Self-Supervised Learning Theory

Weiran Huang（黄维然）

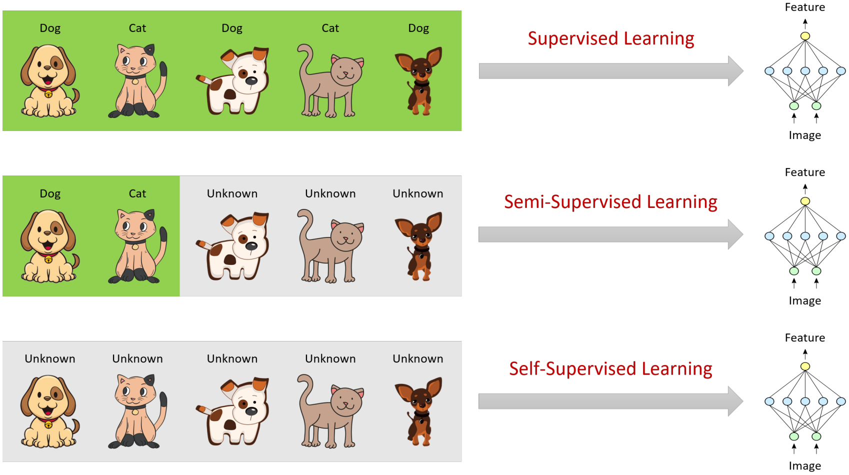
Huawei Noah's Ark Lab

Alumni Forum of IIIS

NOAH'S ARK LAB

Introduction to Self-Supervised Learning

Representation Learning Paradigm Evolution



Supervised Learning

Semi-Supervised Learning

Self-Supervised Learning

What Is Self-Supervised Learning (SSL)?

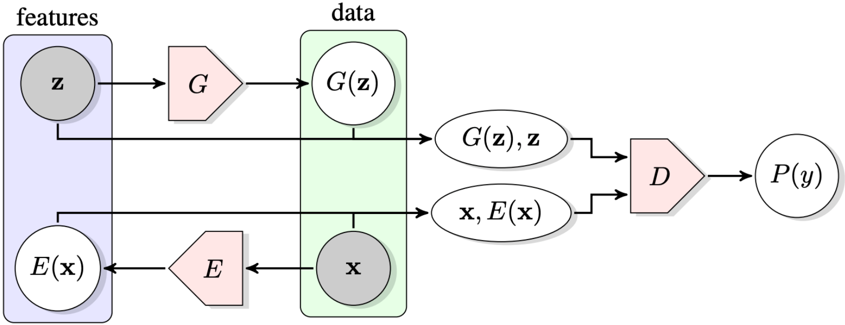
Self-Supervised Learning (SSL) learns data representations through self-supervised tasks, and then use the learned representations for downstream prediction tasks. It has been used in both computer vision [1–3, 11, 12, 17] and natural language processing [7, 8, 10, 15, 16].

There are three common approaches for SSL:

- Generative-Based: learning a bijective mapping between input and representation, e.g., BiGAN [5, 6], BigBiGAN [4].
- Contrastive-Based: maximizing the alignment between the features of positive samples, e.g., SimCLR [1], MoCo [12], Barlow Twins [17].
- Pretext-Based: learning the representation via a handcrafted pretext task, e.g., predicting image rotations [9], GPTs [14].

Generative-Based SSL Examples

BiGAN [5, 6]: match the joint distribution between $(\mathbf{x}, E(\mathbf{x}))$ and $(G(\mathbf{z}), \mathbf{z})$, where E is the feature extractor and G is the generator.



Contrastive-Based SSL Examples

SimCLR [1], Moco [12], BYOL [11], SimSiam [2]: match the representations of different views of the same image.

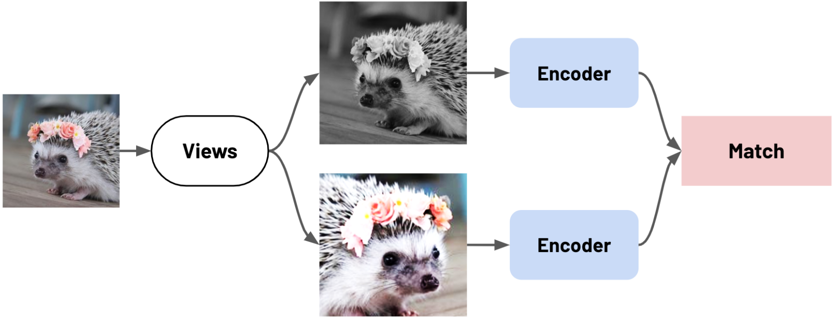
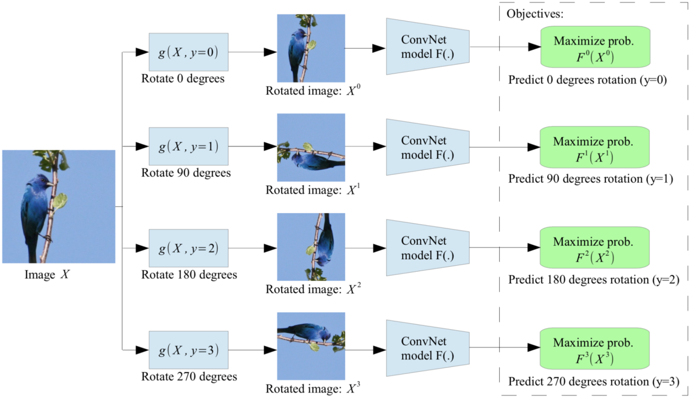


Image from https://ai.stanford.edu/blog/viewmaker/

Pretext-Based SSL Examples

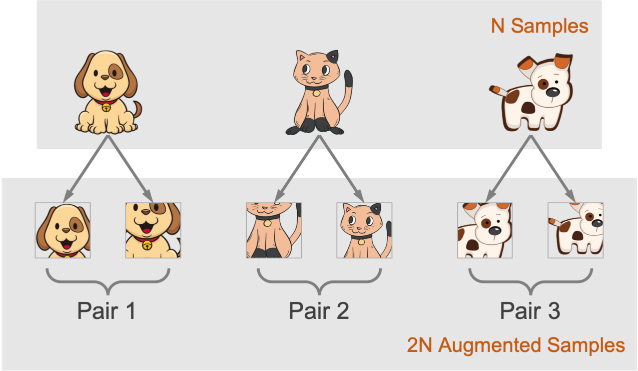
Predicting Image Rotations [9]: manually create labels for input images, and then learn the model as supervised learning usually does.



Contrastive-Based Self-Supervised Learning

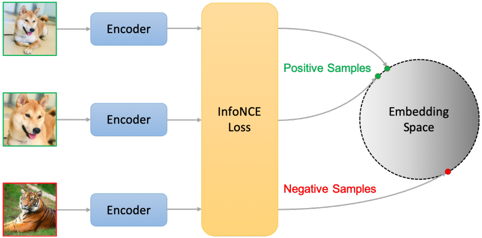
How to Do Contrastive-Based SSL?

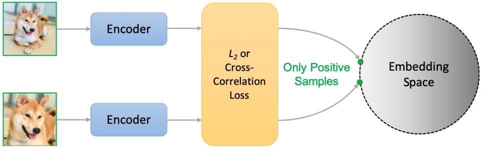
Step 1 of 2: Construct similar sample pairs by data augmentation.



How to Do Contrastive-Based SSL?

Step 2 of 2: Pull the similar sample pairs close to each other in the embedding space (under some regularization to avoid collapse).



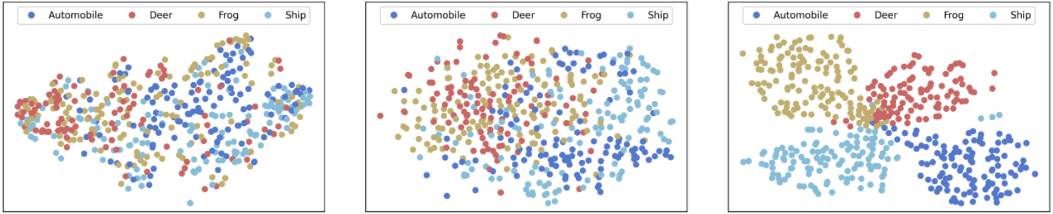


Most contrastive SSL objective can be formulated as

$$\min \mathcal{L}(f) = \mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} \left\| f(\mathbf{x}_1) - f(\mathbf{x}_2) \right\|^2 + \mathcal{L}_{\text{regularization}}(f).$$

Interesting Observations

1. Sample-level alignment leads to class-level clustering;
2. Stronger data augmentation results in more apparent clustering.



(a) Original

(b) Only color distortion

(c) Standard augmentation

Figure: Embedding Space Learned by Contrastive SSL.

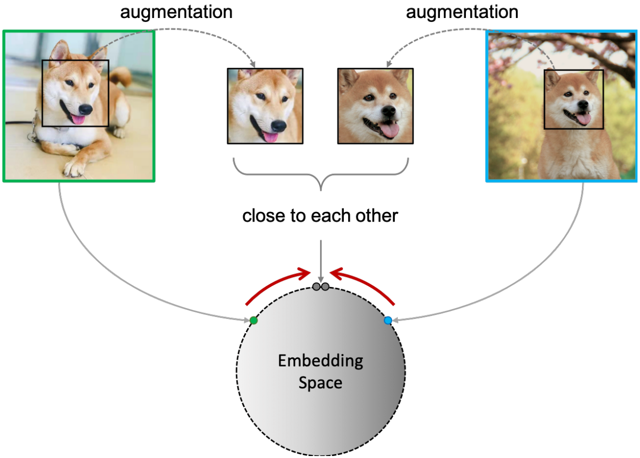
Main Challenge for Performance Analysis

For contrastive-based SSL, data augmentation is the key to success, since data augmentation is the only human knowledge injected.

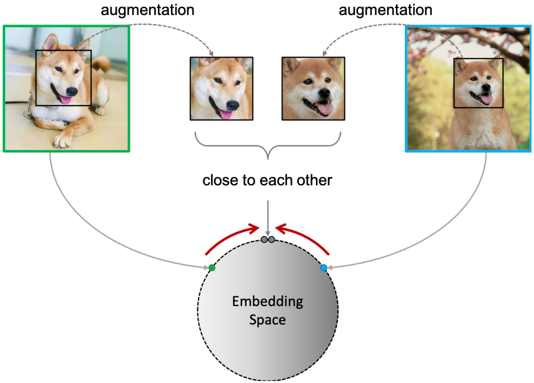
1. How to quantitatively characterize data augmentation?
2. Which kind of embedding space can generalize to downstream tasks?
3. How do the existing methods learn such embedding space?

After addressing the above questions, we can give an explanation for the aforementioned interesting observations.

Data Augmentation Modeling



Data Augmentation Modeling



For a given data augmentation set A, we define the augmented distance between two different samples as

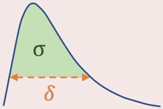
$$d_A(\mathbf{x}_1, \mathbf{x}_2) = \min_{\mathbf{x}_1' \in A(\mathbf{x}_1), \mathbf{x}_2' \in A(\mathbf{x}_2)} \left\| \mathbf{x}_1' - \mathbf{x}_2' \right\| .$$

Data Augmentation Modeling

Definition 1 ((σ, δ)-Augmentation)

The data augmentation set A is called a (σ, δ)-augmentation, if for each class C_k, there exists a subset $C_k^0 \subseteq C_k$ (called the main part of C_k) such that

- $\mathbb{P}[\mathbf{x} \in C_k^0] \geq \sigma \, \mathbb{P}[\mathbf{x} \in C_k]$ where $\sigma \in (0, 1]$,
- $\sup_{\mathbf{x}_1, \mathbf{x}_2 \in C_k^0} d_A(\mathbf{x}_1, \mathbf{x}_2) \leq \delta$.



- Larger σ and smaller δ indicate that the augmented data of each class are more concentrated in terms of the augmented distance.
- For any $A' \supseteq A$, $d_{A'}(\mathbf{x}_1, \mathbf{x}_2) \leq d_A(\mathbf{x}_1, \mathbf{x}_2)$ for any $\mathbf{x}_1, \mathbf{x}_2$. This means that more data augmentations lead to sharper intra-class concentration as δ gets smaller.
- Given δ, we can compute σ by finding the maximum clique of the graph, where each node corresponds to a sample and edge $(\mathbf{x}_1, \mathbf{x}_2)$ exists if $d_A(\mathbf{x}_1, \mathbf{x}_2) \leq \delta$.

Generalization Bound of Contrastive SSL

Theorem 1 (Main Result)

Assume that encoder f with norm r is L-Lipschitz continuous. If the augmentation used in contrastive learning is (σ, δ)-augmented, and

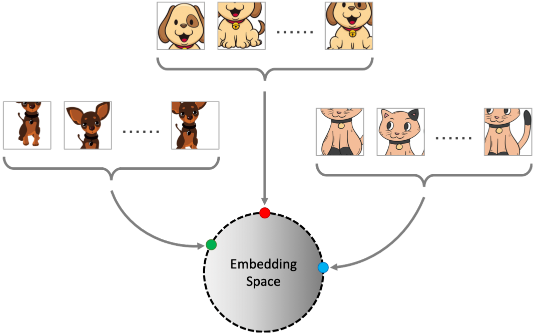
$$\mu_k^\top \mu_\ell < r^2 \left(1 - \rho_{max}(\sigma, \delta, \varepsilon) - \sqrt{2\rho_{max}(\sigma, \delta, \varepsilon)} - \frac{\Delta_\mu}{2}\right)$$

holds for any pair of (ℓ, k) with $\ell \neq k$, then the error rate of downstream classification

$$\mathrm{Err}(G_f) \leq (1 - \sigma) + R_\varepsilon,$$

where $\rho_{max}(\sigma, \delta, \varepsilon) = 2(1 - \sigma) + \frac{R_\varepsilon}{\min_\ell p_\ell} + \sigma \left(\frac{L\delta}{r} + \frac{2\varepsilon}{r}\right)$ and $\Delta_\mu = 1 - \min_{k \in [K]} \frac{\|\mu_k\|^2}{r^2}$.

A Simple Example

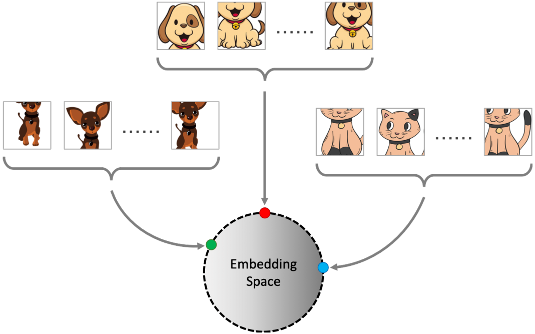


$$\begin{cases} \text{Any two samples from the same class own a same augmented sample } (\sigma = 1, \delta = 0); \\ \text{Each positive pair is embedded to the same point } (\varepsilon = 0, R_\varepsilon = 0). \end{cases}$$

$\Rightarrow$ The samples belonging to the same latent class are mapped to a single point.

$\Rightarrow$ $\frac{\langle \mu_\ell, \mu_k \rangle}{\|\mu_\ell\| \cdot \|\mu_k\|} < 1$ is sufficient to separate the latent classes by the NN classifier.

A Simple Example



In fact, since $\sigma = 1, \delta = 0, \varepsilon = R_\varepsilon = 0$, according to Theorem 1, we have

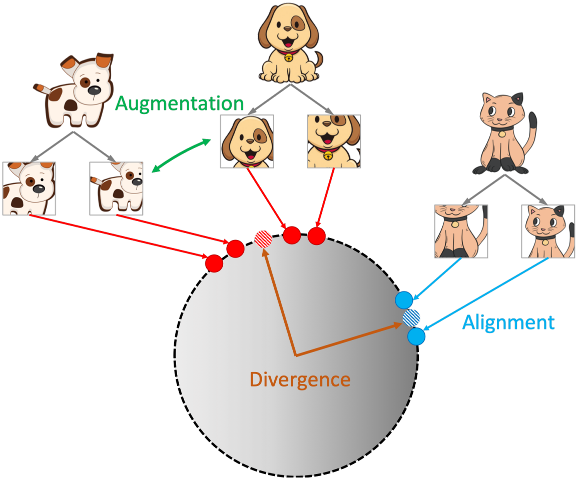
$$\rho_{max}(\sigma, \delta, \varepsilon) = 2(1-\sigma) + \frac{R_\varepsilon}{\min_\ell p_\ell} + \sigma\left(\frac{L\delta}{r} + \frac{2\varepsilon}{r}\right) = 0, \Delta_\mu = 1 - \min_{k\in[K]} \frac{\|\mu_k\|^2}{r^2} = 0.$$

Therefor, $\mu_\ell^\top \mu_k / r^2 < 1 - \rho_{max}(\sigma, \delta, \varepsilon) - \sqrt{2\rho_{max}(\sigma, \delta, \varepsilon)} - \frac{\Delta_\mu}{2} = 1.$

Messages From Theorem 1

1. (Alignment of positive samples) It is the common objective that contrastive algorithms aim to optimize. The good alignment enables the small R_ε, which directly decreases the upper bound of error rate.
2. (Divergence of class centers) The distance between class centers should be large enough. A good alignment property can loosen the divergence condition.
3. (Concentration of augmented data) The augmented data with sharper concentration $(\sigma \to 1, \delta \to 0)$ enable the model to own a smaller upper bound of error rate.

Messages From Theorem 1



Contrastive Loss Functions

- InfoNCE (e.g., SimCLR [1]): pull close positive pairs and push away negative pairs.

$$\mathcal{L}_{\text{InfoNCE}} = -\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} \log \frac{e^{f(\mathbf{x}_1)^\top f(\mathbf{x}_2)}}{e^{f(\mathbf{x}_1)^\top f(\mathbf{x}_2)} + e^{f(\mathbf{x}_1)^\top f(\mathbf{x}^-)}},$$

where $\mathbf{x}, \mathbf{x}'$ are two random samples and A is the data augmentation set.

- Cross-Correlation (e.g., Barlow Twins [17]): decorrelate feature components.

$$\mathcal{L}_{\text{Cross-Corr}} = \sum_{i=1}^{d} (1 - C_{ii})^2 + \lambda \sum_{i=1}^{d} \sum_{i \neq j} C_{ij}^2, \quad \left(\mathbb{E}\left[f(\mathbf{x}_1) f(\mathbf{x}_2)^\top \right] \to I_{d \times d} \right)$$

where $C_{ij} = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} [f_i(\mathbf{x}_1) f_j(\mathbf{x}_2)]$, d is the dimension of encoder f, and f is normalized as $\mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{x}' \in A(\mathbf{x})} [f_i(\mathbf{x}')^2] = 1$ for each dimension.

SSL Loss Functions

The above two losses can be split into two parts:

$$\mathcal{L}(f) = \mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|^2 + \mathcal{L}_{\text{regularization}}(f).$$

- For InfoNCE, we prove that $\mu_k^\top \mu_\ell \lesssim \mathcal{L}_{\text{regularization}}(f)$;
- For Cross-Correlation, we prove that $\mu_k^\top \mu_\ell \lesssim \sqrt{\mathcal{L}_{\text{regularization}}(f)}$.

Therefore, $\mathcal{L}_{\text{regularization}}(f)$ controls the divergence.

Experiments

Augmentations					Accuracy	
(a)	(b)	(c)	(d)	(e)	SimCLR	Barlow Twins
✓	✓	✓	✓	✓	**89.92** ± 0.05	**83.93** ± 0.57
✓	✓	✓	✓	×	88.41 ± 0.11	83.37 ± 0.43
✓	✓	✓	×	×	83.62 ± 0.19	73.70 ± 0.99
✓	✓	×	×	×	62.91 ± 0.25	49.56 ± 0.11
✓	×	×	×	×	62.37 ± 0.09	48.54 ± 0.29

(a) random cropping;
(b) random Gaussian blur;
(c) color dropping (i.e., randomly convert images to grayscale);
(d) color distortion;
(e) random horizontal flipping.

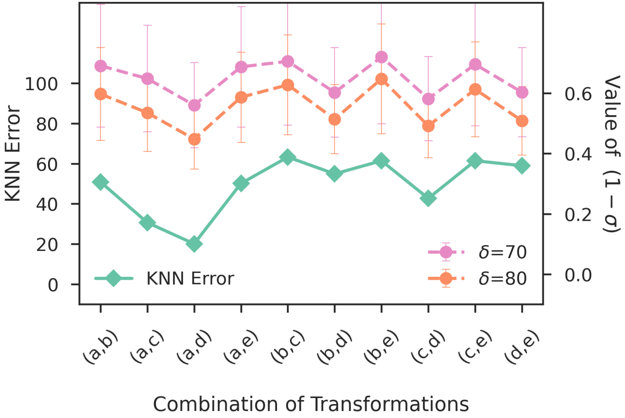
Experiments

Stronger data augmentation results in better performance.

Color Distortion Strength	SimCLR	Barlow Twins
1/8	73.60 ± 0.11	61.13 ± 2.81
1/4	76.25 ± 0.16	68.30 ± 0.15
1/2	78.49 ± 0.09	72.76 ± 1.50
1	$\mathbf{82.64} \pm 0.57$	$\mathbf{78.79} \pm 0.54$

Experiments

Sharper concentration of augmented data (larger σ when fix δ) results in better performance.



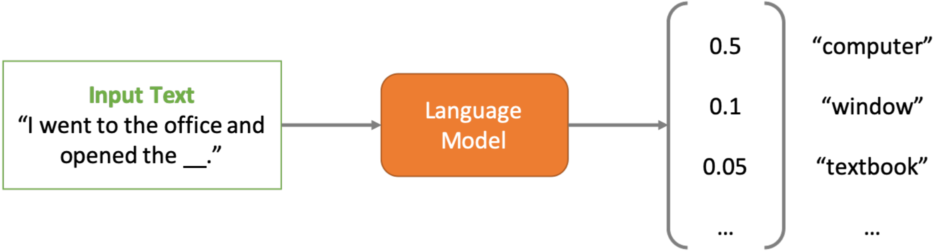
Short Summary

- This work gives a mathematical formulation to model the data augmentation.
- This work provably shows that alignment of positive samples, divergence of class centers and concentration of augmented data are three key factors of contrastive-based SSL generalization.
- This work proves that SimCLR and Barlow Twins implicitly optimize the first two factors.
- Experiments verify that sharper concentration results in better generalization.

Pretext-Based Self-Supervised Learning

What Is Pretext-Based SSL?

Next Word Prediction [14]: use the next word as the label for the given input text.



Why Does Pretext-Based SSL Work?

Lee et al. [13] prove that pretext-based SSL can effectively reduce the sample complexity of downstream tasks under Conditional Independence (CI) between the components of the pretext task conditional on the downstream label.

For example, consider input variable $\mathbf{x}$, pretext label $\mathbf{z}$, and downstream label $\mathbf{y}$ are Gaussian variables.

- If $\mathbf{x} \perp \mathbf{z} \mid \mathbf{y}$, the downstream sample complexity can be reduced to $\tilde{\mathcal{O}}(\dim(\mathbf{y}))$.
- Otherwise, the downstream sample complexity gets worse to $\tilde{\mathcal{O}}(\dim(\mathbf{z}))$.

Can Pretext-Based SSL Be Boosted?

In practice, the CI condition rarely holds, and thus self-supervised learning cannot realize its full potential.

An interesting question raises:

Can we make the CI condition hold with the help of downstream data to boost pretext-based SSL?

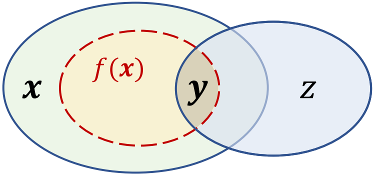


Figure: Applying a function f such that $f(\mathbf{x}) \perp \mathbf{z} \mid \mathbf{y}$.

The Conditions That f Needs to Satisfy

The following two criteria are essential for a meaningful processor f:

$$\mathrm{Cov}[f(\mathbf{x}), \mathbf{z} \mid \mathbf{y}] = 0, \tag{C1}$$

$$\mathbb{E}\left\|\mathbf{y} - W^*_{\mathbf{y},f(\mathbf{x})} f(\mathbf{x})\right\|^2 = \min_{f'} \mathbb{E}\left\|\mathbf{y} - W^*_{\mathbf{y},f'(\mathbf{x})} f'(\mathbf{x})\right\|^2. \tag{C2}$$

Here $W^*_{\mathbf{y},f(\mathbf{x})} \triangleq \arg\min_{W \in \mathbb{R}^{d_y \times d_f}} \mathbb{E}\,\|\mathbf{y} - Wf(\mathbf{x})\|^2$ is defined as the best linear predictor of $\mathbf{y}$ on $f(\mathbf{x})$.

- (C1) is a conditional uncorrelatedness condition, which is a relaxation of the conditional independence condition.
- (C2) ensures that applying function f to input variable $\mathbf{x}$ does not lose the information for predicting $\mathbf{y}$.

Loss Design

We design loss

$$\mathcal{L}(f;\mathcal{P}) \triangleq \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{z},\mathbf{y})\sim\mathcal{P}}\left[\left\|\mathbf{y} - W^*_{\mathbf{y},f(\mathbf{x})} f(\mathbf{x})\right\|^2 - \lambda \left\|\mathbf{z} - W^*_{\mathbf{z},f(\mathbf{x})} f(\mathbf{x})\right\|^2\right] \text{ where } \lambda > 0.$$

Theorem 2 (Rationality of Loss)

Define two sets:

$$\mathcal{A}_{\mathcal{P}} = \left\{ f : f \in \arg\min_{f} \mathcal{L}(f;\mathcal{P}) \right\},$$

$$\mathcal{B}_{\mathcal{P}} = \{ f : f \text{ satisfies Criterion (C1) and Criterion (C2)} \}.$$

Under mild assumptions, there exist a number of population distributions $\{\mathcal{P}\}$'s such that every function in $\mathcal{A}_{\mathcal{P}}$ satisfies Criterion (C1) and Criterion (C2), by choosing a proper parameter λ, i.e., $\mathbb{S} \triangleq \{\mathcal{P} : \mathcal{A}_{\mathcal{P}} \subset \mathcal{B}_{\mathcal{P}}\} \neq \varnothing$.

Insufficient Downstream Samples Provably Fails

To better understand the role of downstream samples, we consider the following loss

$$\mathcal{L}_{\infty,n_0}(f,\mathcal{P}) \triangleq \frac{1}{n_0}\left\|Y_{down} - \widetilde{W}_{down} f(X_{down})\right\|^2 - \lambda \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{z}}\left\|\mathbf{z} - W^*_{\mathbf{z},f(\mathbf{x})} f(\mathbf{x})\right\|^2 .$$

Theorem 3 (Model-Free Lower Bound)

Let

$$\mathcal{A}'_{\mathcal{P}} = \left\{ f : f \in \arg\min_f \mathcal{L}_{\infty,n_0}(f;\mathcal{P}) \right\},$$

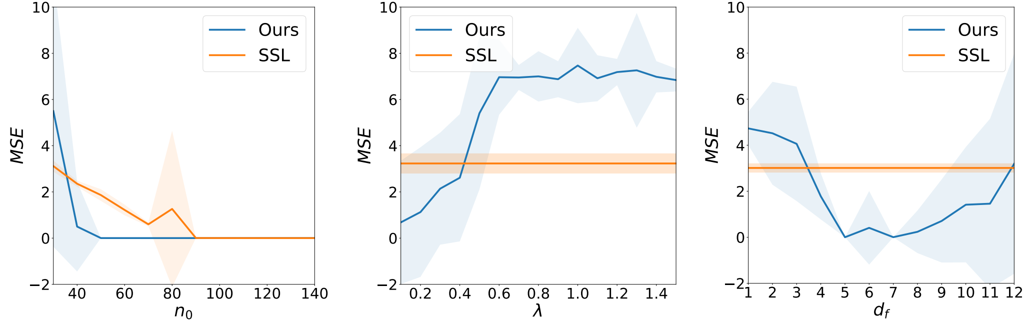
$$\mathcal{B}_{\mathcal{P}} = \{ f : f \text{ satisfies Criterion (C1) and Criterion (C2)} \} .$$

Under mild assumptions, if $n_0 = o(d_f)$, there exists a distribution $\mathcal{P}^0 \in \mathbb{S}$ (i.e., $\mathcal{A}_{\mathcal{P}^0} \subset \mathcal{B}_{\mathcal{P}^0}$), such that $\mathcal{A}'_{\mathcal{P}^0} \cap \mathcal{B}_{\mathcal{P}^0} = \varnothing$.

Remark of Theorem 3

- When $n_0 = o(d_f)$, even if we have infinite pretext data, the criteria cannot be satisfied, resulting in that the downstream performance will get worse. Therefore, it is better NOT to use downstream samples when the downstream samples are insufficient, as the standard self-supervised learning does.
- We can extend the loss in Theorem 3 to a more general loss.
- We can also provide a more precise model-based lower bound of downstream sample size.

Experiments



(a) MSE under different n_0

(b) Large λ hurts MSE

(c) MSE under different d_f

Figure: Downstream performance under different hyperparameters on synthetic data.

Short Summary

We provably answer the question whether we can make the CI condition hold with the help of downstream data to boost pretext-based self-supervised learning.

- It is better NOT to use downstream samples when they are insufficient, as the standard self-supervised learning does.
- We provide both model-free and model-dependent lower bounds of downstream sample size.
- Experiments verify that pretext-based SSL can be boosted with sufficient downstream samples, but will be hurt with insufficient downstream samples.

Possible Future Directions

Self-Supervised Learning is key to human-level intelligence.

— Yann LeCun and Yoshua Bengio (Turing Award Winners)

1. Generalization of SSL: Both IID and OOD.
2. Interpretation of SSL: What kind of features do SSL algorithms learn?
3. Robustness of SSL: How to defend SSL model from an attack?

Thank you!

1. **Towards the Generalization of Contrastive Self-Supervised Learning.**
   Weiran Huang*, Mingyang Yi* (UCAS), Xuyang Zhao* (PKU), arXiv, 2022.
2. **Can Pretext-Based Self-Supervised Learning Be Boosted by Downstream Data? A Theoretical Analysis.**
   Jiaye Teng* (THU), Weiran Huang*, Haowei He* (THU), AISTATS, 2022.

Interns and Visitors Are Welcome

Let’s explore the most cutting-edge and innovative research together!

Reference I

[1] Ting Chen et al. "A simple framework for contrastive learning of visual representations". 2020.

[2] Xinlei Chen and Kaiming He. "Exploring simple siamese representation learning". 2021.

[3] Xinlei Chen et al. "Improved baselines with momentum contrastive learning". 2020.

[4] Jeff Donahue and Karen Simonyan. "Large scale adversarial representation learning". 2019.

[5] Jeff Donahue et al. "Adversarial feature learning". 2017.

[6] Vincent Dumoulin et al. "Adversarially learned inference". 2017.

[7] Hongchao Fang et al. "Cert: Contrastive self-supervised learning for language understanding". 2020.

[8] Tianyu Gao et al. "SimCSE: Simple Contrastive Learning of Sentence Embeddings". 2021.

[9] Spyros Gidaris et al. "Unsupervised representation learning by predicting image rotations". 2018.

[10] John M Giorgi et al. "Declutr: Deep contrastive learning for unsupervised textual representations". 2020.

[11] Jean-Bastien Grill et al. "Bootstrap your own latent: A new approach to self-supervised learning". 2020.

Reference II

[12] Kaiming He et al. “Momentum contrast for unsupervised visual representation learning”. 2020.

[13] Jason D Lee et al. “Predicting what you already know helps: Provable self-supervised learning”. 2020.

[14] Alec Radford et al. “Improving language understanding by generative pre-training”. 2018.

[15] Zhuofeng Wu et al. “Clear: Contrastive learning for sentence representation”. 2020.

[16] Yuanmeng Yan et al. “ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer”. 2021.

[17] Jure Zbontar et al. “Barlow twins: Self-supervised learning via redundancy reduction”. 2021.