + \frac{1}{2}\sum_{j=s+1}^{r} z_j^2$, which is achieved when $\alpha = 1$. So the optimal cannot be achieved when $\alpha < 1/2$.

Next consider the case when $r > 2s$. Then $1/2 > s/r$ and thus $\alpha < s/r$. We also have $\frac{r - \alpha s}{2(r-s)} < 1$ so $\beta = \frac{r - \alpha s}{2(r-s)}$. Let c_j be the coefficients such that $c_j = \alpha$ for $1 \leq j \leq s$ and $c_j = \beta$ for $s < j \leq r$. Then

$$\begin{aligned} g(\{a_{jj}\}) &= \mathbb{E}\left[\ell\left(\sum_{j=1}^{r} c_j z_j^2\right)\right] && (24) \\ &= \mathbb{E}\left[\ell\left(\sum_{j=1}^{r-s} c_j z_j^2 + \sum_{j=r-s+1}^{r} c_j z_j^2\right)\right]. && (25) \end{aligned}$$

Let Π be the set of all permutations of $[r - s]$. Again by symmetry of z_j's and the convexity of $\ell(\cdot)$,

$$\begin{aligned} g(\{a_{jj}\}) &= \frac{1}{|\Pi|}\sum_{\sigma \in \Pi} \mathbb{E}\left[\ell\left(\sum_{j=1}^{r-s} c_{\sigma(j)} z_j^2 + \sum_{j=r-s+1}^{r} c_j z_j^2\right)\right] && (26) \\ &\geq \mathbb{E}\left[\ell\left(\sum_{j=1}^{r-s} \frac{1}{|\Pi|}\sum_{\sigma \in \Pi} c_{\sigma(j)} z_j^2 + \sum_{j=r-s+1}^{r} c_j z_j^2\right)\right] && (27) \\ &= \mathbb{E}\left[\ell\left(\sum_{j=1}^{r-s} \frac{s\alpha + (r-2s)\beta}{r-s} z_j^2 + \sum_{j=r-s+1}^{r} \beta z_j^2\right)\right]. && (28) \end{aligned}$$

When $\alpha < s/r$, since $\beta = \frac{r-\alpha s}{2(r-s)}$, we have $\frac{s\alpha+(r-2s)\beta}{r-s} < 1/2$, so

$$g(\{a_{jj}\}) > \mathbb{E}\left[\ell\left(\sum_{j=1}^{r-s}\frac{1}{2}z_j^2 + \sum_{j=r-s+1}^{r}\beta z_j^2\right)\right] \tag{29}$$

$$\geq \mathbb{E}\left[\ell\left(\sum_{j=1}^{r-s}\frac{1}{2}z_j^2 + \sum_{j=r-s+1}^{r} z_j^2\right)\right] \tag{30}$$

$$= \mathbb{E}\left[\ell\left(\sum_{j=1}^{s} z_j^2 + \frac{1}{2}\sum_{j=s+1}^{r} z_j^2\right)\right]. \tag{31}$$

The right-most hand side is achieved when $\alpha = 1, \beta = 1/2$. So the optimal cannot be achieved when $\alpha < s/r$.

In summary, we have $\alpha \geq \min(1/2, s/r)$. □

Therefore,

$$\mathbb{E}_{(x,x^+)}\left[\ell\left(\phi(x)^\top[\phi(x^+) - \mathbb{E}_{x^-}\phi(x^-)]\right)\right] \geq \mathbb{E}\left[\ell\left(\alpha Z + \beta Z_1\right)\right]. \tag{32}$$

On the other hand, it can be verified that for any ϕ^* with W^*,

$$\mathbb{E}_{(x,x^+)}\left[\ell\left(\phi(x)^\top[\phi(x^+) - \mathbb{E}_{x^-}\phi(x^-)]\right)\right] = \mathbb{E}\left[\ell\left(\alpha Z + \beta Z_1\right)\right]. \tag{33}$$

Therefore, ϕ^* is the optimal solution. □

B.2. Proof of Proposition 2.2

Proof of Proposition 2.2. $\mathbb{E}_{\mathcal{D}_1}[(\hat{y}-y)^2] \geq 0$ and

$$\mathbb{E}_{\mathcal{D}_1}[(\hat{y}-y)^2] = 0 \Leftrightarrow \forall z_{R_1}, \quad u^\top[OA^*, \mathbf{0}]M^{-1}M[z_{R_1}; \mathbf{0}; z_U] = {u_1^*}^\top z_{R_1} \tag{34}$$

$$\Leftrightarrow \forall z_{R_1}, \quad u^\top OA^*[z_{R_1}; \mathbf{0}] = {u_1^*}^\top z_{R_1} \tag{35}$$

$$\overset{(*)}{\Leftrightarrow} A^*_{1:r,1:r}(O^\top)_{1:r,1:k}u = u_1^* \tag{36}$$

$$\Leftrightarrow \forall v \in \mathbb{R}^r, \quad u = O_{1:k,1:r}(A^*_{1:r,1:r})^{-1}u_1^* + O_{1:k,r+1:k}v. \tag{37}$$

The $(*)$ is from non-zero variance for z_{R_1}. The empirical Rademacher complexity and Rademacher complexity of $\mathcal{F}_{\phi^*}$ with m samples are

$$\hat{\mathcal{R}}_m(\mathcal{F}_{\phi^*}) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{f_{u,\phi}\in\mathcal{F}_{\phi^*}}\sum_{i=1}^m \sigma_i f_{u,\phi}(x^{(i)})\right] \tag{38}$$

$$= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\|u\|_2\leq\|u^*\|_2}\sum_{i=1}^m \sigma_i u^\top OA^*[z_{R_1}^{(i)}; \mathbf{0}]\right] \tag{39}$$

$$= \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{\|u\|_2\leq\|u^*\|_2} u^\top \sum_{i=1}^m \sigma_i O_{1:k,1:r}A^*_{1:r,1:r}z_{R_1}^{(i)}\right] \tag{40}$$

$$= \frac{\|u^*\|_2}{m}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i O_{1:k,1:r}A^*_{1:r,1:r}z_{R_1}^{(i)}\right\|_2\right], \tag{41}$$

$$\mathcal{R}_m(\mathcal{F}_{\phi^*}) = \mathbb{E}_{z_R, z_U}\left[\hat{\mathcal{R}}_m(\mathcal{F}_{\phi^*})\right] \tag{42}$$

$$= \frac{\|u^*\|_2}{m}\mathbb{E}_{z_{R_1}^{(i)}}\left[\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i O_{1:k,1:r}A^*_{1:r,1:r}z_{R_1}^{(i)}\right\|_2\right]\right] \tag{43}$$

$$= \frac{\|u^*\|_2}{m}\mathbb{E}_{z_{R_1}^{(i)}}\left[\left\|A^*_{1:r,1:r}\sum_{i=1}^m z_{R_1}^{(i)}\right\|_2\right]. \tag{44}$$

Define $X := A^*_{1:r,1:r} \sum_{i=1}^m z^{(i)}_{R_1}$. Note that for $j \in R_1 \cap R_2$, $X_j = \alpha \sum_{i=1}^m z^{(i)}_j$ is a Gaussian of mean zero and variance $\mathbb{E}[X_j^2] = \alpha \mathbb{E}\left[\left(\sum_{i=1}^m z^{(i)}_j\right)^2\right] = \alpha \mathbb{E}\left[\sum_{i=1}^m \left(z^{(i)}_j\right)^2\right] = m\alpha$. Similarly, for $j \in R_1 \setminus R_2$, $X_j = \beta \sum_{i=1}^m z^{(i)}_j$ is a Gaussian of mean zero and variance $\mathbb{E}[X_j^2] = m\beta$. Since X_j is sub-gaussian, $X_j^2 - m\alpha$ for $j \in R_1 \cap R_2$ and $X_j^2 - m\beta$ for $j \in R_1 \setminus R_2$ are sub-exponential and more precisely

$$\|X_j^2 - m\alpha\|_{\psi_1} \leq C_1 \|X_j^2\|_{\psi_1} = C_1 \|X_j\|^2_{\psi_2} \leq C_2 m\alpha, \quad j \in R_1 \cap R_2, \tag{45}$$

$$\|X_j^2 - m\beta\|_{\psi_1} \leq C_1 \|X_j^2\|_{\psi_1} = C_1 \|X_j\|^2_{\psi_2} \leq C_2 m\beta, \quad j \in R_1 \setminus R_2, \tag{46}$$

where C_1, C_2 are absolute constants and $C_2 > 1$. Let $K = \max(C_2 m\alpha, C_2 m\beta) \leq C_2 m$ and $\mu := m(s\alpha + (r-s)\beta)$. By Bernstein's inequality, we have for every $\gamma \geq 0$ that

$$\mathbb{P}\left\{\left|\frac{1}{r}(\|X\|_2^2 - \mu)\right| \geq \gamma\right\} \leq 2\exp\left[-c\min\left(\frac{\gamma^2}{K^2}, \frac{\gamma}{K}\right) r\right] \tag{47}$$

$$\Rightarrow \mathbb{P}\left\{\left|\frac{\|X\|_2^2}{\mu} - 1\right| \geq \frac{r\gamma}{\mu}\right\} \leq 2\exp\left[-\frac{c}{C_2^2}\min\left(\frac{\gamma^2}{m^2}, \frac{\gamma}{m}\right) r\right], \tag{48}$$

where c is an absolute constant. For all numbers $z \geq 0$, we have $|z-1| \geq \delta \Rightarrow |z^2 - 1| \geq \max(\delta, \delta^2)$. Thus, for any $\delta \geq 0$, we have

$$\mathbb{P}\left\{\left|\frac{\|X\|_2}{\sqrt{\mu}} - 1\right| \geq \delta\right\} \leq \mathbb{P}\left\{\left|\frac{\|X\|_2^2}{\mu} - 1\right| \geq \max(\delta, \delta^2)\right\} \tag{49}$$

$$\leq 2\exp\left[-\frac{c}{C_2^2}\min\left(\left(\frac{\mu\max(\delta,\delta^2)}{mr}\right)^2, \frac{\mu\max(\delta,\delta^2)}{mr}\right) r\right] \tag{50}$$

$$\leq 2\exp\left[-\frac{c}{C_2^2}\left(\frac{\mu}{mr}\right)^2 \min\left(\left(\max(\delta,\delta^2)\right)^2, \max(\delta,\delta^2)\right) r\right] \tag{51}$$

$$= 2\exp\left[-\frac{c}{C_2^2}\frac{\mu^2}{m^2 r}\delta^2\right], \tag{52}$$

where the last inequality is from $\mu \leq mr$. Changing variables to $\theta = \delta\sqrt{\mu}$, we obtain the desired sub-gaussian tail

$$\mathbb{P}\left\{|\|X\|_2 - \sqrt{\mu}| \geq \theta\right\} \leq 2\exp\left[-\frac{c}{C_2^2}\frac{\mu}{m^2 r}\theta^2\right]. \tag{53}$$

By generalization of integral identity, we have

$$|\mathbb{E}\left[\|X\|_2 - \sqrt{\mu}\right]| = \left|\int_0^\infty \mathbb{P}\{\|X\|_2 - \sqrt{\mu} > \theta\} d\theta - \int_{-\infty}^0 \mathbb{P}\{\|X\|_2 - \sqrt{\mu} < \theta\} d\theta\right| \tag{54}$$

$$\leq 2\int_0^\infty \mathbb{P}\{|\|X\|_2 - \sqrt{\mu}| > \theta\} d\theta \tag{55}$$

$$\leq 4\int_0^\infty \exp\left[-\frac{c}{C_2^2}\frac{\mu}{m^2 r}\theta^2\right] d\theta \tag{56}$$

$$\leq C_3 \frac{m\sqrt{r}}{\sqrt{\mu}} \tag{57}$$

$$\leq \sqrt{2m} C_3, \tag{58}$$

where C_3 is an absolute constant and the last inequality is from $\mu = m(s\alpha + (r-s)\beta) = \frac{1}{2}m(s\alpha + r) \geq \frac{1}{2}mr$. Thus, we have

$$\left|\mathcal{R}_m(\mathcal{F}_{\phi^*}) - \sqrt{\frac{1}{m}\left(\frac{1}{\alpha}v_1 + \frac{1}{\beta}v_2\right)(s\alpha + (r-s)\beta)}\right| = \frac{\|u^*\|_2}{m}|\mathbb{E}\left[\|X\|_2 - \sqrt{\mu}\right]| \tag{59}$$

$$\leq O\left(\sqrt{\frac{1}{m}\left(\frac{1}{\alpha}v_1 + \frac{1}{\beta}v_2\right)}\right). \tag{60}$$

□

B.3. Discussion on the Implications

Pre-training and Predicting on One Task. Suppose we only pre-train the representation and then learn a predictor on the task $\mathcal{D}_1$. This is equivalent to setting $r = s$, which leads to $\mathcal{D}_1 = \mathcal{D}_2$. Then by Proposition 2.1, we know that

$$\phi^*(x) = W^* x = O \sum_{j \in R_1} \sqrt{\alpha} z_j e_j$$

where e_j's are the basis vectors. The contrastive loss is

$$\mathbb{E}_{(x,x^+)} \left[\ell \left(\phi(x)^\top [\phi(x^+) - \mathbb{E}_{x^-} \phi(x^-)] \right) \right] = \mathbb{E} \left[\ell \left(\alpha \sum_{j=1}^{r} z_j^2 \right) \right].$$

Since z_j's are standard Gaussians, $\alpha = 1$ in the optimal. So $\phi^*(x) = W^* x = \sum_{j \in R_1} z_j e_j$.

Furthermore, we have $v_1 = \|u_1^*\|_2^2, v_2 = 0$ in Proposition 2.2, then the Rademacher complexity of $\mathcal{F}_{\phi^*}$ satisfies

$$\left| \mathcal{R}_m(\mathcal{F}_{\phi^*}) - \sqrt{\frac{r \|u_1^*\|_2^2}{m}} \right| \le O \left(\sqrt{\frac{\|u_1^*\|_2^2}{m}} \right). \tag{61}$$

Ignoring the low-order term on the right hand side, we have

$$\mathcal{R}_m(\mathcal{F}_{\phi^*}) \approx \sqrt{\frac{r}{m}} \|u_1^*\|_2. \tag{62}$$

Pre-training and Predicting on Two Tasks. Consider the case with $r = 2s$, that is, each task has half of the invariant features being common features, and the other half being task-specific features. Then by Proposition 2.1, we know that

$$\phi^*(x) = W^* x = O \times \left(\sum_{j \in R_1 \cap R_2} \sqrt{\alpha} z_j e_j + \sum_{j \in R \setminus (R_1 \cap R_2)} \sqrt{\beta} z_j e_j \right)$$

where e_j's are the basis vectors.

We now show that $\alpha = 1$ in the optimal. Since $r = 2s$, $\frac{r - \alpha s}{2(r-s)} = \frac{2-\alpha}{2} \le 1$, then $\beta = \min \left(\frac{r-\alpha s}{2(r-s)}, 1 \right) = \frac{2-\alpha}{2}$. The contrastive loss is

$$\begin{aligned}
\mathbb{E}_{(x,x^+)} \left[\ell \left(\phi(x)^\top [\phi(x^+) - \mathbb{E}_{x^-} \phi(x^-)] \right) \right] &= \mathbb{E} \left[\ell \left(\alpha \sum_{j \in R_1 \cap R_2} z_j^2 + \beta \sum_{j \in R_1 \setminus R_2} z_j^2 \right) \right] && (63) \\
&= \mathbb{E} \left[\ell \left(\alpha \sum_{j=1}^{s} z_j^2 + \frac{2-\alpha}{2} \sum_{j=s+1}^{r} z_j^2 \right) \right] && (64) \\
&= \mathbb{E} \left[\ell \left(\alpha Z + \frac{2-\alpha}{2} Z_1 \right) \right] && (65)
\end{aligned}$$

where $Z = \sum_{j=1}^{s} z_j^2$ and $Z_1 = \sum_{j=s+1}^{r} z_j^2$ are i.i.d. χ_s^2 random variables. Let $Q_\alpha = \alpha Z + \frac{2-\alpha}{2} Z_1$. Then it can be shown that Q_1 stochastically dominates Q_α for $\alpha < 1$. Then the loss $\mathbb{E}[\ell(Q_\alpha)]$ is minimized at $\alpha = 1$. So

$$\phi^*(x) = O \times \left(\sum_{j \in R_1 \cap R_2} z_j e_j + \sum_{j \in R \setminus (R_1 \cap R_2)} \sqrt{\frac{1}{2}} z_j e_j \right).$$

Then when $v_1 = v_2 = \|u_1^*\|_2^2 / 2$, by Proposition 2.2, the Rademacher complexity of $\mathcal{F}_{\phi^*}$ satisfies

$$\left| \mathcal{R}_m(\mathcal{F}_{\phi^*}) - \sqrt{\frac{9r \|u_1^*\|_2^2}{8m}} \right| \le O \left(\sqrt{\frac{3 \|u_1^*\|_2^2}{2m}} \right). \tag{66}$$

Ignoring the low-order term on the right hand side, we have

$$\mathcal{R}_m(\mathcal{F}_{\phi^*}) \approx \sqrt{\frac{9r}{8m}}\|u_1^*\|_2. \tag{67}$$

Compared to the case of pre-training only on the target task, pre-training on the target task and a different task can lead to a larger Rademacher complexity and thus worse generalization.

B.4. Analysis for Multiple Tasks

Pre-training on Multiple Tasks. Consider multiple different data distributions for contrastive learning, while the target (downstream) task is one of them. That is, suppose we have T different invariant-feature subsets $R_1, \dots, R_T \subseteq [d]$, where $|R_t| = r$ for any $t \in [T]$. Let $R = R_1 \cup \cdots \cup R_T$ be the set of all invariant features, and $U = [d] \setminus R$ be the set of spurious features. Assume all distributions share a public feature set $S := R_1 \cap \cdots \cap R_T$ of size s, and each distribution owns a private disjoint feature set of size $r - s$, which means $\forall t_i, t_j \in [T], t_i \neq t_j$ we have $R_{t_i} \cap R_{t_j} \setminus S = \emptyset$. For $j \in S, z_j \sim \mathcal{N}(0, \sigma_S^2),\ \sigma_S > 0$. For the data distribution $\mathcal{D}_t$ $(t = 1, \dots, T)$, we have for $j \in R_t \setminus S, z_j \sim \mathcal{N}(0, \sigma_R^2),\ \sigma_R > 0$, while $z_{R \setminus R_t} = 0$, and $y = (u_t^*)^\top z_{R_t}$ is the ground-truth label for task t. Contrastive learning is over an uneven mixture of unlabeled data from the T distributions, where R_t has weights $w_t > 0$ and $\sum_{t=1}^T w_t = 1$. The target downstream task is $\mathcal{D}_1$. Without loss of generality, assume $S = \{1, , 2, \dots, s\}$ and $\forall t \in [T],\ R_t = S \cup \{s + (t-1)(r-s) + 1, s + (t-1)(r-s) + 2, \dots, s + t(r-s)\}$ and $R = \{1, 2, \dots, s + T(r-s)\}$.

Proposition B.2 (General Version of Proposition 2.1). *There exists $\alpha \in [0, 1]$, $\beta = \min\left(1, \frac{r - \alpha s}{T(r-s)}\right) \in [1/T, 1]$ such that the following holds. $\phi^*(x) = W^* x$ is an optimal representation for the loss (1) in contrastive learning with any W^* of the form:*

$$W^* = [OA^*, \boldsymbol{0}]M^{-1} \tag{68}$$

where $O \in \mathbb{R}^{k \times k}$ is any orthnormal matrices, A^* is a $k \times k$ diagonal matrix with

$$A^*_{jj} = \begin{cases} \sqrt{\alpha} & \text{if } j \in S, \\ \sqrt{\beta} & \text{otherwise,} \end{cases} \tag{69}$$

and the matrix of zeros has size $k \times (d - k)$.

Proof. Following the same argument as in the proof of Proposition 2.1,

$$\begin{aligned} \mathbb{E}_{(x,x^+)}\left[\ell\left(\phi(x)^\top[\phi(x^+) - \mathbb{E}_{x^-}\phi(x^-)]\right)\right] &\geq \sum_{t=1}^T w_t \mathbb{E}_{\{z_j\}}\left[\ell\left(\sum_{j \in R_t} a_{jj}^2 z_j^2\right)\right] & (70) \\ &= \sum_{t=1}^T w_t \mathbb{E}_{\{\tilde{z}_j \sim \mathcal{N}(0,1)\}}\left[\ell\left(\sum_{j \in S} a_{jj}^2 \sigma_S^2 \tilde{z}_j^2 + \sum_{j \in R_t \setminus S} a_{jj}^2 \sigma_R^2 \tilde{z}_j^2\right)\right] & (71) \\ &:= g(\{a_{jj}\}), & (72) \end{aligned}$$

where $\tilde{z}_j$ is a random variable draw from standard Gaussian. Following the same argument as in Lemma B.1, we have the following claim: to achieve the minimum, we can set (1) $a_{\ell\ell}^2 = a_{\ell'\ell'}^2 =: \alpha$ for any $\ell \neq \ell' \in S$, (2) $a_{\ell\ell}^2 = a_{\ell'\ell'}^2 := \alpha_t$ for any $\ell \neq \ell' \in R_t \setminus S, \forall t \in [T]$. Let $Z = \sum_{j \in S} \tilde{z}_j^2$ and $Z_t = \sum_{j \in R_t \setminus S} \tilde{z}_j^2, \forall t \in [T]$. By symmetry of $\tilde{z}_j$'s, all Z_t follow the same distribution. Then

$$\begin{aligned} g(\{a_{jj}\}) &= \sum_{t=1}^T w_t \mathbb{E}\left[\ell\left(\alpha \sigma_S^2 Z + \alpha_t \sigma_R^2 Z_t\right)\right] & (73) \\ &= \sum_{t=1}^T w_t \mathbb{E}\left[\ell\left(\alpha \sigma_S^2 Z + \alpha_t \sigma_R^2 Z_1\right)\right] & (74) \\ &\geq \mathbb{E}\left[\ell\left(\alpha \sigma_S^2 Z + \sigma_R^2 Z_1 \sum_{t=1}^T w_t \alpha_t\right)\right]. & (75) \end{aligned}$$

So the minimum is achieved when $\alpha_{t_i} = \alpha_{t_j} := \beta$ for any $t_i, t_j \in [T]$, leading to

$$g(\{a_{jj}\}) = \mathbb{E}\left[\ell\left(\alpha\sigma_S^2 Z + \beta\sigma_R^2 Z_1\right)\right]. \tag{76}$$

Given the constraint $\alpha s + T\beta(r-s) = \sum_j a_{jj}^2 \leq r$, $0 \leq \alpha, \beta \leq 1$. Following the same argument as in the proof of Proposition 2.1, we finish the proof. □

Given this result we can also analyze the Rademacher complexity of the predictor class similarly as that for the case with two tasks and obtain a result similar to Proposition 2.2.

Similar to the two-task setting, $f(\phi) = u^\top \phi$ with $u = O_{1:k,1:r}(A^*_{1:r,1:r})^{-1}u_1^* + O_{1:k,r+1:k}v$ for any $v \in \mathbb{R}^{k-r}$ satisfies $f(\phi^*(x)) = y = (u_1^*)^\top z_{R_1}$, and $u^* = O_{1:k,1:r}(A^*_{1:r,1:r})^{-1}u_1^*$ is the least norm optimal solution. So the predictor class should be the same as Eqn. (5).

Proposition B.3 (General Version of Proposition 2.2)**.** *Suppose $\alpha > 0$. Let $v_1 = \sum_{j=1}^s (u_{1j}^*)^2$ and $v_2 = \sum_{j=s+1}^r (u_{1j}^*)^2$. Then the Rademacher complexity of $\mathcal{F}_{\phi^*}$ in Eqn. (5) satisfies*

$$\left|\mathcal{R}_m(\mathcal{F}_{\phi^*}) - \sqrt{\frac{1}{m}\left(\frac{1}{\alpha}v_1 + \frac{1}{\beta}v_2\right)(s\alpha\sigma_S^2 + (r-s)\beta\sigma_R^2)}\right| \leq O\left(\frac{\max\{\sigma_S^2, \sigma_R^2\}}{\min\{\sigma_S, \sigma_R\}}\sqrt{\frac{1}{m}\left(\frac{1}{\alpha}v_1 + \frac{1}{\beta}v_2\right)}\right). \tag{77}$$

Proof of Proposition B.3. Following the same argument as in the proof of Proposition 2.2, the Rademacher complexity of $\mathcal{F}_{\phi^*}$ with m samples are

$$\mathcal{R}_m(\mathcal{F}_{\phi^*}) = \frac{\|u^*\|_2}{m}\mathbb{E}_{z_{R_1}^{(i)}}\left[\left\|A^*_{1:r,1:r}\sum_{i=1}^m z_{R_1}^{(i)}\right\|_2\right]. \tag{78}$$

Define $X := A^*_{1:r,1:r}\sum_{i=1}^m z_{R_1}^{(i)}$. Following the same argument as in the proof of Proposition 2.2, we have

$$\|X_j^2 - m\alpha\sigma_S^2\|_{\psi_1} \leq C_2 m\alpha\sigma_S^2, \quad j \in S, \tag{79}$$

$$\|X_j^2 - m\beta\sigma_R^2\|_{\psi_1} \leq C_2 m\beta\sigma_R^2, \quad j \in R \setminus S, \tag{80}$$

where C_2 is an absolute constants and $C_2 > 1$. Let $K = \max(C_2 m\alpha\sigma_S^2, C_2 m\beta\sigma_R^2) \leq C_2 m(\sigma_S^2 + \sigma_R^2)$ and $\mu := m(s\alpha\sigma_S^2 + (r-s)\beta\sigma_R^2)$. By Bernstein's inequality, we have for every $\gamma \geq 0$ that

$$\mathbb{P}\left\{\left|\frac{1}{r}(\|X\|_2^2 - \mu)\right| \geq \gamma\right\} \leq 2\exp\left[-c\min\left(\frac{\gamma^2}{K^2}, \frac{\gamma}{K}\right)r\right] \tag{81}$$

$$\Rightarrow \mathbb{P}\left\{\left|\frac{\|X\|_2^2}{\mu} - 1\right| \geq \frac{r\gamma}{\mu}\right\} \leq 2\exp\left[-\frac{c}{C_2^2}\min\left(\frac{\gamma^2}{m^2(\sigma_S^2 + \sigma_R^2)^2}, \frac{\gamma}{m(\sigma_S^2 + \sigma_R^2)}\right)r\right], \tag{82}$$

where c is an absolute constant. We have $\mu \leq m(\sigma_S^2 + \sigma_R^2)r$. Following the same argument as in the proof of Proposition 2.2, for any $\delta \geq 0$, we have

$$\mathbb{P}\left\{\left|\frac{\|X\|_2}{\sqrt{\mu}} - 1\right| \geq \delta\right\} \leq 2\exp\left[-\frac{c}{C_2^2}\frac{\mu^2}{m^2(\sigma_S^2 + \sigma_R^2)^2 r}\delta^2\right]. \tag{83}$$

Changing variables to $\theta = \delta\sqrt{\mu}$, we obtain the desired sub-gaussian tail

$$\mathbb{P}\left\{|\|X\|_2 - \sqrt{\mu}| \geq \theta\right\} \leq 2\exp\left[-\frac{c}{C_2^2}\frac{\mu}{m^2(\sigma_S^2 + \sigma_R^2)^2 r}\theta^2\right]. \tag{84}$$

By generalization of integral identity, following the same argument as in the proof of Proposition 2.2, we have

$$|\mathbb{E}\left[\|X\|_2 - \sqrt{\mu}\right]| \leq C_3\frac{m(\sigma_S^2 + \sigma_R^2)\sqrt{r}}{\sqrt{\mu}} \tag{85}$$

$$\leq \sqrt{2m}C_3\frac{(\sigma_S^2 + \sigma_R^2)}{\min\{\sigma_S, \sigma_R\}}, \tag{86}$$

where the last inequality is from $\mu = m(s\alpha\sigma_S^2 + (r-s)\beta\sigma_R^2) \geq \min\{\sigma_S^2, \sigma_R^2\} m(s\alpha + (r-s)\beta) \geq \min\{\sigma_S^2, \sigma_R^2\}\frac{1}{2}mr$ and C_3 is an absolute constant. Thus, we have

$$\left|\mathcal{R}_m(\mathcal{F}_{\phi^*}) - \sqrt{\frac{1}{m}\left(\frac{1}{\alpha}v_1 + \frac{1}{\beta}v_2\right) m(s\alpha\sigma_S^2 + (r-s)\beta\sigma_R^2)}\right| = \frac{\|u^*\|_2}{m}\left|\mathbb{E}\left[\|X\|_2 - \sqrt{\mu}\right]\right| \tag{87}$$

$$\leq O\left(\frac{\max\{\sigma_S^2, \sigma_R^2\}}{\min\{\sigma_S, \sigma_R\}}\sqrt{\frac{1}{m}\left(\frac{1}{\alpha}v_1 + \frac{1}{\beta}v_2\right)}\right). \tag{88}$$

□

C. More Experiments Details and Results

C.1. Dataset

CIFAR-10. CIFAR-10 (Krizhevsky et al., 2009) dataset consists of 60,000 32 × 32 color images in 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. Each class has 6,000 images. There are 50,000 training images and 10,000 test images.

CINIC-10. CINIC-10 (Darlow et al., 2018) consists of 32 × 32 color images from both CIFAR and ImageNet and has 90,000 training images with ten classes identical to CIFAR-10.

SVHN. The Street View House Numbers (Netzer et al., 2011) contains 10 digits color images of size 32 × 32 in natural scene. It has 73,257 digits for training and 26,032 digits for testing.

GTSRB. The German Traffic Sign Recognition Benchmark (Stallkamp et al., 2012) is a dataset of color images depicting 43 different traffic signs. The images are not of fixed dimensions and have a rich background and varying light conditions as expected of photographed images of traffic signs. The original training set contains 34,799 images, and the original test set contains 12,630 images. We resize each image to 32×32. The dataset has a significant imbalance in the number of sample occurrences across classes. We use data augmentation techniques to enlarge the training data and balance the number of samples in each class. We construct a class preserving data augmentation pipeline consisting of rotation, translation, and projection transforms and apply this pipeline to the training images until each class contains 2,500 examples. So we construct a new training set containing 107,500 images in total. We also construct a new test set by randomly selecting 10,000 images from the original test set for evaluation.

ImageNet32. ImageNet32 (Deng et al., 2009) is a huge dataset made up of small images called the down-sampled version of ImageNet. ImageNet32 is composed of 1,281,167 training data and 50,000 test data with 1,000 labels.

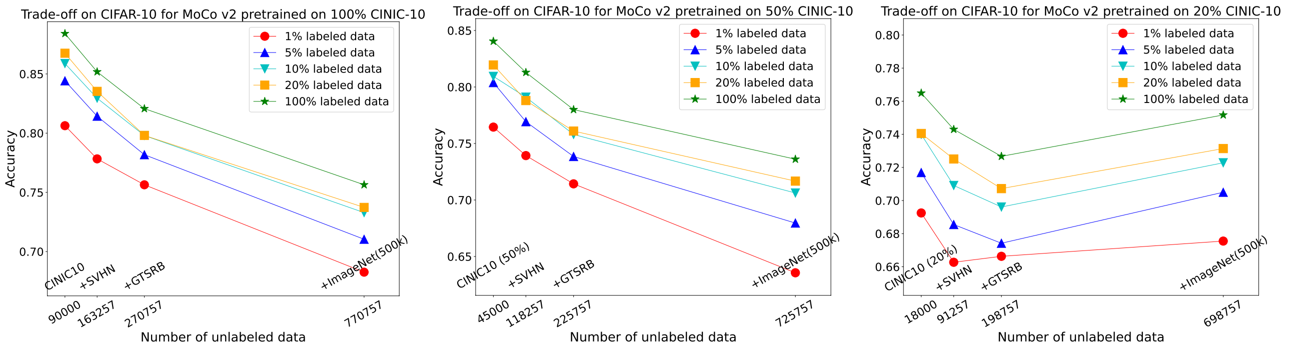
MNIST. The Modified National Institute of Standards and Technology (LeCun et al., 1998) is a database of handwritten gray-scale digits of size 28 × 28. It contains 60,000 training images and 10,000 testing images.

EMNIST. Extended MNIST (Cohen et al., 2017) includes images of handwritten letters and digits. The images in EMNIST were converted into the same size 28 × 28 by the same process as MNIST. EMNIST-Letters has 145,600 lower case characters with 26 balanced classes, and EMNIST-Digits has 280,000 characters with ten balanced classes.

Fashion-MNIST. Fashion-MNIST (Xiao et al., 2017) is a dataset of 28 × 28 gray-scale images with ten classes: T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, ankle boot. The training set size is 60,000, and the test set size is 10,000.

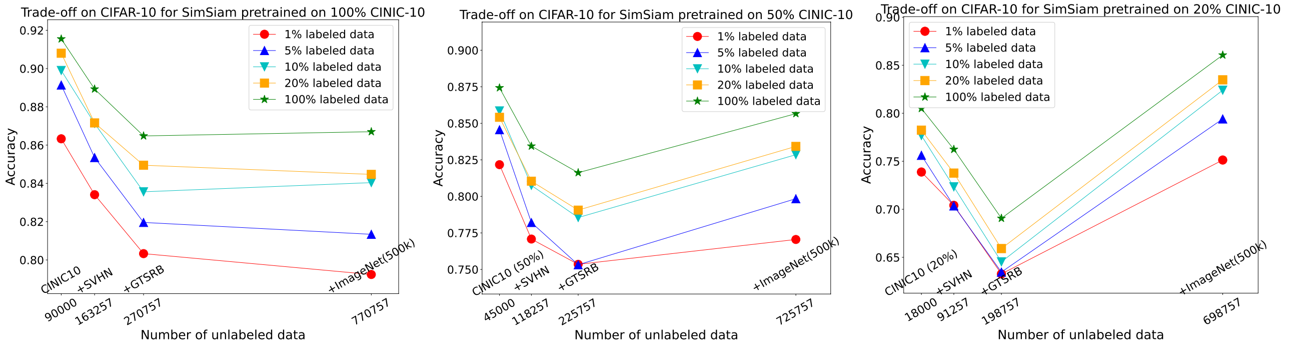
C.2. Ablation Study

Varying Target-Relevant pre-training Data Percentage. In Fig. 3 and 4, we use (a) 100% (b) 50% (c) 20% CINIC-10 to train MoCo v2 and SimSiam, and keep the same setting as Fig. 1. For Fig. 3 (b) with 50% CINIC-10, test accuracy drops, e.g., the test accuracy of 1% CIFAR-10 in Fig. 3 (a) 80.63% vs. (b) 76.45%. We can also see the decreasing curve in Fig. 3 (b). On the other hand, we also have test accuracy drops in Fig. 3 (c) and Fig. 4 (b) (c). However, we can see a U-curve rather than a strictly decreasing curve in Fig. 3 (c) and Fig. 4 (b) (c). ImageNet32 is more relevant with CIFAR-10 than SVHN and GTSRB, consistent with human intuition. When we have a small partition of CINIC-10 which does not cover all target relevant features, the feature extractor can learn these missing features from ImageNet32. Although there are many irrelevant features in ImageNet32, the positive effect is larger than the negative effect, and so it plots a U-curve. It is



(a) (b) (c)

Figure 3. Trade-off on CIFAR-10 for MoCo v2 with varying target relevant pre-training data percentage.



(a) (b) (c)

Figure 4. Trade-off on CIFAR-10 for SimSiam with varying target relevant pre-training data percentage.

consistent with our statement that we need a large and target relevant pre-training dataset rather than a diverse irrelevant one.

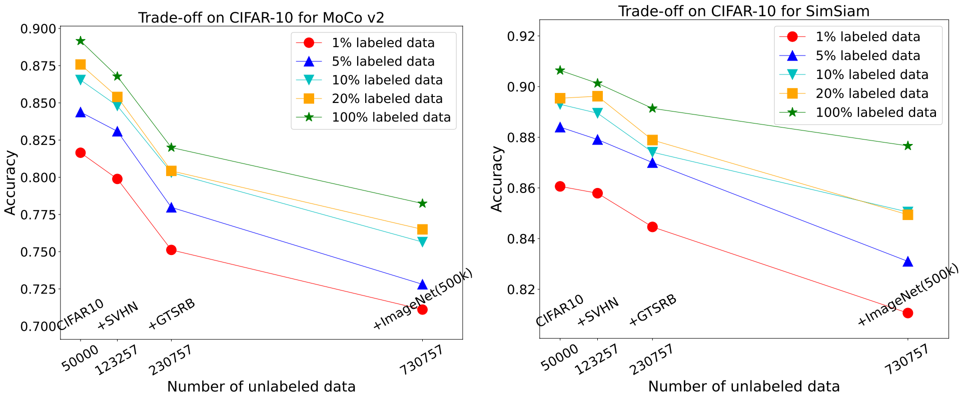


Figure 5. Trade-off on CIFAR-10 for MoCo v2 and SimSiam pre-trianed on CIFAR-10.

Replacing CINIC-10 With CIFAR-10. In Fig. 5, we keep the same setting as Fig. 1 except we replace CINIC-10 with CIFAR-10. Note that our downstream task is still CIFAR-10. In Fig. 5, we can see the same phenomena and similar performance as Fig. 1. Thus, if we have a good choice of a task-relevant pre-training dataset, we can get similar performance as pre-training on the downstream task domain directly.