An Augmentation-Aware Theory for Self-Supervised Contrastive Learning

Jingyi Cui[1], Hongwei Wen[2], Yisen Wang[#,1]

1 Peking University; 2 University of Twente

Correspondence to yisen.wang@pku.edu.cn

ICML International Conference On Machine Learning

北京大学 PEKING UNIVERSITY

UNIVERSITY OF TWENTE.

Background

- Self-supervised contrastive learning has emerged as a powerful tool to learn meaningful representations from unlabeled data.
- However, in the existing theoretical research, the role of data augmentation is still under-exploited.
- The effects of specific augmentation types such as random cropping and random color distortion are unexplained.

Our Contributions

- We for the first time propose an augmentation-aware error bound for self-supervised contrastive learning, which explicitly includes the quality of data augmentation in the bound without any additional assumptions.
- By proposing a novel semantic label assumption, we analyze specific types of data augmentation including random resized crop and color distortion.
- We conduct experiments to verify our theoretical conclusions.

Mathematical Formulations

Notations. $\bar{x} \in \mathcal{X}$: original input image (with bar notation).
$x := a(\bar{x})$: augmented image (without bar notation).
$a \in \mathcal{A}$: random data augmentation.
$C \in \mathbb{N}$: the number of classes; $[C] := \{1, \dots, C\}$.
$c \in [C] \sim \pi_c$: the class label of $\bar{x}$;
$\pi_c := \mathrm{P}(y = c)$; $\boldsymbol{\pi} = \{\pi_c\}_{c=1}^C$; $\rho_c := \mathrm{P}(\cdot \mid y = c)$.

Data generation process of unsupervised contrastive learning.

(i) draw positive/negative classes: $c, \{c_k\}_{k=1}^K \sim \boldsymbol{\pi}^{K+1}$;
(ii) draw an original sample for the anchor and positives $\bar{x} \sim \rho_c$;
(iii) draw original samples for the negatives $\bar{x}_k \sim \rho_{c_k}$, $k = 1, \dots, K$;
(iv) draw data augmentations $a, a', \{a_k\}_{k=1}^K \sim \mathcal{A}^{K+1}$.

Then we have: anchor $x = a(\bar{x})$, positive sample $x' = a'(\bar{x})$, and negative samples $x_k = a_k(\bar{x}_k)$, $k = 1, \dots, K$.

InfoNCE loss function.

$$\mathcal{L}^{\mathrm{un}}(x, x', \{x_k\}_{k=1}^K; f) := -\log\left(\frac{e^{f(x)^\top f(x')}}{e^{f(x)^\top f(x')} + \sum_{k=1}^K e^{f(x)^\top f(x_k)}}\right).$$

Unsupervised risk.

$$\mathcal{R}^{\mathrm{un}}(f) := \mathbb{E}_{c,\{c_k\}_{k\in[K]}} \mathbb{E}_{\bar{x}\sim\rho_c, \bar{x}_k\sim\rho_{c_k}} \mathbb{E}_{a,a',\{a_k\}_{k=1}^K} \mathcal{L}^{\mathrm{un}}(x, x', \{x_k\}_{k=1}^K; f).$$

Downstream supervised classification. For evaluation, given the learned representation $f : \mathcal{X} \to \mathbb{R}^d$, we train a linear classifier $g = \boldsymbol{W} f : \mathbb{R}^d \to \mathbb{R}^C$ on top of f with $\boldsymbol{W} \in \mathbb{R}^{C\times d}$. Specifically, we use the mean classifier where $\boldsymbol{W} := [\mu_1, \dots, \mu_C]^\top$, $\mu_c := \mathbb{E}_{\bar{x}\sim\rho_c} f(\bar{x}), c \in [C]$ with cross entropy loss function $\mathcal{L}^{\mathrm{sup}}(\bar{x}, c; f)$.

Supervised risk.

$$\mathcal{R}^{sup}(f) := \mathbb{E}_{c\sim\boldsymbol{\pi}} \mathbb{E}_{\bar{x}\sim\rho_c} \mathcal{L}^{\mathrm{sup}}(\bar{x}, c; f).$$

Main Theorem

Theorem 1 (Augmentation-Aware Error Bound).

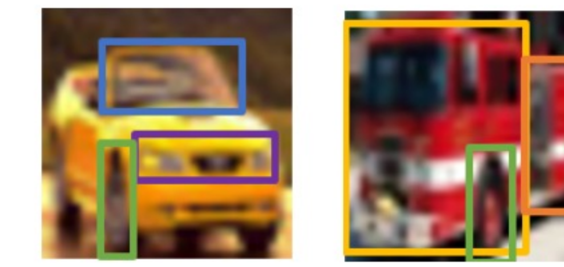
$$\mathcal{R}^{\mathrm{sup}} \le \frac{1}{1-\tau_K}\Big[\mathcal{R}^{\mathrm{un}} - \tau_K \mathbb{E}_{c,\{c_k\}_{k\in[K]}} \log(\mathrm{Col}+1)\Big] + \mathbb{E}_c \mathbb{E}_{\bar{x},\bar{x}'\sim\rho_c} \mathbb{E}_a \min_{a'} \big\| f(a(\bar{x})) - f(a'(\bar{x}')) \big\| + 5\mathbb{E}_c \mathbb{E}_{\bar{x}\sim\rho_c} \max_{a,a'} \big\| f(a(\bar{x})) - f(a'(\bar{x})) \big\|.$$

- ➢ The bound is composed of the unsupervised contrastive risk, CURL's class collision term, and two distance terms.
- ➢ The first distance term represents the minimum distance between two augmented same-class (different) images. It measures how well the same-class images are connected.
- ➢ The second distance term represents the maximum distance between the two augmentations of the same images. It could be understood as the range or variance of data augmentation.
- ➢ The result holds without any further assumptions, especially without the conditional independence assumption of CURL.
- ➢ Under a mild centered representation assumption ($\mathbb{E}_a f(a(\bar{x})) = f(\bar{x})$), the coefficient 5 can be improved to 1.
- ➢ Under the Lipschitz continuous assumption, the distance terms can be on the pixel-level with coefficients c_L (Lipschitz constant).

Impacts of Data Augmentations

Semantic Label Assumption. An image can have several semantic areas with their corresponding semantic labels, i.e., each pixel $\xi_{j,\ell}$ has a semantic label s related to image class y.

Fig. Semantic labels. (a) An *automobile* image has semantic labels *windshield*, *headlights*, and *wheels*; (b) a *truck* image has semantic labels *truck cab*, *cargo box*, and *wheels.*



(a) Automobile. (b) Truck.

If $a(\bar{x})$ contains only same-semantic label pixels (semantic label s),

$$\mathbb{E}_a \min_{a'} \| a(\bar{x}) - a'(\bar{x}') \|_F = 2\left[d^2 \sum_{i\in[3]} \left(\sigma_s^{(i)}\right)^2\right]^{1/2} := 2\sigma.$$

If $a(\bar{x})$ has more than one semantic labels, $\mathbb{E}_a \min_{a'} \| a(\bar{x}) - a'(\bar{x}') \|_F$

$$= 2\sigma + \left[\sum_{j,\ell\in[d], i\in[3]} \mathbf{1}\left[s(\bar{\xi}_{j,\ell}) \ne s_{\max}\right] \left(\mu_s^{(i)} - \mu_{s_{\max}}^{(i)}\right)^2\right]^{1/2}.$$

- ➢ With larger crop size, the cropping area intersects more often with the semantic boundary, i.e., larger MinSameClassDist.
- ➢ Smaller crop size gives a larger variance, i.e., MaxSameImageDist.
- ➢ A trade-off between the distances w.r.t. augmentation strength.

Experimental verification

- We verify the distance trade-offs on TinyImagenet.
- The optimal distance sums corresponds to best accuracy.