On the Similarities of Embeddings in Contrastive Learning

Chungpa Lee, Sehee Lim, Kibok Lee, Jy-yong Sohn

YONSEI UNIVERSITY

ICML International Conference On Machine Learning

To understand contrastive learning, we analyze the *embeddings of positive and negative pairs* through *the lens of cosine similarity*. In full-batch settings, perfect alignment of positive pairs is *unattainable* when the similarities of negative pairs fall below a threshold. This misalignment can be mitigated by incorporating within-view negative pairs into the loss. In mini-batch settings, smaller batch sizes lead to the *increased variance* in the similarities of negative pairs—a distinctive characteristic absent in full-batch settings and a potential contributor to performance degradation in mini-batch settings. To explore this, we introduce an auxiliary loss that reduces this variance, leading to improved performance in small-batch settings.

Contrastive Learning (CL)

In CL, a normalized encoder $f \in \mathbb{R}^d$ is trained so that:

- embeddings of ***positive pairs*** $(\mathbf{u}_i, \mathbf{v}_i) = (f(\mathbf{x}_i), f(\mathbf{y}_i))$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are augmented views of the same instance, are mapped into similar embeddings (i.e., $\mathbf{u}_i \approx \mathbf{v}_i$),
- ***negative pairs*** $(\mathbf{u}_i, \mathbf{v}_j)$ where $i \neq j$ are pushed apart.

Two formulations of contrastive losses used in practice.

Def.3.1. The InfoNCE-Based Loss $\mathscr{L}_{\text{info-sym}}(\mathbf{U}, \mathbf{V})$:

$$\sum_{i\in[n]} \psi\Big(c_1 \sum_{j\in[n]\setminus\{i\}} \phi\big((\mathbf{v}_j - \mathbf{v}_i)^\top \mathbf{u}_i\big) + c_2 \sum_{j\in[n]\setminus\{i\}} \phi\big((\mathbf{u}_j - \mathbf{v}_i)^\top \mathbf{u}_i\big)\Big)$$

$$+\sum_{i\in[n]} \psi\Big(c_1 \sum_{j\in[n]\setminus\{i\}} \phi\big((\mathbf{u}_j - \mathbf{u}_i)^\top \mathbf{v}_i\big) + c_2 \sum_{j\in[n]\setminus\{i\}} \phi\big((\mathbf{v}_j - \mathbf{u}_i)^\top \mathbf{v}_i\big)\Big)$$

for some constants $c_1, c_2 \in \{0,1\}$,

where ϕ and ψ are some convex and increasing functions.

Ex. InfoNCE (Oord et al., Representation learning with contrastive predictive coding. arXiv, 2018.), SimCLR (Chen et al. A simple framework for contrastive learning of visual representations. ICML, 2020.), DCL (Yeh et al., Decoupled contrastive learning. ECCV, 2022.), and DHEL (Koromilas et al., Bridging mini-batch and asymptotic analysis in contrastive learning. ICML, 2024.).

Def.3.2. The Independently Additive Loss $\mathscr{L}_{\text{ind-add}}(\mathbf{U}, \mathbf{V})$:

$$-\sum_{i\in[n]} \phi(\mathbf{u}_i^\top \mathbf{v}_i) + c_1 \sum_{i\neq j\in[n]} \psi(\mathbf{u}_i^\top \mathbf{v}_j) + c_2 \sum_{i\neq j\in[n]} \big(\psi(\mathbf{u}_i^\top \mathbf{u}_j) + \psi(\mathbf{v}_i^\top \mathbf{v}_j)\big)$$

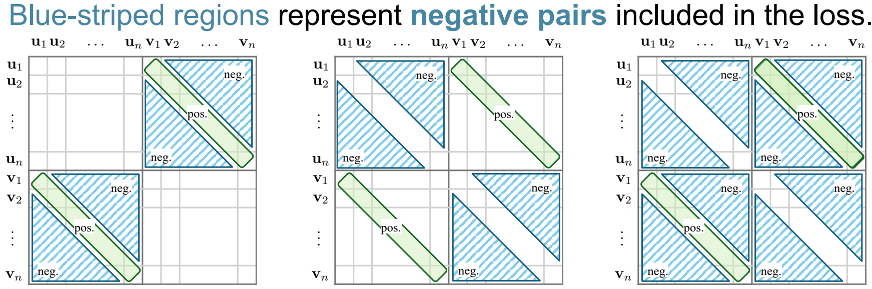
for some constants $c_1, c_2 \in \{0,1\}$,

where ϕ : concave, increasing, and ψ : convex, increasing.

Ex. SigLIP (Zhai et al., Sigmoid loss for language image pre-training. ICCV, 2023) and Spectral CL (HaoChen et al., Provable guarantees for self-supervised deep learning with spectral contrastive loss. NeurIPS, 2021.).

Negative pair considered in the loss formulations, by (c_1, c_2).

Each grid shows all possible pairs of embeddings in $\mathbf{U}$ and $\mathbf{V}$. Blue-striped regions represent **negative pairs** included in the loss.



(a) $(c_1, c_2) = (1,0)$ Cross-view negatives. (b) $(c_1, c_2) = (0,1)$ Within-view negatives. (c) $(c_1, c_2) = (1,1)$ All negatives.

Cosine Similarity of Embeddings

Similarities between embeddings of a positive / negative pair.

Def.4.1. The ***positive-pair similarity*** for the encoder f is

$$s_{\text{pos}}(f) := f(\mathbf{x})^\top f(\mathbf{y}) \quad \text{for } (\mathbf{x}, \mathbf{y}) \sim \hat{p}_{\text{pos}},$$

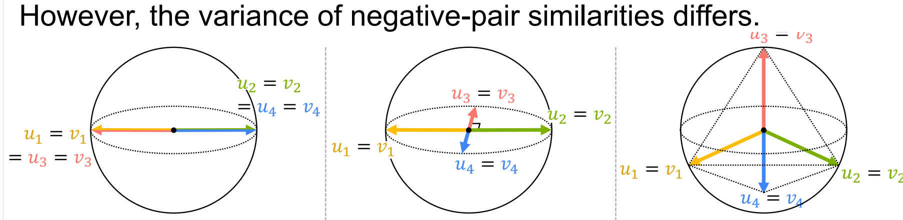
and the ***negative-pair similarity*** for the encoder f is

$$s_{\text{neg}}(f) := f(\mathbf{x})^\top f(\mathbf{y}) \quad \text{for } (\mathbf{x}, \mathbf{y}) \sim \hat{p}_{\text{neg}}.$$

Three examples of 8 embeddings.

In all cases, $s_{\text{pos}}(f) = 1$ and $\mathbb{E}[s_{\text{neg}}(f)] = -1/3$.

However, the variance of negative-pair similarities differs.



(a) $\text{Var}[s_{\text{neg}}(f)] = 8/9$ (b) $\text{Var}[s_{\text{neg}}(f)] = 2/9$ (c) $\text{Var}[s_{\text{neg}}(f)] = 0$

Embedding Learned in Full-Batch CL

Thm.5.2. For any normalized encoder f,

$$\mathbb{E}[s_{\text{pos}}(f)] \leq 1 + \big(\mathbb{E}[s_{\text{neg}}(f)] + 1/(n-1)\big),$$

where n is the size of the training dataset.

⚠️ **Excessive separation of negative pairs in full-batch CL:** When the average of negative-pair similarities drops below $-1/(n-1)$, positive pairs cannot be fully aligned ($\mathbb{E}[s_{\text{pos}}(f)] < 1$).

Define the optimal encoder as $f^\star := \arg\min_f \mathbb{E}\big[\mathscr{L}(\mathbf{U}, \mathbf{V})\big]$.

Thm.5.3. If we use the loss of $\mathscr{L}_{\text{ind-add}}(\mathbf{U}, \mathbf{V})$ with $(c_1, c_2) = (1,0)$ and $\phi'(1) \ll \psi'(-1/(n-1))$, then

$$s_{\text{pos}}(f^\star) < 1 \quad \text{and} \quad s_{\text{neg}}(f^\star) < -1/(n-1).$$

Thm.5.1. If we use the loss of (1) $\mathscr{L}_{\text{info-sym}}(\mathbf{U}, \mathbf{V})$ or (2) $\mathscr{L}_{\text{ind-add}}(\mathbf{U}, \mathbf{V})$ with $(c_1, c_2) \in \{(0,1), (1,1)\}$, then

$$s_{\text{pos}}(f^\star) = 1 \quad \text{and} \quad s_{\text{neg}}(f^\star) = -1/(n-1).$$

✅ **Including the within-view negative pairs** (by set $c_2 = 1$ of $\mathscr{L}_{\text{info-sym}}(\mathbf{U}, \mathbf{V})$) **mitigates the misalignment** of positive pairs.

Embedding Learned in Mini-Batch CL

For the batch size m, define the mini-batch optimal encoder as

$$f^\star_{\text{batch}} := \arg\min_f \mathbb{E}\Big[\mathscr{L}(\mathbf{U}_1, \mathbf{V}_1) + \cdots + \mathscr{L}(\mathbf{U}_b, \mathbf{V}_b)\Big],$$

where $\cup_{k\in[b]} \mathbf{U}_k = \mathbf{U}$, $\cup_{k\in[b]} \mathbf{V}_k = \mathbf{V}$, and $|\mathbf{U}_k| = |\mathbf{V}_k| = m$.

Thm.5.5. If we use the loss of (1) $\mathscr{L}_{\text{info-sym}}(\mathbf{U}, \mathbf{V})$ or (2) $\mathscr{L}_{\text{ind-add}}(\mathbf{U}, \mathbf{V})$ with $(c_1, c_2) \in \{(0,1), (1,1)\}$, then

$$s_{\text{pos}}(f^\star_{\text{batch}}) = 1, \quad \mathbb{E}[s_{\text{neg}}(f^\star_{\text{batch}})] = -1/(n-1),$$

and $\text{Var}\big[s_{\text{neg}}(f^\star_{\text{batch}})\big] = O(1/m)$.

⚠️ **Excessive separation of negative pairs in mini-batch CL:** The effect of using mini-batch is in the increased variance of negative-pair similarities ($\text{Var}[s_{\text{neg}}(f^\star_{\text{batch}})] = O(1/m)$), caused by stronger separation among negative pairs within each batch.

Def.5.7. For the batch size m, define the auxiliary loss as

$$\mathscr{L}_{\text{VRNS}}(\mathbf{U}_k, \mathbf{V}_k) := \frac{1}{m(m-1)} \sum_{i\neq j\in[m]} \Big(\mathbf{u}_i^\top \mathbf{v}_j + \frac{1}{n-1}\Big)^2.$$

✅ **We introduce an auxiliary loss term $\mathscr{L}_{\text{VRNS}}(\mathbf{U}, \mathbf{V})$ which reduces the variance of negative-pair similarities.**

Empirical Validation

Excessive Separation of Negative Pairs in mini-batch CL

ResNet-18 on CIFAR-10, varying batch sizes.

Batch size	Variance of negative-pair similarities	
	SimCLR	SimCLR + Ours
32	0.1649	0.1008
64	0.1505	0.0952
128	0.1444	0.0929
256	0.1404	0.0921
512	0.1396	0.0917

Effect of Variance Reduction on Classification Performance