Generalization Analysis for Supervised Contrastive Representation Learning under Non-IID Settings

Nong Minh Hieu, Antoine Ledent

School of Computing and Information Systems, SMU Singapore

Contrastive learning overview

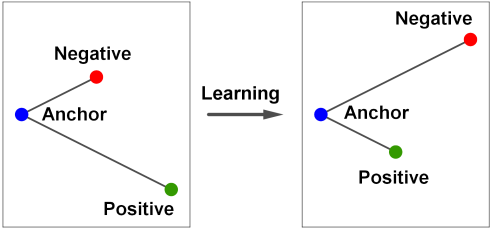

Figure: Visual illustration of contrastive learning

Contrastive Learning Overview

Problem Settings: Let $\mathcal{X}$ be the input space and $\mathcal{C}$ be the finite set of labels. Furthermore, denote

- ρ as the class distribution over $\mathcal{C}$.
- $\mathcal{D}_c$ as the class-conditional distribution given label $c \in \mathcal{C}$.
- $\bar{\mathcal{D}}_c$ as the class-conditional distribution given labels $\mathcal{C} \setminus \{c\}$.

Definition (Unsupervised Risk): Given a contrastive loss $\ell : \mathbb{R}^k \to [0, \mathcal{M}]$, the unsupervised risk is defined as

$$\mathrm{L_{un}}(f) = \mathop{\mathbb{E}}_{c \sim \rho} \mathop{\mathbb{E}}_{\substack{\mathbf{x}, \mathbf{x}^+ \sim \mathcal{D}_c^2 \\ \mathbf{x}_{1:k}^- \sim \bar{\mathcal{D}}_c^k}} \left[\ell\left(\left\{ f(\mathbf{x})^\top \left[f(\mathbf{x}^+) - f(\mathbf{x}_i^-) \right] \right\}_{i=1}^k \right) \right]. \tag{1}$$

Contrastive Learning Overview

We are interested in the excess risk $\mathrm{ER_{un}}(\widehat{f}_n) = \mathrm{L_{un}}(\widehat{f}_n) - \inf_{f \in \mathcal{F}} \mathrm{L_{un}}(f)$ where $\widehat{f}_n$ is an empirical risk minimizer:

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} \ell\left(\left\{f(\mathbf{x}_j)^\top \left[f(\mathbf{x}_j^+) - f(\mathbf{x}_{ji}^-)\right]\right\}_{i=1}^{k}\right). \tag{2}$$

Previous works often assume that the tuples $\left\{(\mathbf{x}_j, \mathbf{x}_j^+, \mathbf{x}_{j1:k}^-)\right\}_{j=1}^{n}$ are i.i.d. - A condition rarely met in practice.

Reality: Tuples are sub-sampled from a labeled dataset $\mathcal{S} = \left\{(\mathbf{x}_j, \mathbf{y}_j)\right\}_{j=1}^{N}$.

Contrastive Learning Overview

In this work, we are interested in the excess risk bounds of $\widehat{f}_{\mathrm{sub}} = \arg\min_{f \in \mathcal{F}} \widehat{\mathcal{L}}(f; \mathcal{T}_{\mathrm{sub}})$ where:

$$\widehat{\mathcal{L}}(f; \mathcal{T}_{\mathrm{sub}}) = \frac{1}{n} \sum_{j=1}^{n} \ell\Big(\Big\{ f(\mathbf{z}_j)^\top \Big[f(\mathbf{z}_j^+) - f(\mathbf{z}_{ji}^-) \Big] \Big\}_{i=1}^{k} \Big). \tag{3}$$

Where $\mathcal{T}_{\mathrm{sub}} = \Big\{ (\mathbf{z}_j, \mathbf{z}_j^+, \mathbf{z}_{j1:k}^-) \Big\}_{j=1}^{M}$ and each tuple is selected independently as:

- Select $c \in \mathcal{C}$ with probability $\widehat{\rho}(c) = \frac{N_c^+}{N}$.
- Select $\mathbf{z}_j, \mathbf{z}_j^+$ from class c without replacement.
- Select $\mathbf{z}_{j1}^-, \dots, \mathbf{z}_{jk}^-$ outside of class c without replacement.

U-Statistics Formulation

For each $c \in \mathcal{C}$ let $\mathcal{S}_c$ be the set of data points from class c and $\bar{\mathcal{S}}_c = \mathcal{S} \setminus \mathcal{S}_c$. Then, we define the set of $(k+2)$-tuples $\mathcal{T}_c$ as follows:

$$\mathcal{T}_c = \Big\{(\mathbf{z}, \mathbf{z}^+, \mathbf{z}^-_{1:k}) : \mathbf{z}, \mathbf{z}^+ \in \mathcal{S}_c,\ \mathbf{z}^-_1, \ldots, \mathbf{z}^-_k \in \bar{\mathcal{S}}_c\Big\}. \tag{4}$$

Then, we use the following U-Statistics to estimate $\mathrm{L}_{\mathrm{un}}(f)$:

$$\mathcal{U}_N(f) = \sum_{c \in \mathcal{C}} \frac{N_c^+}{N} \mathcal{U}_N(f|c), \tag{5}$$

$$\mathcal{U}_N(f|c) = \frac{1}{\binom{N_c^+}{2} \times \binom{N-N_c^+}{k}} \sum_{(\mathbf{z},\mathbf{z}^+,\mathbf{z}^-_{1:k}) \in \mathcal{T}_c} \ell\bigg(\Big\{f(\mathbf{z})^\top \Big[f(\mathbf{z}^+) - f(\mathbf{z}^-_i)\Big]\Big\}_{i=1}^{k}\bigg). \tag{6}$$

Contributions

Our contributions are as follows:

- We derive excess risk bounds for $\widehat{f}_{\mathcal{U}} = \arg\min_{f \in \mathcal{F}} \mathcal{U}_N(f)$ and $\widehat{f}_{\text{sub}} = \arg\min_{f \in \mathcal{F}} \widehat{\mathcal{L}}(f; \mathcal{T}_{\text{sub}})$.
- We apply our bounds to linear maps and neural networks:

$$\mathcal{F}_{\text{lin}} = \left\{ \mathbf{x} \mapsto A\mathbf{x} : A \in \mathbb{R}^{d \times m}, \|A^\top\|_{2,1} \leqslant a, \|A\|_\sigma \leqslant s \right\}, \tag{7}$$

$$\mathcal{F}_{\text{nn}} = \mathcal{F}_L \circ \mathcal{F}_{L-1} \circ \cdots \circ \mathcal{F}_1, \tag{8}$$

$$\text{where } \mathcal{F}_l = \left\{ \mathbf{x} \mapsto \varphi_l\left(A^{(l)}\mathbf{x}\right) : A^{(l)} \in \mathbb{R}^{d_{l-1} \times d_l}, \|A^{(l)}\|_\sigma \leqslant s_l \right\}.$$

Assumption: We assume that the contrastive loss $\ell : \mathbb{R}^k \to [0, \mathcal{M}]$ is ℓ^∞-Lipschitz with constant $\eta > 0$.

Main Results

- **Theorem 5.1**: Let $\widehat{f}_{\mathcal{U}} = \arg\min_{f \in \mathcal{F}} \mathcal{U}_N(f)$. For any $\delta \in (0,1)$, with probability of at least $1-\delta$:

$$\mathrm{ER}_{\mathrm{un}}(\widehat{f}_{\mathcal{U}}) \leqslant \mathcal{O}\left(\sum_{c \in \mathcal{C}} \rho(c) \frac{K_{\mathcal{F},c}}{\sqrt{\widetilde{N}}} + \mathcal{M}\sqrt{\frac{\ln |\mathcal{C}|/\delta}{\widetilde{N}}}\right)$$

  where $\widetilde{N} = N \min\left(\frac{\rho_{\min}}{2}, \frac{1-\rho_{\max}}{k}\right)$ and each term $K_{\mathcal{F},c}$ involves the covering number of $\mathcal{F}$.
- **Applications**: Bounds for linear maps and neural networks.

Minimizer	Function Class	Generalization Bound
$\widehat{f}_{\mathcal{U}}$	Linear Maps	$\widetilde{\mathcal{O}}\left(\frac{\eta s a b^2}{\sqrt{\widetilde{N}}} + \mathcal{M}\sqrt{\frac{\ln \lvert\mathcal{C}\rvert/\delta}{\widetilde{N}}}\right)$
$\widehat{f}_{\mathcal{U}}$	Neural Nets	$\widetilde{\mathcal{O}}\left(\frac{\mathcal{M}\mathcal{W}^{\frac{1}{2}}}{\sqrt{\widetilde{N}}} + \mathcal{M}\sqrt{\frac{\ln \lvert\mathcal{C}\rvert/\delta}{\widetilde{N}}}\right)$

Table: Generalization bounds for the U-Statistics minimizer.

Main Results

- **Theorem 5.2**: Let $\widehat{f}_{\mathrm{sub}} = \arg\min_{f\in\mathcal{F}} \widehat{\mathcal{L}}(f;\mathcal{T}_{\mathrm{sub}})$. For any $\delta \in (0,1)$, with probability of at least $1-\delta$:

$$\mathrm{ER}_{\mathrm{un}}(\widehat{f}_{\mathrm{sub}}) \leqslant \mathcal{O}\left(\widehat{\mathfrak{R}}_{\mathcal{T}_{\mathrm{sub}}}(\ell\circ\mathcal{F}) + \sum_{c\in\mathcal{C}} \rho(c)\frac{K_{\mathcal{F},c}}{\sqrt{\widetilde{N}}} + \mathcal{M}\left[\sqrt{\frac{\ln|\mathcal{C}|/\delta}{\widetilde{N}}} + \sqrt{\frac{\ln 1/\delta}{M}}\right]\right)$$

  where $\widetilde{N} = N\min\left(\frac{\rho_{\min}}{2}, \frac{1-\rho_{\max}}{k}\right)$ and each term $K_{\mathcal{F},c}$ involves the covering number of $\mathcal{F}$.
- **Applications**: Bounds for linear maps and neural networks.

Minimizer	Function Class	Generalization Bound
$\widehat{f}_{\mathrm{sub}}$	Linear Maps	$\tilde{\mathcal{O}}\left(\eta sab^2\left[\frac{1}{\sqrt{\widetilde{N}}}+\frac{1}{\sqrt{M}}\right]+\mathcal{M}\left[\sqrt{\frac{\ln|\mathcal{C}|/\delta}{\widetilde{N}}}+\sqrt{\frac{\ln 1/\delta}{M}}\right]\right)$
$\widehat{f}_{\mathrm{sub}}$	Neural Nets	$\tilde{\mathcal{O}}\left(\mathcal{M}\mathcal{W}^{\frac{1}{2}}\left[\frac{1}{\sqrt{\widetilde{N}}}+\frac{1}{\sqrt{M}}\right]+\mathcal{M}\left[\sqrt{\frac{\ln|\mathcal{C}|/\delta}{\widetilde{N}}}+\sqrt{\frac{\ln 1/\delta}{M}}\right]\right)$

Table: Generalization bounds for the sub-sampled empirical risk minimizer.