CS 839: Theoretical Foundations of Deep Learning **Spring 2023**

Lecture 18 Trade-off of Efficiency and Universality in Contrastive Learning

Instructor: Yingyu Liang *Date:* *Scriber: Yingyu Liang*

To analyze the trade-off, we compare two scenarios:

- Specific representation: the representation is pre-trained only on the unlabeled data from one target task.
- Universal representation: the representation is pre-trained on the mixture of several tasks.

For both representations, we would like to evaluate two performance criteria as follows

- Label efficiency on the target task: measured by the prediction accuracy (generalization) of learning a classifier on top of the representation, using labeled data from the target task.
- Universality: measure by the average accuracy on all the tasks, i.e., for each task, learn a classifier on top of the representation using labeled data from the task, evaluate the accuracy, and then take the average over all tasks.

We will need to specify the data distribution in each task, how the representation is learned via contrastive learning, and how the evaluation is done.

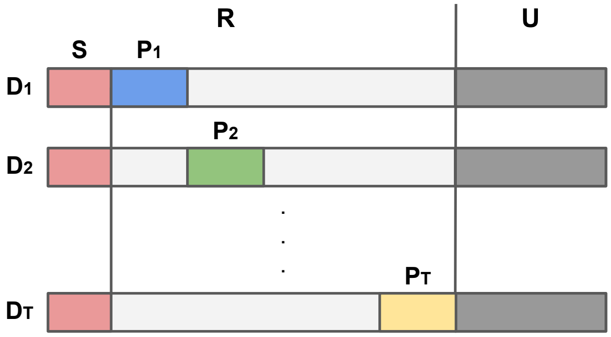

Figure 1: Illustration of the features in our data distributions.

1 Problem Setup

Data distributions. Suppose we have T tasks, with data distributions $\mathcal{D}_t(t = 1, \dots, T)$. The data distribution is specified by first sampling a hidden representation $z \in \mathbb{R}^d$, and then generating the input $x \in \mathbb{R}^d$ and the label $y \in \mathbb{R}$ by linear functions on z.

All tasks share a public feature S of size s, and each task $\mathcal{D}_t$ additionally owns a private disjoint feature set P_t of size $r - s$, i.e., $P_t \cap S = \emptyset$ and $P_{t_1} \cap P_{t_2} = \emptyset$ for $t_1 \neq t_2$ (Fig. 1). Then

$R_t = S \cup P_t$ are called the invariant features for $\mathcal{D}_t$. All invariant features are $R = \cup_{t=1}^T R_t$. The other features are called spurious features $U = [d] \setminus R$. For simplicity, consider the case $s = 1$ and $r = 2$, i.e., there is only one shared feature, and each task has only one private feature.

In task $\mathcal{D}_t$, the (x, x^+) are generated as follows:

$$z_{R_t} \sim \mathcal{N}(0, I), \; z_{R \setminus R_t} = 0, \; z_U \sim \mathcal{N}(0, I), z = [z_R; z_U], \quad x = g(z), \tag{1}$$

$$z_U^+ \sim \mathcal{N}(0, I), z^+ = [z_R; z_U^+], \quad x^+ = g(z^+), \tag{2}$$

and x^- is simply an i.i.d. copy from the same distribution as x.

The generating function g is a linear function, i.e., $x = g(z) = Mz$ where $M \in \mathbb{R}^{d \times d}$ is an orthonormal dictionary. Since linear probing has strong performance on pre-trained representations, we thus assume that the label in each task t is linear in its invariant features $y = (u_t^*)^\top z_{R_t}$ for some $u_t^* \in \mathbb{R}^r$.

Constrative learning. In practice, multiple independent negative examples are used, and thus we consider the following contrastive loss

$$\min_{\phi \in \Phi} \; \mathbb{E}_{(x,x^+)} \left[\ell \left(\phi(x)^\top (\phi(x^+) - \mathbb{E}_{x^-} \phi(x^-)) \right) \right] \tag{3}$$

for a convex and decreasing $\ell(t)$ to pre-train a representation ϕ.

Assume the representations are linear functions with weights of bounded spectral/Frobenius norms:

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

Here the norm bounds are chosen to be the minimum values to allow recovering the invariant features in the target task, i.e., there exists $\phi \in \Phi$ such that $\phi(x) = [z_{R_t}; \mathbf{0}]$.

Evaluation. Then, when using ϕ for prediction in the target task $\mathcal{D}_t$, the predictor class should contain a predictor matching the ground-truth label:

$$\mathcal{F}_{\phi,t} = \{f(z) = u^\top z : u \in \mathbb{R}^k, \|u\| \leq B_{\phi,t}\} \tag{4}$$

where $B_{\phi,t}$ is the minimum value such that there exists $u_t \in \mathcal{F}_{\phi,t}$ with $y = u_t^\top \phi(x)$ on $\mathcal{D}_t$.

2 Analysis of the Representation Learned

First, we know the following useful property of the constrative loss.

For each $\mathcal{D}_t$,

$$\mathbb{E}_{(x,x^+)} \left[\ell \left(\phi(x)^\top [\phi(x^+) - \mathbb{E}_{x^-} \phi(x^-)] \right) \right] \tag{5}$$

$$= \mathbb{E}_{(z,z^+)} \left[\ell \left((WMz)^\top (WMz^+ - \mathbb{E}_{z^-}[WMz^-]) \right) \right] \tag{6}$$

$$= \mathbb{E}_{(z,z^+)} \left[\ell \left(z^\top (M^\top W^\top WM)(z^+ - \mathbb{E}_{z^-}[z^-]) \right) \right] \tag{7}$$

$$\geq \mathbb{E}_{z_R} \left[\ell \left((\mathbb{E}_{z_U}[z])^\top M^\top W^\top WM (\mathbb{E}_{z_U^+}[z^+] - \mathbb{E}_{z^-}[z^-]) \right) \right] \tag{8}$$

$$= \mathbb{E}_{z_R} \left[\ell \left([z_R; \mathbf{0}]^\top M^\top W^\top WM([z_R; \mathbf{0}] - 0) \right) \right] \tag{9}$$

$$= \mathbb{E}_{z_R} \left[\ell \left(\|WM[z_R; \mathbf{0}]\|^2 \right) \right] \tag{10}$$

where the inequality comes from the convexity of $\ell(t)$ and Jensen's inequality. The equality holds if and only if WMz does not depend on z_U and WMz^+ does not depend on z_U^+, so the optimal solution should satisfy this condition.

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_S, and z_{P_t}, without loss of generality, we can assume $A_R = QA$ where A is a diagonal matrix with diagonal entries a_{jj}'s and Q is any orthonormal matrix. Furthermore, $A_U = 0$ in the optimal solution since it does not affect the loss but only decreases the norm bound on A_R. So on data from the task $\mathcal{D}_t$,

$$\mathbb{E}_{\mathcal{D}_t}\left[\ell\left(\|WM[z_R; \mathbf{0}]\|^2\right)\right] = \mathbb{E}_{z_{R_t}}\left[\ell\left(\sum_{j\in R_t} a_{jj}^2 z_j^2\right)\right]. \tag{11}$$

Proposition 1. The representation ϕ^* obtained on an even mixture of data from all the tasks $\{\mathcal{D}_t : 1 \le t \le T\}$ satisfies $\phi^* \circ g(z) = Q\left(\sum_{j\in S} z_j e_j + \sum_{j\in R\setminus S}\sqrt{\frac{1}{T}} z_j e_j\right)$, where e_j's are the basis vectors and Q is any orthonormal matrix.

Proof. On the mixture,

$$\begin{aligned}
&\mathbb{E}_{(x,x^+)}\left[\ell\left(\phi(x)^\top[\phi(x^+) - \mathbb{E}_{x^-}\phi(x^-)]\right)\right] && (12)\\
\ge&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\{z_j\}}\left[\ell\left(\sum_{j\in R_t} a_{jj}^2 z_j^2\right)\right] && (13)\\
=&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\{z_j\}}\left[\ell\left(\sum_{j\in S} a_{jj}^2 z_j^2 + \sum_{j\in P_t} a_{jj}^2 z_j^2\right)\right] && (14)\\
:=&g(\{a_{jj}\}). && (15)
\end{aligned}$$

Let $Z = \sum_{j\in S} z_j^2, Z_t = \sum_{j\in P_t} z_j^2$. Let α denote the coefficient α_{jj}^2 for $j \in S$, and α_t denote that α_{jj}^2 for $j \in P_t$. Then $Z \sim \chi_s^2 = \chi_1^2$ and $Z_t \sim \chi_{r-s}^2 = \chi_1^2$, and we have:

$$\begin{aligned}
g(\{a_{jj}\}) &= g(\{\alpha, \alpha_t\}) && (16)\\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\ell\left(\alpha Z + \alpha_t Z_t\right)\right] && (17)\\
&\ge \mathbb{E}\left[\ell\left(\alpha Z + Z_1\sum_{t=1}^{T}\frac{1}{T}\alpha_t\right)\right]. && (18)
\end{aligned}$$

The inequality comes from the convexity of $\ell(t)$ and Jensen's inequality. So the minimum is achieved when $\alpha_t := \beta$ for any $t \in [T]$, leading to

$$g(\{\alpha, \alpha_t\}) = \mathbb{E}\left[\ell\left(\alpha Z + \beta Z_1\right)\right] \tag{19}$$

subject to the constraints $\alpha s + T\beta(r - s) = \alpha + T\beta \le r$, $0 \le \alpha, \beta \le 1$. Then we get $\phi^* \circ g(z) = W^*Mz = Q\left(\sum_{j\in S}\sqrt{\alpha} z_j e_j + \sum_{j\in R\setminus S}\sqrt{\beta} z_j e_j\right)$ for some $\alpha = 1,\ \beta = \frac{1}{T}$, where e_j's are the basis vectors and Q is any orthonormal matrix. □

Proposition 2. The representation ϕ_t^* obtained on data from $\mathcal{D}_t$ satisfies $\phi_t^* \circ g(z) = Q\left(\sum_{j \in R_t} z_j e_j\right)$ where e_j's are the basis vectors and Q is any orthonormal matrix.

Proof. With only one shared and only one private feature, following a similar argument as above, we get $\phi_t^* \circ g(z) = Q\left(\sum_{j \in R_t} z_j e_j\right)$, where e_j's are the basis vectors and Q is any orthonormal matrix. □

We can see that universal representation encodes all useful features but down-weights the private feature for the target task, and thus leads to the trade-off.

The analysis for more general cases can be seen in [1].

References

[1] Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha. The trade-off between universality and label efficiency of representations from contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.