Blessing of Class Diversity in Pre-training

Yulai Zhao, Jianshu Chen, Simon S. Du

Princeton University, Tencent AI Lab, University of Washington

ABSTRACT

- Pre-training refers to training a model on a few or many tasks to help it learn parameters that can be used in other tasks.
- We present a new statistical analysis aiming to explain the recent superior achievements of the pre-training techniques in natural language processing (NLP).
- We prove that when the classes of the pre-training task (e.g., different words in the masked language model task) are sufficiently diverse, in the sense that the least singular value of the last linear layer in pre-training is large, then pre-training can significantly improve the sample efficiency of downstream tasks.
- Our proof relies on a vector-form Rademacher complexity chain rule for disassembling composite function classes and a modified self-concordance condition. These techniques can be of independent interest.

PROBLEM SETUP

This work is in line with previous transfer learning theories (Du et al., 2021; Tripuraneni et al., 2020) that first pre-train on a large corpus to get a good representation, which, could be future utilized by various downstream tasks. Formally, the procedure is divided into two stages: the pre-training stage to find a representation function and the downstream training stage to obtain a predictor for the downstream task.

◆ In the first stage, we have one pre-training task with n samples, $\{x_i^p, y_i^p\}_{i=1}^n$, where $x_i^p \in X^p \subset R^d$ is the input and $y_i^p \in \{0,1\}^{k-1}$ is the one-hot label for k-class classification.

➢ We aim to obtain a good representation function $\hat{h}$ within a function class $H \subset \{R^d \to R^r\}$ where r is the embedding dimension (often equals to 768, 1024, 2048 in NLP pre-training).

➢ For example, one popular choice of the representation function $\hat{h}$ in NLP applications is the Transformer model and its variants (Vaswani et al., 2017; Devlin et al., 2019). On top of the representation, we predict the labels using function f^p within $F^p \subset \{R^r \to R^{k-1}\}$.

➢ To train the representation function and predictor in pretraining stage, we consider the Empirical Risk Minimization (ERM) procedure

$$\hat{h} = \arg\min_{h\in H} \min_{f^p\in F^p} \hat{R}_p(f^p, h) = \arg\min_{h\in H} \min_{f^p\in F^p} \frac{1}{n} \sum_{i=1} l(f^p \circ h(x_i^p), y_i^p).$$

◆ In the second stage, we assume there are m samples $\{x_i^d, y_i^d\}_{i=1'}^m$ where $x_i^d \in X^d \subset R^d$ is the input and $y_i^d \in \{0,1\}^{k'-1}$ is the one-hot label for k' -class classification.

➢ In most real-world applications, $n \gg m$ and $k \gg k'$, e.g. sentiment analysis.

➢ For the downstream task, we fix the representation function learned from the pre-training task and train the task-dependent predictor within $F^d \subset \{R^r \to R^{k'-1}\}$.

$$\hat{f}^d = \arg\min_{f^p\in F^p} \hat{R}_d(f^d, \hat{h}) = \arg\min_{f^d\in F^d} \frac{1}{m} \sum_{i=1} l(f^d \circ \hat{h}(x_i^d), y_i^d).$$

◆ To this end, we finally obtain a predictor and a representation. We use the following risk to measure the performance of predictor and representation

$$\mathrm{E}_{x^d,y^d}\, l(\hat{f}^d \circ \hat{h}(x^d), y^d) - \mathrm{E}_{x^d,y^d}\, l(g^d(x^d), y^d),$$

where g^d is the optimal predictor for the downstream.

We call this term **Transfer Learning Risk**, which serves as the main objective to be minimized during the optimization process.

Assumptions and Definitions

➢ Throughout the paper, we make the following realizability assumption, which is also a standard assumption in the classical PAC learning

Assumption 3.1 (Realizability). There exist $h \in \mathcal{H}$, $f^{\mathrm{p}} \in \mathcal{F}^{\mathrm{p}}$, $f^{\mathrm{d}} \in \mathcal{F}^{\mathrm{d}}$ such that $g^{\mathrm{p}} = f^{\mathrm{p}} \circ h$ and $g^{\mathrm{d}} = f^{\mathrm{d}} \circ h$.

➢ We make the following assumption on both pre-training and downstream tasks to describe how the underlying data are generated

Assumption 3.2 (Multinomial Logistic Data). For a K-class classification task with q samples, $\{x_i, y_i\}_{i=1}^q$, where $x_i \in \mathcal{X}$ is the input and $y_i \in \{0,1\}^{K-1}$ is the one-hot label. Let f and h be the true underlying predictor layer and representation function. Then the output is $f \circ h(x) \in \mathbb{R}^{K-1}$. Assume each label $\{y\}_i$ is generated from a conditional distribution of a multinomial logistic regression model: $y \sim \mathcal{P}(\cdot|f \circ h(x))$,

$$\mathcal{P}(y|f \circ h(x)) = e^{y^\top f\circ h(x) - \Phi(f\circ h(x))}$$

where $\Phi(x) = \log(1 + \sum_{s=1}^{K-1} e^{x_s})$, $x \in \mathbb{R}^{K-1}$ and y is an one-hot label.

➢ To measure the "closeness" between the learned representation and true underlying feature representation, we use the following metric, following Tripuraneni et al. (2020)

Definition 3.4. Let $h \in \mathcal{H}$ be the optimal representation function and $h' \in \mathcal{H}$ be any representation function. Let $f^{\mathrm{p}} \in \mathcal{F}^{\mathrm{p}}$ be the optimal pre-training predictor on top of h. The pre-training representation difference is defined as:

$$d_{\mathcal{F}^{\mathrm{p}}, f^{\mathrm{p}}}(h'; h) = \inf_{f'\in\mathcal{F}^{\mathrm{p}}} \mathop{\mathbb{E}}_{x^{\mathrm{p}},y^{\mathrm{p}}} [\ell(f' \circ h'(x^{\mathrm{p}}), y^{\mathrm{p}}) - \ell(f^{\mathrm{p}} \circ h(x^{\mathrm{p}}), y^{\mathrm{p}})]$$

where the expectation is over the pre-training data distribution.

➢ For transfer learning, we also need to introduce a similar concept on the downstream task.

$$d_{\mathcal{F}^{\mathrm{d}}}(h'; h) = \sup_{f^{\mathrm{d}}\in\mathcal{F}^{\mathrm{d}}} \inf_{f'\in\mathcal{F}^{\mathrm{d}}} \mathop{\mathbb{E}}_{x^{\mathrm{d}},y^{\mathrm{d}}} [\ell(f' \circ h'(x^{\mathrm{d}}), y^{\mathrm{d}}) - \ell(f^{\mathrm{d}} \circ h(x^{\mathrm{d}}), y^{\mathrm{d}})]$$

➢ Finally, we introduce the key notion of diversity, which measures how well a learned representation, say h', from the pre-training task can be transferred

Definition 3.6. Let $h \in \mathcal{H}$ be the optimal representation function. Let $f^{\mathrm{p}} \in \mathcal{F}^{\mathrm{p}}$ be the optimal pre-training predictor on top of h. The **diversity parameter** $\nu > 0$ is the largest constant that satisfies

$$d_{\mathcal{F}^{\mathrm{d}}}(h'; h) \le \frac{d_{\mathcal{F}^{\mathrm{p}}, f^{\mathrm{p}}}(h'; h)}{\nu}, \forall h' \in \mathcal{H}. \quad (1)$$

One of our key technical contribution is to show **the least singular value of the last linear layer serves as a lower bound of the diversity**, when predictors are linear.

RESULTS

◆ First, we present our generic end-to-end transfer learning guarantee for multi-class transfer learning problems. We only impose the following mild regularity assumptions to ensure the bounds are general.

Assumption 4.1 (Regularity Conditions)

1. In pre-training, loss l is B^p-bounded and $l(\cdot, y)$ is L^p-Lipschitz for any y.
2. In downstream task, loss l is B^d-bounded and $l(\cdot, y)$ is L^d-Lipschitz for any y.
3. Any pre-training predictor $f \in F^p$ is $L(F^p)$-Lipschitz.
4. Bounded predictors: $||f \circ h(x)|| \le D_{X^p}, \forall x \in X^p, h \in H, f \in F^p$. Similarly, the following holds: $||f \circ h(x)|| \le D_{X^d}, \forall x \in X^d, h \in H, f \in F^d$

Under the assumptions, for a given fixed failure probability δ, with probability at least $1-\delta$, we have the **Transfer Learning Risk** bounded by

$$O\left(\frac{1}{\nu}\left\{L^{\mathrm{p}}\left[\log(n)[L(\mathcal{F}^{\mathrm{p}})G_n(\mathcal{H}) + \bar{G}_n(\mathcal{F}^{\mathrm{p}})] + \frac{\sqrt{k}D_{\mathcal{X}^{\mathrm{p}}}}{n^2}\right] + B^{\mathrm{p}}\sqrt{\frac{\log(1/\delta)}{n}}\right\} + L^{\mathrm{d}}\bar{G}_m(\mathcal{F}^{\mathrm{d}}) + B^{\mathrm{d}}\sqrt{\frac{\log(1/\delta)}{m}}\right).$$

◆ We then proceed to the setting that is of most interest to NLP pre-training, where the loss functions are cross-entropy and the F^p, F^d are sets of linear functions. Assume we have a necessary assumption that enables us to relate the diversity parameter with concrete quantity in the network

Assumption 4.3 (Lower Bounded Least Eigenvalue). Let the optimal linear predictor at the last layer for pre-training be $\alpha^{\mathrm{p}} \in \mathbb{R}^{r\times(k-1)}$, $\tilde{\nu} \triangleq \sigma_r(\alpha^{\mathrm{p}}(\alpha^{\mathrm{p}})^\top) > 0$ where σ_r is the r-biggest eigenvalue.

Intuitively, this assumption ensures that the pre-training task matrix spans the entire r-dimensional space and thus covers the output of the optimal representation $h(\cdot) \in R^r$.

Under the assumptions, for a given fixed failure probability δ, with probability at least $1-\delta$, we have the **Transfer Learning Risk** bounded by

$$O\left(\frac{1}{\tilde{\nu}}\left\{\sqrt{k}\left[\log(n)[\sqrt{k}G_n(\mathcal{H}) + \bar{G}_n(\mathcal{F}^{\mathrm{p}})] + \frac{\sqrt{k}D_{\mathcal{X}^{\mathrm{p}}}}{n^2}\right] + D_{\mathcal{X}^{\mathrm{p}}}\sqrt{\frac{\log(1/\delta)}{n}}\right\} + \sqrt{k'}\mathop{\mathbb{E}}_{\mathcal{X}^{\mathrm{d}}}\hat{G}_m(\mathcal{F}^{\mathrm{d}}|\hat{h}\circ x^{\mathrm{d}}) + \sigma\sqrt{\frac{\log(1/\delta)}{m}} + D_{\mathcal{X}^{\mathrm{d}}}\sqrt{\frac{\log(1/\delta)}{m}}\right)$$

◆ To get a better impression of this bound, we instantiate the quantities in a specific setting, in which not only the predictors, but the underlying representation is also linear, i.e., $h(x) = B^\top x, B \in R^{d\times r}$.

Under several assumptions, for a given fixed failure probability δ, with probability at least $1-\delta$, we have the **Transfer Learning Risk** bounded by

$$O\left(\frac{1}{\tilde{\nu}}\left[\sqrt{k}\log(n)\left(\sqrt{\frac{kdr^2}{n}} + k\sqrt{\frac{r}{n}}\right) + \frac{k}{n^2} + \sqrt{\frac{\log(1/\delta)}{n}}\right] + (k')^{\frac{3}{2}}\sqrt{\frac{r}{m}} + k'\sqrt{\frac{\log(1/\delta)}{m}}\right)$$

To interpret this bound, consider the practically relevant scenario where k' = O(1) (e.g., sentiment analysis), $m, k \ll n, \mathrm{r} \ll d$, and in the benign case $\tilde{\nu} = \Omega(k)$, the risk simplifies to $\tilde{O}\left(\sqrt{\frac{dr^2}{n}} + \sqrt{\frac{r}{m}}\right)$, the first term accounts for using all pre-training data to learn the representation function and the second term accounts for using the downstream data to learn the last linear layer. This is significantly better than not using pre-training, in which case the risk scales as $O\left(\sqrt{\frac{d}{m}}\right)$. The improvement showcases the power of pre-training.

CONCLUSIONS

- We formally prove the benefit of multi-class pre-training using the notion of class diversity.
- Our proof uses the vector-form Rademacher complexity chain rule and a modified self-concordance condition.

Future Work

☐ Our work is based on realizability assumptions (cf. Assumption 3.1 and 3.2) that are commonly adopted in transfer learning and classical PAC learning framework. We believe our theorems can be extended to agnostic versions by relaxing these assumptions.

☐ If the target task is well-aligned with the source tasks, one can define more fine-grained notions to capture the task relevance.

☐ Based on the techniques presented by this work, develop theories that could explain some more recent pre-training algorithm showing that one can do pre-training with the downstream dataset itself, and still get good results. Direct application of the transfer learning scheme is precluded in such setting.

REFERENCES

- S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-Shot Learning via Learning the Representation, Provably. In International Conference on Learning Representations, 2021.
- N. Tripuraneni, M. Jordan, and C. Jin. On the theory of transfer learning: The importance of task diversity. Advances in Neural Information Processing Systems, 2020
- A. Maurer. A vector-contraction inequality for Rademacher complexities. In International Conference on Algorithmic Learning Theory, 2016.
- A. Maurer, M. Pontil, and B. Romera-Paredes. The Benefit of Multitask Representation Learning. Journal of Machine Learning Research, 2016
- F. Bach. Self-concordant analysis for logistic regression. Electronic Journal of Statistics, 2010
- J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning. Advances in Neural Information Processing Systems, 2021

ACKNOWLEDGEMENTS

This work was supported in part by NSF CCF 2212261, NSF IIS 2143493, NSF DMS-2134106, NSF CCF 2019844, NSF IIS 2110170, and a gift funding from Tencent.