Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss

Mehmet F. Demirel - Nov 3, 2021

SSL

- Data augmentation to get two views of the same image.
- Despite empirical success, theoretical understanding lacks.
- Some papers (e.g. Arora et al. (2019)) provide guarantees under the assumption that two views are somewhat conditionally independent given the label or a hidden variable.
- However, this is not true in practical cases, i.e., two views are augmentations of a natural image, and usually have strong correlation.
    - They are not independent conditioned on the label.

This paper

- Presents a theoretical framework for self-supervised learning without requiring conditional independence
- Designs a principled, practical loss function for learning neural net representations that resembles state-of-the-art contrastive learning methods.
- Proves that, under a simple and realistic data assumption, linear classification using representations learned on a polynomial number of unlabeled data samples can recover the ground-truth labels of the data with high accuracy.

Idea

- “The fundamental data property that we leverage is a notion of continuity of the population data within the same class. Though a random pair of examples from the same class can be far apart, the pair is often connected by (many) sequences of examples, where consecutive examples in the sequences are close neighbors within the same class. This property is more salient when the neighborhood of an example includes many different types of augmentations.”

Setting

- $\bar{X}$: set of all natural data (raw input without augmentation)
  - Assumed to be finite but exponentially large
- Each $\bar{x} \in \bar{X}$ belongs to one of r classes
- $y : \bar{X} \to [r]$ is the labeling function
- $P_{\bar{X}}$: population distribution over $\bar{X}$ from which we draw training data.
- $A(\cdot \mid \bar{x})$: distribution of the augmentations of a given $\bar{x} \in \bar{X}$
- X : set of all augmented data
  - Assumed to be finite but exponentially large, $|X| = N$
- We will learn an embedding function $f : X \to \mathbb{R}^k$ and evaluate its quality using linear probe g with weights $B \in \mathbb{R}^{k \times r}$
  - $g_{f,B}(x) = \arg\max_{i \in [r]} \; (f(x)^\top B)_i$
  - Then given raw data $\bar{x}$, we ensemble the predictions on augmented data and predict $\bar{g}_{f,B}(\bar{x}) = \arg\max_{i \in [r]} \; \underset{x \sim A(\cdot \mid \bar{x})}{P} [g_{f,B}(x) = i]$
- $\mathcal{E}(f) = \min_{B \in \mathbb{R}^{k \times r}} \; \underset{\bar{x} \sim P_{\bar{X}}}{Pr} [y(\bar{x}) \neq \bar{g}_{f,B}(\bar{x})]$: linear probe error (the error of the best possible linear classifier on the representations)

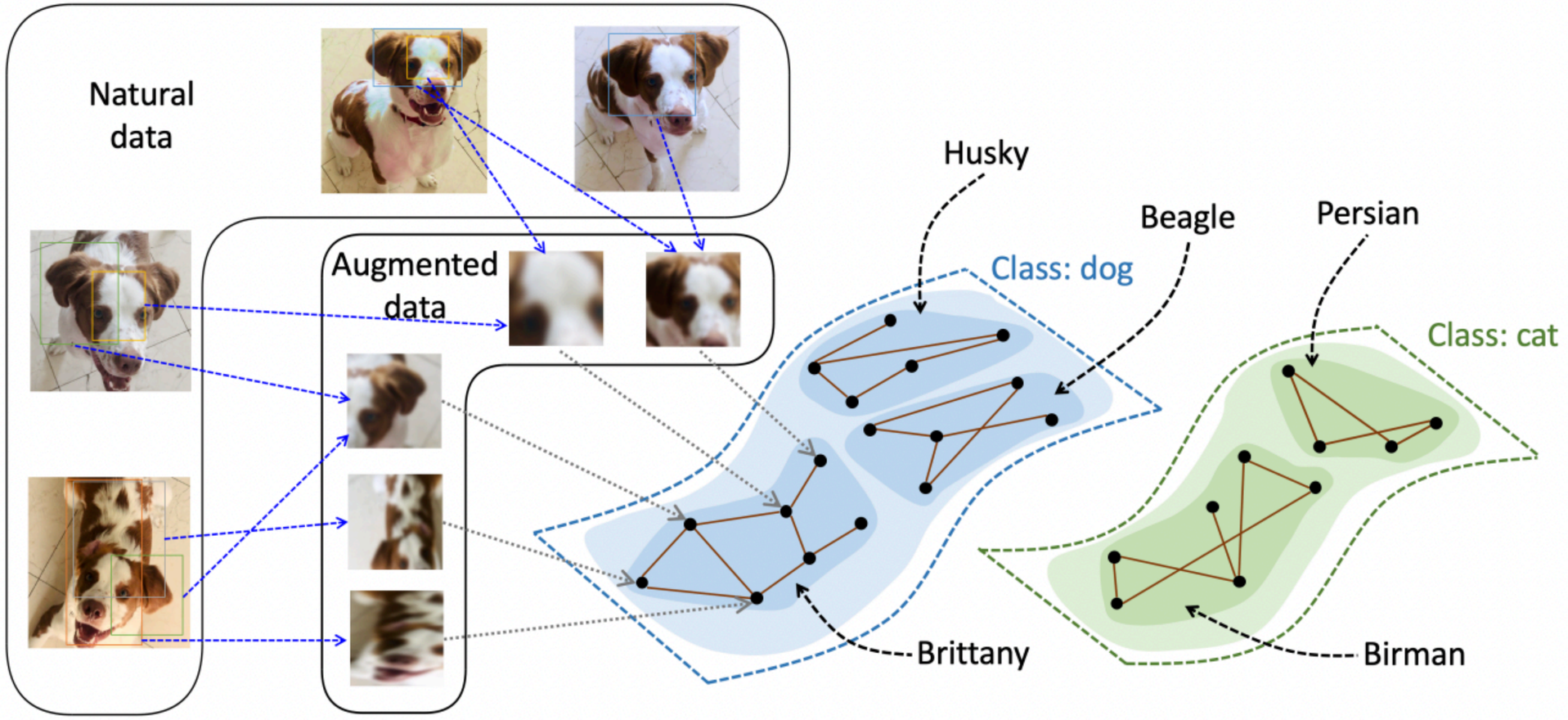
Augmentation Graph

- Population augmentation graph $G(X, w)$, where the vertex set is all augmentation data X, and w denotes the edge weights.
- For any two augmented data $x, x' \in X$, $w_{xx'}$ denotes the marginal probability of generating the pair x and x' from a random natural data $\bar{x} \sim P_{\bar{X}}$:

$$w_{xx'} = \mathbb{E}_{\bar{x} \sim P_{\bar{X}}}[A(x \mid \bar{x}) A(x' \mid \bar{x})]$$

- $$\sum_{x, x' \in X} w_{xx'} = 1$$

Augmentation Graph



Spectral Decomposition

- $w_x = \sum_{x' \in X} w_{xx'}$
- $A \in \mathbb{R}^{N \times N}$ is adjacency matrix with $A_{xx'} = w_{xx'}$
- $D \in \mathbb{R}^{N \times N}$ is a diagonal matrix with $D_{xx} = w_x$
- $\bar{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ is the normalized adjacency matrix.

Spectral Decomposition

Standard spectral graph theory approaches produce vertex embeddings as follows. Let $\gamma_1, \gamma_2, \cdots, \gamma_k$ be the k largest eigenvalues of $\overline{A}$, and $v_1, v_2, \cdots, v_k$ be the corresponding unit-norm eigenvectors. Let $F^* = [v_1, v_2, \cdots, v_k] \in \mathbb{R}^{N \times k}$ be the matrix that collects these eigenvectors in columns, and we refer to it as the eigenvector matrix. Let $u_x^* \in \mathbb{R}^k$ be the x-th row of the matrix F^*. It turns out that u_x^*'s can serve as desirable embeddings of x's because they exhibit clustering structure in Euclidean space that resembles the clustering structure of the graph $G(\mathcal{X}, w)$.

Spectral Contrastive Learning

Standard spectral graph theory approaches produce vertex embeddings as follows. Let $\gamma_1, \gamma_2, \cdots, \gamma_k$ be the k largest eigenvalues of $\overline{A}$, and $v_1, v_2, \cdots, v_k$ be the corresponding unit-norm eigenvectors. Let $F^* = [v_1, v_2, \cdots, v_k] \in \mathbb{R}^{N \times k}$ be the matrix that collects these eigenvectors in columns, and we refer to it as the eigenvector matrix. Let $u_x^* \in \mathbb{R}^k$ be the x-th row of the matrix F^*. It turns out that u_x^*'s can serve as desirable embeddings of x's because they exhibit clustering structure in Euclidean space that resembles the clustering structure of the graph $G(\mathcal{X}, w)$.

- The embeddings u_x^* obtained by eigendecomposition are non-parametric.
- The embedding matrix F^* cannot be stored efficiently.
- **Then, parametrize the rows of F^* as a neural net function, and assume that embeddings u_x^* can be represented by $f(x)$ for some $f \in \mathcal{F}$**

Spectral Contrastive Learning

Standard spectral graph theory approaches produce vertex embeddings as follows. Let $\gamma_1, \gamma_2, \cdots, \gamma_k$ be the k largest eigenvalues of $\overline{A}$, and $v_1, v_2, \cdots, v_k$ be the corresponding unit-norm eigenvectors. Let $F^* = [v_1, v_2, \cdots, v_k] \in \mathbb{R}^{N \times k}$ be the matrix that collects these eigenvectors in columns, and we refer to it as the eigenvector matrix. Let $u_x^* \in \mathbb{R}^k$ be the x-th row of the matrix F^*. It turns out that u_x^*'s can serve as desirable embeddings of x's because they exhibit clustering structure in Euclidean space that resembles the clustering structure of the graph $G(\mathcal{X}, w)$.

- Design a loss function for f such that minimizing it could recover F^* up to some linear transformation.

$$\min_{F \in \mathbb{R}^{N \times k}} \mathcal{L}_{\mathrm{mf}}(F) := \left\| \overline{A} - FF^\top \right\|_F^2 .$$

Spectral Contrastive Learning

By the classical theory on low-rank approximation (Eckart–Young–Mirsky theorem (Eckart and Young, 1936)), any minimizer $\widehat{F}$ of $\mathcal{L}_{\text{mf}}(F)$ contains scaling of the largest eigenvectors of $\overline{A}$ up to a right transformation—for some orthonormal matrix $R \in \mathbb{R}^{k \times k}$, we have $\widehat{F} = F^* \cdot \text{diag}([\sqrt{\gamma_1}, \ldots, \sqrt{\gamma_k}])R$. Fortunately, multiplying the embedding matrix by any matrix on the right and any diagonal matrix on the left does not change its linear probe performance, which is formalized by the following lemma.

Spectral Contrastive Learning

Lemma 3.1. *Consider an embedding matrix $F \in \mathbb{R}^{N \times k}$ and a linear classifier $B \in \mathbb{R}^{k \times r}$. Let $D \in \mathbb{R}^{N \times N}$ be a diagonal matrix with positive diagonal entries and $Q \in \mathbb{R}^{k \times k}$ be an invertible matrix. Then, for any embedding matrix $\widetilde{F} = D \cdot F \cdot Q$, the linear classifier $\tilde{B} = Q^{-1}B$ on $\widetilde{F}$ has the same prediction as B on F. As a consequence, we have*

$$\mathcal{E}(F) = \mathcal{E}(\widetilde{F}). \tag{5}$$

where $\mathcal{E}(F)$ denotes the linear probe performance when the rows of F are used as embeddings.

Proof of Lemma 3.1. Let $D = \mathrm{diag}(s)$ where $s_x > 0$ for $x \in \mathcal{X}$. Let $u_x, \tilde{u}_x \in \mathbb{R}^k$ be the x-th row of matrices F and $\widetilde{F}$, respectively. Recall that $g_{u,B}(x) = \arg\max_{i \in [r]} (u_x^\top B)_i$ is the prediction on an augmented datapoint $x \in \overline{\mathcal{X}}$ with representation u_x and linear classifier B. Let $\widetilde{B} = Q^{-1}B$, it's easy to see that $g_{\tilde{u},\widetilde{B}}(x) = \arg\max_{i \in [r]} (s_x \cdot u_x^\top B)_i$. Notice that $s_x > 0$ doesn't change the prediction since it changes all dimensions of $u_x^\top B$ by the same scale, we have $g_{\tilde{u},\widetilde{B}}(x) = g_{u,B}(x)$ for any augmented datapoint $x \in \mathcal{X}$. The equivalence of loss naturally follows. □

Spectral Contrastive Learning

$$\min_{F \in \mathbb{R}^{N \times k}} \mathcal{L}_{\mathrm{mf}}(F) := \left\| \overline{A} - FF^\top \right\|_F^2 .$$

- $\mathscr{L}_{mf}$ is based on the rows u_x of F.
- Then $(FF^\top)_{xx'} = u_x^\top u_{x'}$
- $\mathscr{L}_{mf}$ can be decomposed into a sum of N^2 terms with $u_x^\top u_{x'}$

Spectral Contrastive Learning

- $\bar{x}, \bar{x}' \sim P_{\bar{X}}$ are a random raw data points
- Define positive pair (x, x^+) where $x, x^+ \sim A(\,\cdot\,|\,\bar{x})$
- Define negative pair (x, x^-) where $x^- \sim A(\,\cdot\,|\,\bar{x}')$

Spectral Loss

$$\min_{F \in \mathbb{R}^{N \times k}} \mathcal{L}_{\mathrm{mf}}(F) := \left\| \overline{A} - FF^\top \right\|_F^2 .$$

Lemma 3.2 (Spectral contrastive loss). *Recall that u_x is the x-th row of F. Let $u_x = w_x^{1/2} f(x)$ for some function f. Then, the loss function $\mathcal{L}_{\mathrm{mf}}(F)$ is equivalent to the following loss function for f, called spectral contrastive loss, up to an additive constant:*

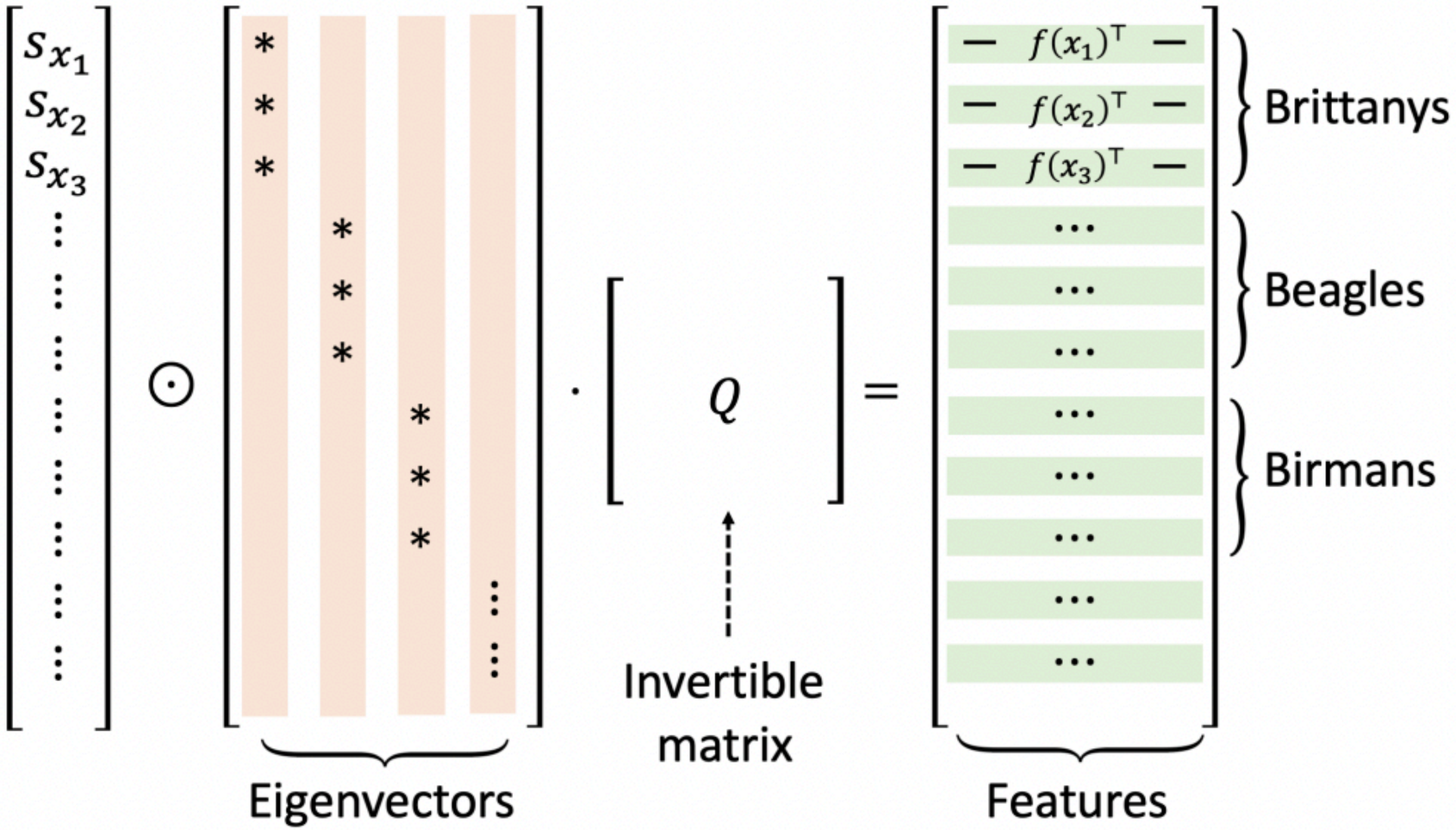
$$\mathcal{L}_{\mathrm{mf}}(F) = \mathcal{L}(f) + \text{const}$$

$$\text{where } \mathcal{L}(f) \triangleq -2 \cdot \mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] + \mathbb{E}_{x,x^-}\left[\left(f(x)^\top f(x^-)\right)^2\right] \tag{6}$$

Proof of Lemma 3.2. We can expand $\mathcal{L}_{\mathrm{mf}}(F)$ and obtain

$$\begin{aligned}
\mathcal{L}_{\mathrm{mf}}(F) &= \sum_{x,x' \in \mathcal{X}} \left(\frac{w_{xx'}}{\sqrt{w_x w_{x'}}} - u_x^\top u_{x'} \right)^2 \\
&= \sum_{x,x' \in \mathcal{X}} \left(\frac{w_{xx'}^2}{w_x w_{x'}} - 2 \cdot w_{xx'} \cdot f(x)^\top f(x') + w_x w_{x'} \cdot \left(f(x)^\top f(x') \right)^2 \right)
\end{aligned} \tag{7}$$

Spectral Loss



Guarantees for spectral loss on pop. data

Definition 3.3 (Dirichlet conductance). *For a graph $G = (\mathcal{X}, w)$ and a subset $S \subseteq \mathcal{X}$, we define the Dirichlet conductance of S as*

$$\phi_G(S) := \frac{\sum_{x \in S, x' \notin S} w_{xx'}}{\sum_{x \in S} w_x}.$$

We note that when S is a singleton, there is $\phi_G(S) = 1$ due to the definition of w_x. We introduce the sparsest m-partition to represent the number of edges between m disjoint subsets.

Definition 3.4 (Sparsest m-partition). *Let $G = (\mathcal{X}, w)$ be the augmentation graph. For an integer $m \in [2, |\mathcal{X}|]$, we define the sparsest m-partition as*

$$\rho_m := \min_{S_1, \cdots, S_m} \max\{\phi_G(S_1), \ldots, \phi_G(S_m)\}$$

where $S_1, \cdots, S_m$ are non-empty sets that form a partition of $\mathcal{X}$.

Guarantees for spectral loss on pop. data

Assumption 3.5 (Labels are recoverable from augmentations)**.** *Let $\bar{x} \sim \mathcal{P}_{\overline{\mathcal{X}}}$ and $y(\bar{x})$ be its label. Let the augmentation $x \sim \mathcal{A}(\cdot|\bar{x})$. We assume that there exists a classifier g that can predict $y(\bar{x})$ given x with error at most α. That is, $g(x) = y(\bar{x})$ with probability at least $1 - \alpha$.*

We also introduce the following assumption which states that some universal minimizer of the population spectral contrastive loss can be realized by the hypothesis class.

Assumption 3.6 (Realizability)**.** *Let $\mathcal{F}$ be a hypothesis class containing functions from $\mathcal{X}$ to $\mathbb{R}^k$. We assume that at least one of the global minima of $\mathcal{L}(f)$ belongs to $\mathcal{F}$.*

Theorem 3.7. *Assume the representation dimension $k \geq 2r$ and Assumption 3.5 holds for $\alpha > 0$. Let $\mathcal{F}$ be a hypothesis class that satisfies Assumption 3.6 and let $f^*_{\text{pop}} \in \mathcal{F}$ be a minimizer of $\mathcal{L}(f)$. Then, we have*

$$\mathcal{E}(f^*_{\text{pop}}) \leq \widetilde{O}\left(\alpha/\rho^2_{\lfloor k/2 \rfloor}\right).$$

Finite-sample generalization bounds

training dataset $\{\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n\}$ with $\bar{x}_i \sim \mathcal{P}_{\overline{\mathcal{X}}}$

$$\widehat{\mathcal{L}}_n(f) := -\frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{x\sim\mathcal{A}(\cdot|\bar{x}_i)\\ x^+\sim\mathcal{A}(\cdot|\bar{x}_i)}} \left[f(x)^\top f(x^+)\right] + \frac{1}{n(n-1)}\sum_{i\neq j} \mathbb{E}_{\substack{x\sim\mathcal{A}(\cdot|\bar{x}_i)\\ x^-\sim\mathcal{A}(\cdot|\bar{x}_j)}} \left[\left(f(x)^\top f(x^-)\right)^2\right].$$

unbiased estimator of $\mathscr{L}(f)$ where $\mathcal{L}(f) \triangleq -2\cdot \mathbb{E}_{x,x^+}\left[f(x)^\top f(x^+)\right] + \mathbb{E}_{x,x^-}\left[\left(f(x)^\top f(x^-)\right)^2\right]$

Finite-sample generalization bounds

define the Rademacher complexity of $\mathcal{F}$ on n data as

$\widehat{\mathcal{R}}_n(\mathcal{F}) := \max_{x_1,\cdots,x_n\in\mathcal{X}} \mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F},i\in[k]} \frac{1}{n}\left(\sum_{j=1}^n \sigma_j f_i(x_j)\right)\right]$, where σ is a uniform random vector in $\{-1,1\}^n$ and $f_i(z)$ is the i-th dimension of $f(z)$.

Theorem 4.1. *For some $\kappa > 0$, assume $\|f(x)\|_\infty \le \kappa$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$. Let $f^*_{\text{pop}} \in \mathcal{F}$ be a minimizer of the population loss $\mathcal{L}(f)$. Given a random dataset of size n, let $\hat{f}_{\text{emp}} \in \mathcal{F}$ be a minimizer of empirical loss $\widehat{\mathcal{L}}_n(f)$. Then, when Assumption 3.6 holds, with probability at least $1-\delta$ over the randomness of data, we have*

$$\mathcal{L}(\hat{f}_{\text{emp}}) \le \mathcal{L}(f^*_{\text{pop}}) + c_1 \cdot \widehat{\mathcal{R}}_{n/2}(\mathcal{F}) + c_2 \cdot \left(\sqrt{\frac{\log 2/\delta}{n}} + \delta\right),$$

where constants $c_1 \lesssim k^2\kappa^2 + k\kappa$ and $c_2 \lesssim k\kappa^2 + k^2\kappa^4$.

Finite-sample generalization bounds

Theorem 4.2. *Assume representation dimension $k \geq 4r + 2$, Assumption 3.5 holds for $\alpha > 0$ and Assumption 3.6 holds. Recall γ_i be the i-th largest eigenvalue of the normalized adjacency matrix. Then, for any $\epsilon < \gamma_k^2$ and $\hat{f}_{\text{emp}} \in \mathcal{F}$ such that $\mathcal{L}(\hat{f}_{\text{emp}}) < \mathcal{L}(f^*_{\text{pop}}) + \epsilon$, we have:*

$$\mathcal{E}(\hat{f}_{\text{emp}}) \lesssim \frac{\alpha}{\rho^2_{\lfloor k/2 \rfloor}} \cdot \log k + \frac{k\epsilon}{\Delta_\gamma^2},$$

where $\Delta_\gamma := \gamma_{\lfloor 3k/4 \rfloor} - \gamma_k$ is the eigenvalue gap between the $\lfloor 3k/4 \rfloor$-th and the k-th eigenvalue.

Guarantee for learning linear probe with labeled data

$$\ell((z, y(\bar{x})), B) := \sum_{i=1}^{r} \min\left\{ \left(B^\top z - \vec{y}(\bar{x})\right)_i^2, 1 \right\}$$

Theorem 5.1. *In the setting of Theorem 3.7, assume $\gamma_k \geq C_\lambda$ for some $C_\lambda > 0$. Learn a linear probe $\widehat{B} \in \arg\min_{\|B\|_F \leq 1/C_\lambda} \sum_{i=1}^{n} \ell((f^*_{\text{pop}}(x_i), y(\bar{x}_i)), B)$ by minimizing the capped quadratic loss subject to a norm constraint. Then, with probability at least $1 - \delta$ over random data, we have*

$$\Pr_{\bar{x} \sim \mathcal{P}_{\bar{\mathcal{X}}}} \left(\bar{g}_{f^*_{\text{pop}}, \widehat{B}}(\bar{x}) \neq y(\bar{x}) \right) \lesssim \frac{\alpha}{\rho^2_{\lfloor k/2 \rfloor}} \cdot \log k + \frac{r}{C_\lambda} \cdot \sqrt{\frac{k}{n}} + \sqrt{\frac{\log 1/\delta}{n}}.$$

Here the first term is the population error from Theorem 3.7. The last two terms are the generalization gap from standard concentration inequalities for linear classification and are small when the number of labeled data n is polynomial in the feature dimension k. We note that this result reveals a trade-off when choosing the feature dimension k: when n is fixed, a larger k decreases the population contrastive loss while increases the generalization gap for downstream linear classification. The proof of Theorem 5.1 is in Section E.

Experiments

Datasets	CIFAR-10			CIFAR-100			Tiny-ImageNet		
Epochs	200	400	800	200	400	800	200	400	800
SimCLR (repro.)	83.73	87.72	90.60	54.74	61.05	63.88	**43.30**	**46.46**	48.12
SimSiam (repro.)	87.54	**90.31**	91.40	61.56	64.96	65.87	34.82	39.46	46.76
Ours	**88.66**	90.17	**92.07**	**62.45**	**65.82**	**66.18**	41.30	45.36	**49.86**

Table 1: Top-1 accuracy under linear evaluation protocal.

	SimCLR	BYOL	MoCo v2	SimSiam	Ours
acc. (%)	66.5	66.5	67.4	68.1	66.97

Table 2: ImageNet linear evaluation accuracy with 100-epoch pre-training. All results but ours are reported from (Chen and He, 2020). We use batch size 384 during pre-training.