The Generalization of Contrastive Self-Supervised Learning

Weiran Huang[1*] **Mingyang Yi**[2*] **Xuyang Zhao**[3*] **Zihao Jiang**[1]




[1]Qing Yuan Research Institute, Shanghai Jiao Tong University

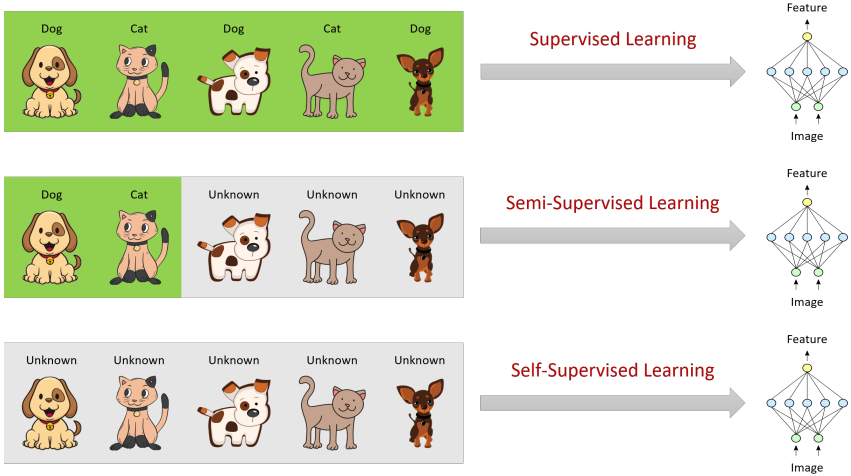
[2]Huawei Noah's Ark Lab

[3]School of Mathematical Sciences, Peking University

ICLR 2023

Introduction to Contrastive Self-Supervised Learning

Representation Learning Paradigm Evolution



What Is Self-Supervised Learning (SSL)?

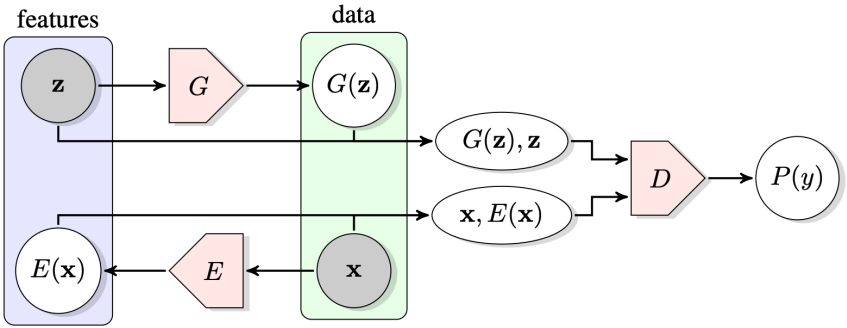
Self-Supervised Learning (SSL) learns data representations through self-supervised tasks, and then use the learned representations for downstream prediction tasks. It has been used in both computer vision [2–4, 12, 14, 18] and natural language processing [8, 9, 11, 16, 17].

There are three common approaches for SSL:

- Generative-Based: learning a bijective mapping between input and representation, e.g., BiGAN [6, 7], BigBiGAN [5].
- Pretext-Based: learning the representation via a handcrafted pretext task, i.e., image colorization [19], predicting image rotations [10].
- Contrastive-Based: maximizing the alignment between the features of positive samples, e.g., SimCLR [2], MoCo [14], Barlow Twins [18].

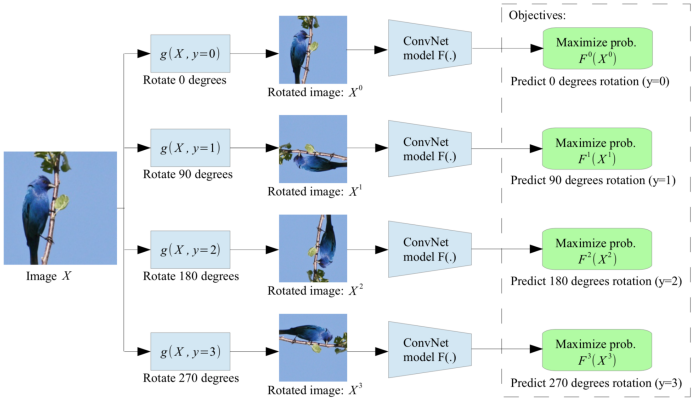
Approaches for SSL (Generative-Based)

BiGAN [6, 7]: 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.



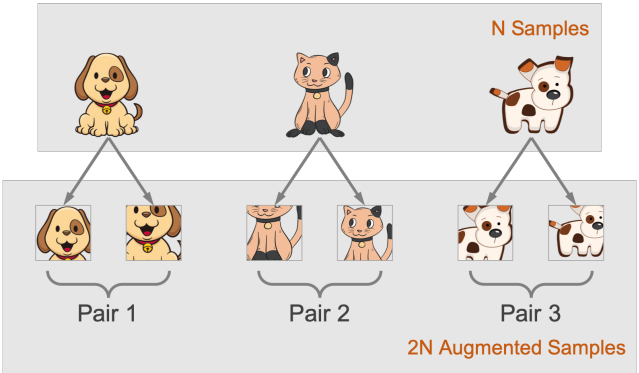
Approaches for SSL (Pretext-Based)

Predicting Image Rotations [10]: manually create labels for input images, and then learn the model as supervised learning usually does.



Approaches for SSL (Contrastive-Based)

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



Approaches for SSL (Contrastive-Based)

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

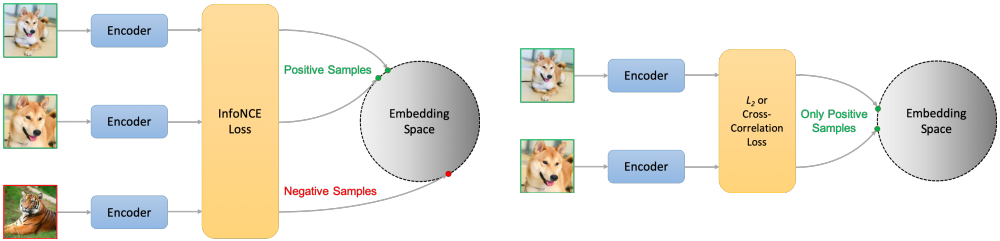


Figure: Left: Contrastive learning w/ negative samples (e.g., SimCLR [2]). Right: Contrastive learning w/o negative samples (e.g., Barlow Twins [18]).

Approaches for SSL (Contrastive-Based)

- InfoNCE Loss: 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 Loss: decorrelate the components of representation.

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

  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.

How to Evaluate Self-Supervised Learned Representations?

Given a training dataset $D = X \times Y$,

1. Do self-supervised training only using (augmented) X, and obtain an encoder f.
2. Train a linear classifier W on the top of encoder f using $D = X \times Y$.

The performance of self-supervised learned representation f is measured by the test accuracy of $W \circ f(\cdot)$.

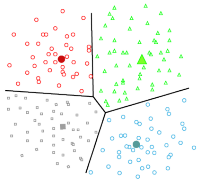


Figure: Clustered structure can be well classified by a linear layer.

Interesting Observations in Contrastive SSL

1. Aligning positive samples (augmented from the "same data point") is able to gather the samples from the "same latent class" into a cluster.

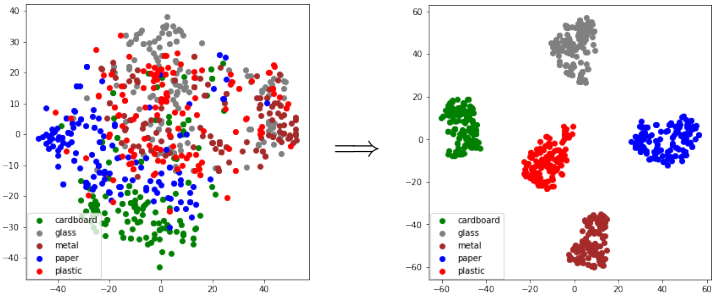
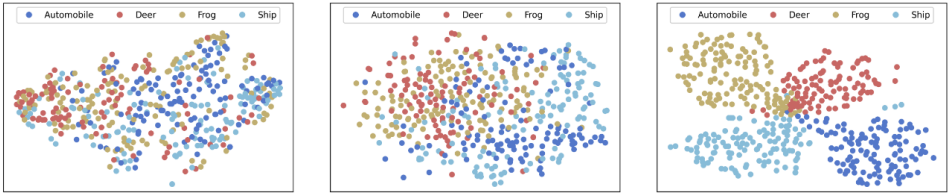


Figure: Embedding Space (https://github.com/mwdhont/SimCLRv1-keras-tensorflow).

Interesting Observations in Contrastive SSL

2. Richer data augmentation leads to a more clustered structure in the embedding space.



(a) Original **(b)** Only color distortion **(c)** Standard augmentation

Figure: SimCLR's embedding space with different richnesses of data augmentations.

Interesting Observations in Contrastive SSL

3. The best composition of augmentations: random cropping and random color distortion.

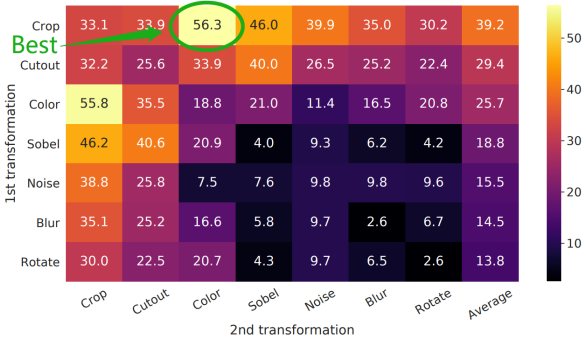


Figure: Experimental results reported in SimCLR paper.

Interesting Observations in Contrastive SSL

4. Barlow Twins decorrelates components of representation instead of directly optimizing the geometry of embedding space, but it still results in the clustered structure.

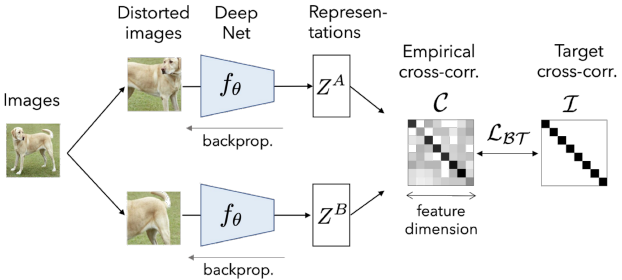


Figure: Barlow Twins aims to decorrelate the components of representation.

Existing Work for Understanding Contrastive SSL

(Empirical Understanding) Wang et al. [15] propose two factors: alignment and uniformity. They empirically verify that two factors are highly correlated to the downstream performance.

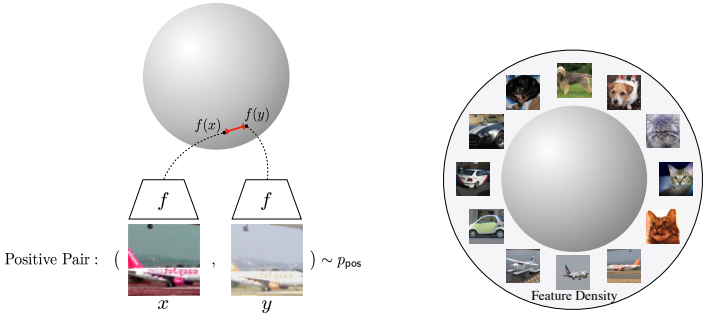


Figure: Alignment and Uniformity

Existing Work for Understanding Contrastive SSL

(Theoretical Understanding) Arora et al. [1] provide a generalization bound of contrastive SSL, by hypothesizing that positive pairs are sampled from the same latent class, i.e.,

$$D_{\text{positive}}(\mathbf{x}_1, \mathbf{x}_2) = \mathbb{E}_{C_k} D_{C_k}(\mathbf{x}_1) D_{C_k}(\mathbf{x}_2),$$
$$D_{\text{negative}}(\mathbf{x}^-) = \mathbb{E}_{C_k} D_{C_k}(\mathbf{x}^-).$$

Main Questions

Existing work avoids to characterize the role of data augmentation in contrastive SSL, which is the key to success since data augmentation is the only human knowledge injected.

In this report, we propose a quantitative description of data augmentation, which enables us to provably answer the following two questions:

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

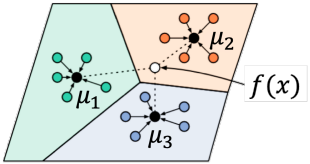
After addressing the above two questions, we can give an explanation for the aforementioned interesting observations.

Mathematical Formulation

Nearest Neighbor (NN) Classifier

Nearest neighbor classifier is defined as

$$G_f(\mathbf{x}) = \arg\min_{k \in [K]} \|f(\mathbf{x}) - \mu_k\|,$$



where $\mu_k := \mathbb{E}_{\mathbf{x} \in C_k} \mathbb{E}_{\mathbf{x}' \in A(\mathbf{x})}[f(\mathbf{x}')]$ is the center of class C_k.

It can be regarded as a special linear classifier, i.e.,

$$G_f(\mathbf{x}) = \arg\max_{k \in [K]} (Wf(\mathbf{x}) + b)_k,$$

by setting the k-th row of W to be μ_k and $b_k = -\frac{1}{2}\|\mu_k\|^2$.

- A directly learned linear classifier should have better performance than the nearest neighbor classifier G_f.

Connection Between the Self-Supervised and Downstream Tasks

To analyze the generalization of contrastive SSL, we need to measure how well the samples are clustered by classes in the embedding space.

However, the learning objective of self-supervised task can not include label information. Most contrastive SSL objective can be formulated as

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

- There is a gap between $\mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|^2$ and $\mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2 \in C_k} \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|^2$.
- What guarantees that different samples from the same latent class are pulled close?

Connection Between the Self-Supervised and Downstream Tasks

To analyze the generalization of contrastive SSL, we need to measure how well the samples are clustered by classes in the embedding space.

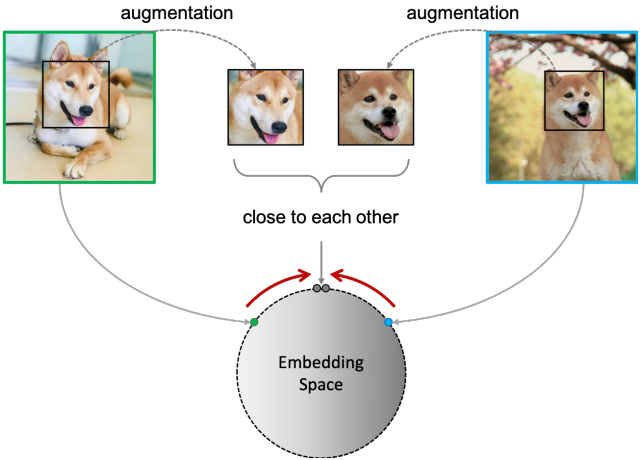
However, the learning objective of self-supervised task can not include label information. Most contrastive SSL objective can be formulated as

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

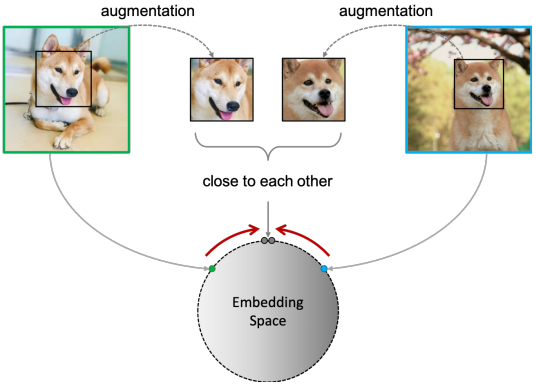
- There is a gap between $\mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|^2$ and $\mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2 \in C_k} \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|^2$.
- What guarantees that different samples from the same latent class are pulled close?

Data Augmentation!

Data Augmentation Modeling



Data Augmentation Modeling



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

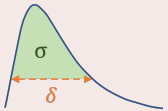
$$d_A(\mathbf{x}_1, \mathbf{x}_2) = \min_{\mathbf{x}_1' \in A(\mathbf{x}_1), \mathbf{x}_2' \in A(\mathbf{x}_2)} \left\| \mathbf{x}_1' - \mathbf{x}_2' \right\|.$$

Data Augmentation Modeling

Definition 1 ((σ, δ)-Augmentation)

The collection of augmented data $A(\mathbf{x})$ is (σ, δ)-augmented, if for each class C_k, there exists a subset $C_k^0 \subseteq C_k$ (called main part of C_k) such that

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

Main Part Samples With Good Alignment

Our analysis focus on the samples located in the main part $C_1^0 \cup \cdots \cup C_K^0$.

When contrastive learning finishes, most of such samples are expected to have good alignment property.

Thus, we can write the samples in the main part with good alignment property as

$$(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon,$$

where $S_\varepsilon := \left\{ \mathbf{x} \in \cup_{k=1}^K C_k \colon \forall \mathbf{x}_1, \mathbf{x}_2 \in A(\mathbf{x}), \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\| \leq \varepsilon \right\}$.

Lemma 1 (Error Rate)

If samples in $(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon(f)$ can be correctly classified by a classifier G, then its downstream error rate is

$$\mathrm{Err}(G) := \sum_{k=1}^K \mathbb{P}[G(\mathbf{x}) \neq k, \forall \mathbf{x} \in C_k] \leq (1-\sigma) + \mathbb{P}\left[\overline{S_\varepsilon}\right].$$

Main Part Samples With Good Alignment

To make the lemma useful, we need to

- Upper bound $\mathbb{P}\left[\overline{S_\varepsilon}\right] =: R_\varepsilon$;
- Explore when samples in $(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon(f)$ can be correctly classified by a NN classifier G_f.

Main Part Samples With Good Alignment

To make the lemma useful, we need to

- Upper bound $\mathbb{P}\left[\overline{S_\varepsilon}\right] =: R_\varepsilon$;
- Explore when samples in $(C_1^0 \cup \cdots \cup C_K^0) \cap S_\varepsilon(f)$ can be correctly classified by a NN classifier G_f.

Theorem 1 (Upper Bound of R_ε)

If encoder f is L-Lipschitz continuous, then

$$R_\varepsilon^2 \leq \eta(\varepsilon)^2 \cdot \mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|^2 = \eta(\varepsilon)^2 \cdot \mathcal{L}_{\text{align}}(f),$$

where $\eta(\varepsilon) = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$.

Generalization Bound

Theorem 2 (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

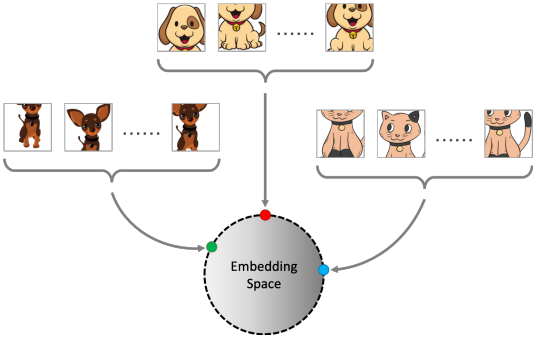
$$\mu_\ell^\top \mu_k < r^2 \left(1 - \rho_{max}(\sigma, \delta, \varepsilon) - \sqrt{2\rho_{max}(\sigma, \delta, \varepsilon)} - \frac{\Delta_\mu}{2}\right)$$

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

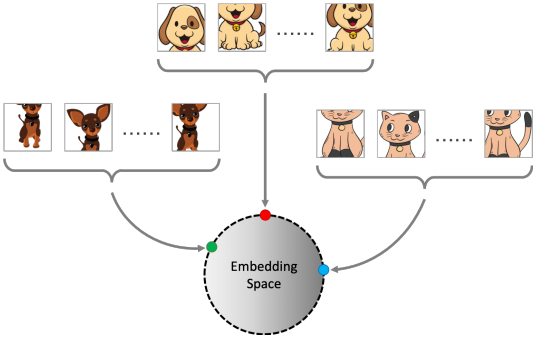


$$\begin{cases} \text{Any two samples from the same class own a same augmented sample } (\sigma = 1, \delta = 0); \\ \text{Each positive pair is embedded to the same point } (\varepsilon = 0, R_\varepsilon = 0). \end{cases}$$

$\Rightarrow$ 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 2, we have

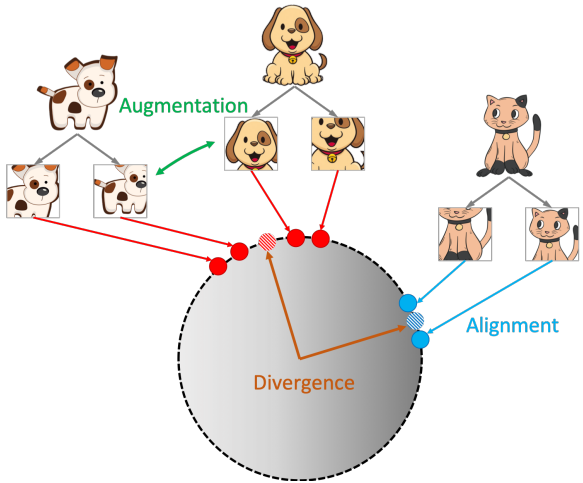
$$\rho_{max}(\sigma, \delta, \varepsilon) = 2(1-\sigma) + \frac{R_\varepsilon}{\min_\ell p_\ell} + \sigma\left(\frac{L\delta}{r} + \frac{2\varepsilon}{r}\right) = 0, \Delta_\mu = 1 - \min_{k\in[K]} \frac{\|\mu_k\|^2}{r^2} = 0.$$

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 the Theorems

1. (Alignment) The pulled closely enough positive samples in the embedding space leads to a small R_ε, which directly decrease the upper bound of error rate;
2. (Divergence) The intra-class centers μ_k in the embedding space should be distinguishable enough, i.e., the minimal $\mu_\ell^\top \mu_k$ of each pair of $\ell \neq k$ should be smaller than a threshold;
3. (Augmentation) The data augmentations directly affect the upper bound of error rate. The augmented data with sharper intra-class concentration space (i.e., corresponds with $\sigma \to 1, \delta \to 0$) enables the model to own a smaller error rate on the downstream classification task.

Messages From the Theorems



Compared With Alignment and Uniformity in [15]

- Both have the same meaning of "alignment", since it is the common objective that algorithms aim to optimize.
- We propose "divergence" instead of "uniformity" to better characterize the sufficient condition of generalization. For example, one can conclude that a good alignment property can loose the divergence condition.
- "Alignment and uniformity" are empirical indicators for generalization, while "alignment and divergence" have explicit theoretical guarantee for generalization.
- Perfect "alignment and uniformity" (if exists) can minimize the InfoNCE loss. However, they are not guaranteed to minimize other existing effective losses. Therefore, they may not be the necessary properties for SSL generalization. In contrast, we will show that both the InfoNCE and cross-correlation loss (implicitly) optimize the "alignment and divergence" property.

SSL Algorithm Analysis

SSL Loss Functions

We now take a close look at two canonical contrastive learning algorithms, SimCLR [2] and Barlow Twins [18].

We will show that their losses can be split into two parts:

$$\mathcal{L}(f) = \mathcal{L}_{\text{positive}}(f) + \mathcal{L}_{\text{regularization}}(f),$$

where $\mathcal{L}_{\text{positive}}(f)$ controls the alignment property and $\mathcal{L}_{\text{regularization}}(f)$ prevents the collapse of representation.

An effective regularizer should be able to make the angles between different classes large.

SimCLR

The InfoNCE loss can be written as ($\|f\| = 1$):

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

SimCLR

The InfoNCE loss can be written as ($\|f\| = 1$):

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

$$= \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} \left[-f(\mathbf{x}_1)^\top f(\mathbf{x}_2) + \log \left(e^{f(\mathbf{x}_1)^\top f(\mathbf{x}_2)} + e^{f(\mathbf{x}_1)^\top f(\mathbf{x}^-)} \right) \right]$$

SimCLR

The InfoNCE loss can be written as ($\|f\| = 1$):

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

$$= \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} \left[-f(\mathbf{x}_1)^\top f(\mathbf{x}_2) + \log \left(e^{f(\mathbf{x}_1)^\top f(\mathbf{x}_2)} + e^{f(\mathbf{x}_1)^\top f(\mathbf{x}^-)} \right) \right]$$

$$= \underbrace{\frac{1}{2} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} \|f(\mathbf{x}_1) - f(\mathbf{x}_2)\|^2 - 1}_{\mathcal{L}_1(f)} + \underbrace{\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} \left[\log \left(e^{f(\mathbf{x}_1)^\top f(\mathbf{x}_2)} + e^{f(\mathbf{x}_1)^\top f(\mathbf{x}^-)} \right) \right]}_{\mathcal{L}_2(f)}.$$

Divergence of SimCLR

Theorem 3

Assume that encoder f with norm 1 is L-Lipschitz continuous. If the augmented data used in SimCLR is (σ, δ)-augmented, then for any $\varepsilon > 0$ and $k \neq \ell$,

$$\mu_k^\top \mu_\ell \leq \log\left(\exp\left\{\frac{\mathcal{L}_2(f) + \tau(\varepsilon, \sigma, \delta)}{p_k p_\ell}\right\} - \exp(1-\varepsilon)\right),$$

where $\tau(\sigma, \delta, \varepsilon, R_\varepsilon)$ is a non-negative term, decreasing with smaller $\varepsilon, R_\varepsilon$ or sharper concentration of augmented data, and $\tau(\sigma, \delta, \varepsilon, R_\varepsilon) = 0$ when $\sigma = 1, \delta = 0, \varepsilon = 0, R_\varepsilon = 0$.

Divergence of SimCLR

Theorem 3

Assume that encoder f with norm 1 is L-Lipschitz continuous. If the augmented data used in SimCLR is (σ, δ)-augmented, then for any $\varepsilon > 0$ and $k \neq \ell$,

$$\mu_k^\top \mu_\ell \leq \log\left(\exp\left\{\frac{\mathcal{L}_2(f) + \tau(\varepsilon, \sigma, \delta)}{p_k p_\ell}\right\} - \exp(1-\varepsilon)\right),$$

where $\tau(\sigma, \delta, \varepsilon, R_\varepsilon)$ is a non-negative term, decreasing with smaller $\varepsilon, R_\varepsilon$ or sharper concentration of augmented data, and $\tau(\sigma, \delta, \varepsilon, R_\varepsilon) = 0$ when $\sigma = 1, \delta = 0, \varepsilon = 0, R_\varepsilon = 0$.

- Divergence $\mu_k^\top \mu_\ell$ can be controlled by $\mathcal{L}_2(f)$.
- $\tau(\varepsilon, \sigma, \delta)$ depends on the alignment property.

Simple Contrastive Loss

The form of InfoNCE loss is critical to meet the requirement of divergence.

- If we reformulate the contrastive loss in a linear form such that

$$\mathcal{L}'(f) = \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} \left[-f(\mathbf{x}_1)^\top f(\mathbf{x}_2) + \lambda f(\mathbf{x}_1)^\top f(\mathbf{x}^-) \right] = \mathcal{L}_1(f) + \lambda \mathcal{L}_2'(f),$$

where $\mathcal{L}_2'(f) = \mathbb{E}_{\mathbf{x},\mathbf{x}'} \mathbb{E}_{\substack{\mathbf{x}_1 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} f(\mathbf{x}_1)^\top f(\mathbf{x}^-) = \| \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{x}_1 \in A(\mathbf{x})} [f(\mathbf{x}_1)] \|^2.$

Simple Contrastive Loss

The form of InfoNCE loss is critical to meet the requirement of divergence.

- If we reformulate the contrastive loss in a linear form such that

$$\mathcal{L}'(f) = \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} \left[-f(\mathbf{x}_1)^\top f(\mathbf{x}_2) + \lambda f(\mathbf{x}_1)^\top f(\mathbf{x}^-) \right] = \mathcal{L}_1(f) + \lambda \mathcal{L}_2'(f),$$

  where $\mathcal{L}_2'(f) = \mathbb{E}_{\mathbf{x},\mathbf{x}'} \mathbb{E}_{\substack{\mathbf{x}_1 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} f(\mathbf{x}_1)^\top f(\mathbf{x}^-) = \| \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{x}_1 \in A(\mathbf{x})} [f(\mathbf{x}_1)] \|^2$.

- It is equivalent to InfoNCE with infinite temperature.

Simple Contrastive Loss

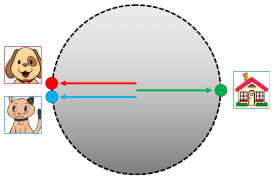
The form of InfoNCE loss is critical to meet the requirement of divergence.

- If we reformulate the contrastive loss in a linear form such that

$$\mathcal{L}'(f) = \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} \left[-f(\mathbf{x}_1)^\top f(\mathbf{x}_2) + \lambda f(\mathbf{x}_1)^\top f(\mathbf{x}^-) \right] = \mathcal{L}_1(f) + \lambda \mathcal{L}_2'(f),$$

  where $\mathcal{L}_2'(f) = \mathbb{E}_{\mathbf{x},\mathbf{x}'} \mathbb{E}_{\substack{\mathbf{x}_1 \in A(\mathbf{x}) \\ \mathbf{x}^- \in A(\mathbf{x}')}} f(\mathbf{x}_1)^\top f(\mathbf{x}^-) = \| \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{x}_1 \in A(\mathbf{x})} [f(\mathbf{x}_1)] \|^2$.

- It is equivalent to InfoNCE with infinite temperature.
- Minimizing $\mathcal{L}_2'(f)$ only leads to f with zero mean in the embedding space.



Barlow Twins

The cross-correlation loss decorrelates the different vector components of $f(\mathbf{x})$:

$$\mathcal{L}_{\text{Cross-Corr}} = \sum_{i=1}^{d} \left(1 - \mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1, \mathbf{x}_2 \in A(\mathbf{x})} [f_i(\mathbf{x}_1) f_i(\mathbf{x}_2)] \right)^2 + \lambda \sum_{i \neq j} \left(\mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1, \mathbf{x}_2 \in A(\mathbf{x})} [f_i(\mathbf{x}_1) f_j(\mathbf{x}_2)] \right)^2$$

Barlow Twins

The cross-correlation loss decorrelates the different vector components of $f(\mathbf{x})$:

$$\mathcal{L}_{\text{Cross-Corr}} = \sum_{i=1}^{d} \left(1 - \mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} [f_i(\mathbf{x}_1) f_i(\mathbf{x}_2)]\right)^2 + \lambda \sum_{i \neq j} \left(\mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} [f_i(\mathbf{x}_1) f_j(\mathbf{x}_2)]\right)^2$$

$$= (1-\lambda) \underbrace{\sum_{i=1}^{d} \left(1 - \mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} [f_i(\mathbf{x}_1) f_i(\mathbf{x}_2)]\right)^2}_{\mathcal{L}_1(f)} + \lambda \underbrace{\left\| \mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})} [f(\mathbf{x}_1) f(\mathbf{x}_2)^\top] - I_d \right\|^2}_{\mathcal{L}_2(f)}.$$

Alignment of Barlow Twins

Since $\mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{x}' \in A(\mathbf{x})}[f_i(\mathbf{x}')^2] = 1$, we have

Lemma 2

For a given encoder f, the expected distance between embedded positive samples $\mathcal{L}_{\text{positive}}(f)$ is upper bounded via $\mathcal{L}_1(f)$, namely,

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

Divergence of Barlow Twins

Theorem 4

Assume that encoder f is L-Lipschitz continuous. If the augmented data used in Barlow Twins is (σ, δ)-augmented, then for any $k \neq \ell$, we have

$$\mu_k^\top \mu_\ell \leq \sqrt{\frac{2}{p_k p_\ell}\left(\mathcal{L}_2(f) + \tau(\varepsilon, \sigma, \delta) - \frac{d-K}{2}\right)},$$

where $\tau(\varepsilon, \sigma, \delta)$ satisfies

$$\left\| \mathop{\mathbb{E}}_{\mathbf{x}} \mathop{\mathbb{E}}_{\mathbf{x}_1, \mathbf{x}_2 \in A(\mathbf{x})} \left[f(\mathbf{x}_1) f(\mathbf{x}_2)^\top \right] - \sum_{k=1}^{K} p_k \mu_k \mu_k^\top \right\|^2 \leq \tau(\varepsilon, \sigma, \delta).$$

Proof Idea

Recall that $\mathcal{L}_2(f) := \left\| \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2 \in A(\mathbf{x})}[f(\mathbf{x}_1)f(\mathbf{x}_2)^\top] - I_d \right\|^2$ and we need $\mu_k^\top \mu_\ell$.

First, we approximate $\mathbb{E}\left[f(\mathbf{x}_1)f(\mathbf{x}_2)^\top\right] \approx \mathbb{E}\left[f(\mathbf{x}_1)f(\mathbf{x}_1)^\top\right] \approx \sum_{k=1}^K p_k \mu_k \mu_k^\top$.

Second, we need to connect $\sum_{k=1}^K p_k \mu_k \mu_k^\top$ to $\mu_k^\top \mu_\ell$:

$$
\begin{aligned}
\left\| \sum_{k=1}^K p_k \mu_k \mu_k^\top - I_d \right\|^2 &= \operatorname{tr}(UU^\top UU^\top - 2UU^\top + I_d) \\
&= \operatorname{tr}(U^\top UU^\top U - 2U^\top U + I_K) + d - K \qquad (1) \\
&= \left\| U^\top U - I_K \right\|^2 + d - K \\
&= \sum_{k=1}^K \sum_{\ell=1}^K (\sqrt{p_k p_\ell}\mu_k^\top \mu_\ell - \delta_{k\ell})^2 + d - K \\
&\geq p_k p_\ell (\mu_k^\top \mu_\ell)^2 + d - K.
\end{aligned}
$$

Application of Theory

Experiments

Dataset	Transformations					Accuracy			
	(a)	(b)	(c)	(d)	(e)	SimCLR	Barlow Twins	MoCo	SimSiam
CIFAR-10	✓	✓	✓	✓	✓	**89.76 ± 0.12**	**86.91 ± 0.09**	**90.12 ± 0.12**	**90.59 ± 0.11**
	✓	✓	✓	✓		88.48 ± 0.22	85.38 ± 0.37	89.69 ± 0.11	89.34 ± 0.09
	✓	✓	✓			83.50 ± 0.14	82.00 ± 0.59	86.78 ± 0.07	85.38 ± 0.09
	✓	✓				63.23 ± 0.05	67.83 ± 0.94	75.12 ± 0.28	63.27 ± 0.30
	✓					62.74 ± 0.18	67.77 ± 0.69	74.94 ± 0.22	61.47 ± 0.74
CIFAR-100	✓	✓	✓	✓	✓	**57.74 ± 0.12**	**57.99 ± 0.29**	**64.19 ± 0.14**	**63.48 ± 0.16**
	✓	✓	✓	✓		55.43 ± 0.10	55.22 ± 0.25	62.50 ± 0.28	60.31 ± 0.41
	✓	✓	✓			45.10 ± 0.25	50.40 ± 0.64	57.04 ± 0.21	51.42 ± 0.14
	✓	✓				28.01 ± 0.18	34.11 ± 0.59	40.18 ± 0.04	26.26 ± 0.30
	✓					27.95 ± 0.09	34.05 ± 1.13	39.63 ± 0.31	25.90 ± 0.83

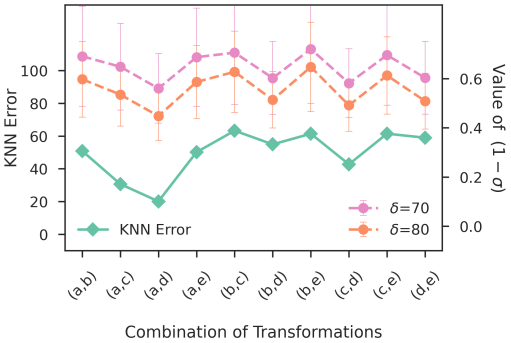
(a) random cropping; (b) random Gaussian blur;
(c) color dropping; (d) color distortion;
(e) random horizontal flipping.

Experiments

Stronger data augmentation indeed leads to the better performance of contrastive self-supervised learning.

Dataset	Color Distortion Strength	Accuracy			
		SimCLR	Barlow Twins	MoCo	SimSiam
CIFAR-10	1	$\mathbf{82.75 \pm 0.24}$	$\mathbf{82.58 \pm 0.25}$	$\mathbf{86.68 \pm 0.05}$	$\mathbf{82.50 \pm 1.05}$
	1/2	78.76 ± 0.18	81.88 ± 0.25	84.30 ± 0.14	81.80 ± 0.15
	1/4	76.37 ± 0.11	79.64 ± 0.34	82.76 ± 0.09	78.80 ± 0.17
	1/8	74.23 ± 0.16	77.96 ± 0.16	81.20 ± 0.12	76.09 ± 0.50
CIFAR-100	1	$\mathbf{46.67 \pm 0.42}$	$\mathbf{50.39 \pm 1.09}$	$\mathbf{58.50 \pm 0.51}$	$\mathbf{49.94 \pm 2.01}$
	1/2	40.21 ± 0.05	48.76 ± 0.25	55.08 ± 0.09	46.27 ± 0.46
	1/4	36.67 ± 0.08	46.22 ± 0.71	52.09 ± 0.18	42.02 ± 0.34
	1/8	34.75 ± 0.20	44.72 ± 0.26	49.43 ± 0.16	36.26 ± 0.34

Experiments



(a) random cropping;

(b) random Gaussian blur;

(c) color dropping;

(d) color distortion;

(e) random horizontal flipping.

- Fix one transformation as (a), we observe that $(a, d) < (a, c) < (a, e) \approx (a, b)$;
- Composition (a, d) has the sharpest concentration and best performance.

Experiments

We revise the InfoNCE loss by

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\mathbf{x}_2\in A(\mathbf{x}) \\ \mathbf{x}^-\in A(\mathbf{x}')}} \|f(\mathbf{x}_1)-f(\mathbf{x}_2)\|^2 + \lambda \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'} \mathop{\mathbb{E}}_{\substack{\mathbf{x}_1,\in A(\mathbf{x}) \\ \mathbf{x}_i^-\in A(\mathbf{x}')}} \left[\log\left(\sum_i e^{f(\mathbf{x}_1)^\top f(\mathbf{x}_i^-)/\tau}\right)\right].$$

Epoch	200	400	600	800
SimCLR ($\tau = 1$)	76.8	80.6	84.3	86.8
SimCLR ($\tau = 0.5$)	80.4	84.1	87.4	89.4
Ours ($\tau = 0.2$)	**83.0**	**86.2**	**88.7**	**90.3**

Summary

- We provide a mathematical formulation to model the data augmentation.
- We show that alignment of positive samples, divergence of class centers and concentration of augmented data are three key factors of self-supervised contrastive learning.
- We prove that SimCLR and Barlow Twins implicitly optimize the first two factors.
- We empirically verify that sharper concentration results in better generalization.

PS: Can Masked Auto-Encoder (MAE [13]) be analyzed by the proposed framework?

Thank you!

Interns and visitors are welcome!

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

For Further Reading I

[1] Sanjeev Arora et al. “A theoretical analysis of contrastive unsupervised representation learning”. 2019.

[2] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. 2020.

[3] Xinlei Chen and Kaiming He. “Exploring simple siamese representation learning”. 2021.

[4] Xinlei Chen et al. “Improved baselines with momentum contrastive learning”. 2020.

[5] Jeff Donahue and Karen Simonyan. “Large scale adversarial representation learning”. 2019.

[6] Jeff Donahue et al. “Adversarial feature learning”. 2017.

[7] Vincent Dumoulin et al. “Adversarially learned inference”. 2017.

[8] Hongchao Fang et al. “Cert: Contrastive self-supervised learning for language understanding”. 2020.

[9] Tianyu Gao et al. “SimCSE: Simple Contrastive Learning of Sentence Embeddings”. 2021.

[10] Spyros Gidaris et al. “Unsupervised representation learning by predicting image rotations”. 2018.

[11] John M Giorgi et al. “Declutr: Deep contrastive learning for unsupervised textual representations”. 2020.

For Further Reading II

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

[13] Kaiming He et al. “Masked autoencoders are scalable vision learners”. 2022.

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

[15] Tongzhou Wang and Phillip Isola. “Understanding contrastive representation learning through alignment and uniformity on the hypersphere”. 2020.

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

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

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

[19] Richard Zhang et al. “Colorful image colorization”. 2016.