Blessing of Class Diversity in Pre-training

Yulai Zhao	Jianshu Chen	Simon Du
Princeton University	Tencent AI Lab	UW

Backgrounds: pre-training

- 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. Pre-training techniques revolutionized NLP, with dramatic improvements for various downstream tasks
- Basically under the scheme of **transfer learning**
- Commonly used in various domains, including natural language processing (NLP) and computer vision
- Example: Masked Language Model (MLM) pre-training task

Backgrounds: pre-training

Benefits

✓Pre-training techniques revolutionized NLP

✓Dramatic improvements for various downstream tasks

✓Better performance on downstream tasks with limited data

✓Utilizes knowledge learned from related tasks

Challenges

- ☐ Computationally expensive and time-consuming
- ☐ Choice of pre-training tasks and data? What is needed the most?
- ☐ Theoretical foundations?

Backgrounds: multi-task learning

A common viewpoint to study pre-training is through the language of multi-task learning.

- Past works [Caruana, 1997, Baxter, 2000, Maurer et al., 2016, Du et al., 2021, Tripuraneni et al., 2020a,b, Thekumparampil et al., 2021] studied multi-task training. A notion ***diversity of tasks*** is shown to be crucial to allow pre-trained model to be useful for downstream tasks.

- Informally the results say: **we should make pre-training tasks more diverse**, then the **worst-case** transfer risk (for downstream task) is controlled when the pre-training representation difference is small.
- However, often require **a large number** of diverse tasks

Backgrounds: multi-task learning

- Unfortunately, this line of theory cannot be used to explain the success of pre-training in NLP since they require **a large number** of diverse tasks, whereas BERT is pretrained on **very few** tasks. To this end, *diversity of tasks* is invalid.

- An example is the BERT which only has two types of pre-training tasks： Masked Language Model and Next Sentence Prediction), whereas it is also required by multi-task learning literature that the number of tasks is **comparable** with the dimension of representation.

Motivation

Can we go beyond multi-task training and develop a theory explaining the success of pre-training that has very few tasks?

- As mentioned, multi-task theories do not apply to the setting.

Problem Setup

- Follow transfer learning notations, divide the procedure into two phases: pre-training and downstream learning. We assume all tasks are classification problems
  1. Train the representation function and prediction function in pre-training

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

✓ This is the standard Empirical Risk Minimization (ERM) procedure.

✓ MLM is supervised learning. x: data, y:label

✓ H is the representation class, i.e., neural networks or transformers in practice

✓ F is the predictor

✓ l is the loss function. For pre-training, we study cross entropy with k classes, ~30K with BPE sub-word units

Problem Setup

- Follow transfer learning notations, divide the procedure into two phases
  2. Fix representation $\hat{h}$ and train the classifier for the downstream task

$$\hat{f}^{down} = arg \min_{f^{down} \in F^{down}} \frac{1}{m} \sum_{i=1 \cdots m} l(f^{down} \circ \hat{h}(x_i^{down}), y_i^{down})$$

Remark

✓Downstream task is often data-limited, so we have n >> m

✓l is the loss function. Assume downstream task has k' classes

✓In sentiment analysis, k'= 2("positive" or "negative") so k >> k'

Problem Setup

- Final metric: obtain a small Transfer Learning Risk

$$E\left[l\left(\hat{f}^{down} \circ \hat{h}\left(x^{down}\right), y^{down}\right)\right] - \min E\left[l\left(f^{down} \circ h\left(x^{down}\right), y^{down}\right)\right]$$

■ Here the expectation is over downstream task data generation.

■ Such kind of PAC metric measures the gap between the representation and predictor we learned with the 'optimal' representation and predictor in classes.

■ A common practice in theories is assuming the true underlying functions are truly captured by the adopted function classes — The realizability assumption

Preliminaries

- 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:

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

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

Intuitively, the quantity demonstrates the performance difference between the optimal predictor and the best possible predictor given a representation h’ .

Preliminaries

- Downstream learning also requires a similar concept

Definition 3.5. Let $h \in \mathcal{H}$ be the optimal representation function and $h' \in \mathcal{H}$ be any representation function. For the downstream task, for a function class $\mathcal{F}^{\mathrm{d}}$, let $f^{\mathrm{d}} \in \mathcal{F}^{\mathrm{d}}$ be the optimal pre-training predictor on top of a specific h. We define the worst-case representation difference between h and $h' \in \mathcal{H}$ as:

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

where the expectation is over the data distribution of the downstream task. Here, the supremum is taken over $\{f^d | f^{\mathrm{d}} \in \mathcal{F}^{\mathrm{d}}, f^d \text{ is the optimal predictor on } h \in \mathcal{H}\}$.

Intuitively, the quantity demonstrates the performance difference between the optimal predictor and the best possible predictor given a representation h’ .

Preliminaries

- We now introduce the key notion of *diversity of classes*, which measures how well a learned representation, say h’, from the pre-training task can be transferred to the downstream task.

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

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

- To interpret, diversity parameter ν is a task-relatedness parameter.
- We note that these definitions are **naturally** defined from inspecting the pre-training procedure. BUT deriving their values statistically is **not natural**.
- In particular, one of our key technical challenge is to show: when last layer classifiers are linear, the least singular value of the linear param would serve as a lower bound of ν.

Main Theorem: generic guarantees for classification pre-training

- Theorem: Under standard regularity conditions, we prove the upper bound of transfer learning risk is characterized by ν (diversity parameter), Lipschitz parameters, and model complexities.

Assumption 1 (Realizability). *There exist $h \in \mathcal{H}$, $f^{\text{pre}} \in \mathcal{F}^{\text{pre}}$, $f^{\text{down}} \in \mathcal{F}^{\text{down}}$ such that $g^{\text{pre}} = f^{\text{pre}} \circ h$ and $g^{\text{down}} = f^{\text{down}} \circ h$.*

Assumption 2 (Regularity conditions). *We assume the following regularity conditions hold:*

- *In pre-training, $\ell(\cdot,\cdot)$ is B^{pre}-bounded, and $\ell(\cdot, y)$ is L^{pre}-Lipschitz for all y.*
- *In downstream task, $\ell(\cdot, y)$ is B^{down}-bounded and L^{down}-Lipschitz for all y.*
- *Any predictor $f \in \mathcal{F}^{\text{pre}}$ is $L(\mathcal{F}^{\text{pre}})$-Lipschitz with respect to the Euclidean distance.*
- *Predictors are bounded: $\|f \circ h(x)\| \leq D_{\mathcal{X}^{\text{pre}}}$ for any $x \in \mathcal{X}^{\text{pre}}, h \in \mathcal{H}, f \in \mathcal{F}^{\text{pre}}$. Similarly $\|f \circ h(x)\| \leq D_{\mathcal{X}^{\text{down}}}$ for any $x \in \mathcal{X}^{\text{down}}, h \in \mathcal{H}, f \in \mathcal{F}^{\text{down}}$.*

Theorem 1. *Under Assumption 1 and 2, for a given fixed failure probability δ, with probability at least $1-\delta$ we have the Transfer Learning Risk upper bounded by:*

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

- Note this is exactly the desired theoretical guarantee because 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.

A Special Setting: linear predictors and representations

- Theorem (Informal) Let H, f^{pre}, f^{down} be linear mappings. Assume several regularity conditions, then diversity parameter is lower bounded

$$\nu = \Omega(\tilde{\nu}\ ),$$

where $\tilde{\nu} = \sigma_r\left(\alpha^{pre}(\alpha^{\mathrm{pre}})^{\mathrm{T}}\right)$ is the singular value of the linear parameter.

In the benign case where $\tilde{\nu} = \Omega(k)$, transfer learning risk is bounded by $O\left(\sqrt{\frac{d\,r^2}{n}} + \sqrt{\frac{r}{m}}\right)$.

Remark

- r and d are dimensions of representation and raw input, so $r < d$.
- Also note downstream task has much fewer data, so we have n >> m
- This is significantly better than not using pre-training, where the risk scales $O\left(\sqrt{\frac{d}{m}}\right)$.
- Therefore we showcase the power of using pre-training.

Contributions

1. New notion of diversity of classes

2. Provable improvements in statistical efficiency for downstream tasks

3. Our proof uses a vector-form Rademacher complexity chain rule and a modified self-concordance condition, both could be of independent interests

4. First set of theoretical results for standard NLP pre-training without strong conditions

Future work

1. Lower and Upper bounds, how to improve?

2. Pre-training with the downstream itself (Krishna et al., 2022), not transfer learning anymore, how to justify?

3. How to apply theoretical findings to practice?

Thank you and some more information

POSTER 105

PAPER

SCAN ME

Yulai Zhao
PhD student @ Princeton