Theoretical Insights into Self-Supervised Contrastive Learning

Weiran Huang

Associate professor

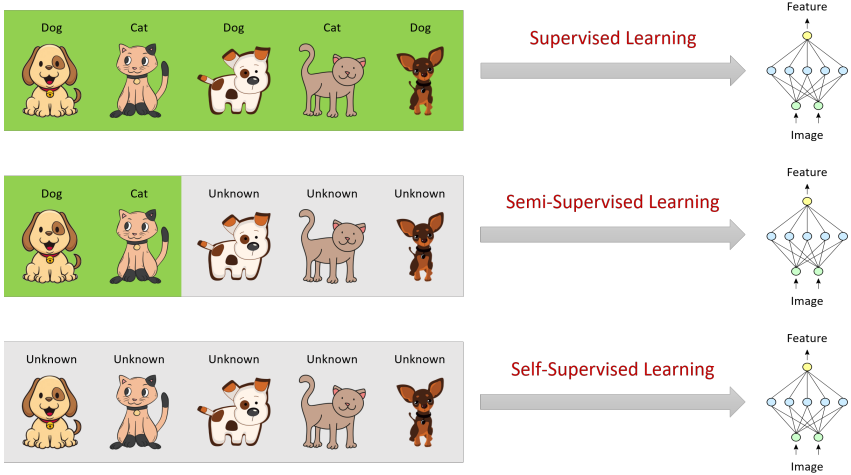
Qing Yuan Research Institute

Shanghai Jiao Tong University

December, 2023

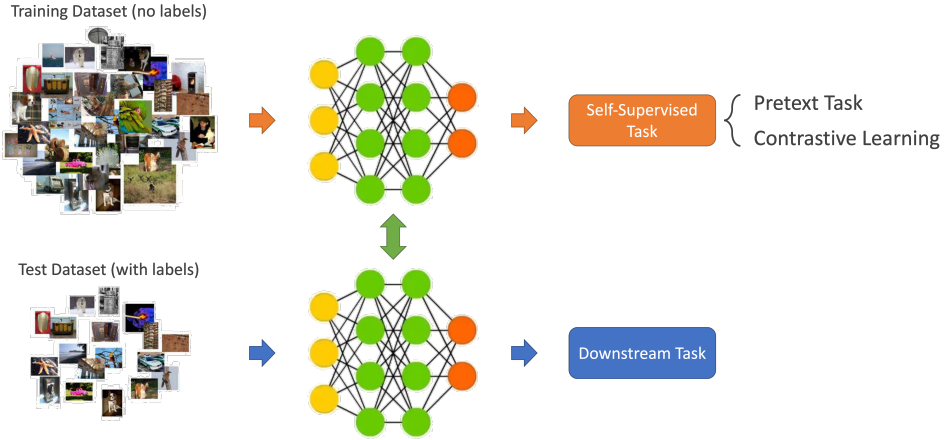
Introduction to Self-Supervised Contrastive Learning

Representation Learning Paradigm Evolution



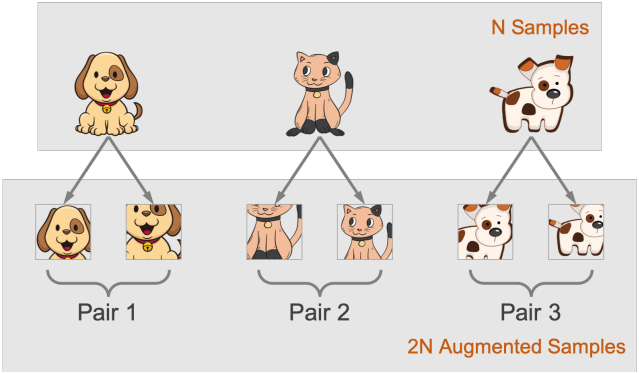
Self-Supervised Learning

Self-Supervised Learning learns data representations through manually designed supervision signals, and then uses the learned representations for downstream tasks.



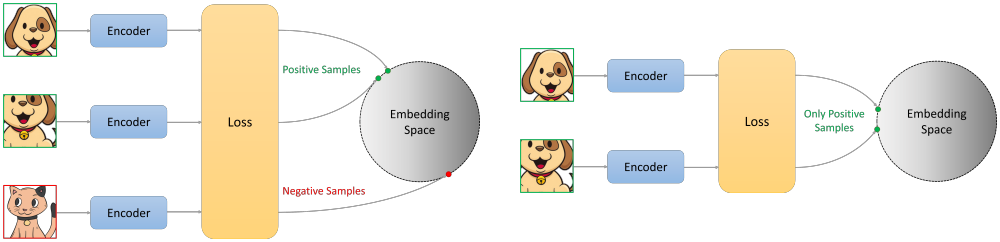
Self-Supervised Contrastive Learning

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



Self-Supervised Contrastive Learning

Step 2 of 2: Pull the similar sample pairs close to each other in the embedding space.



The objectives of most contrastive learning algorithms (including SimCLR, MoCo, Barlow Twins, etc.) can be re-formulated as

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

Observations in Self-Supervised Contrastive Learning

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

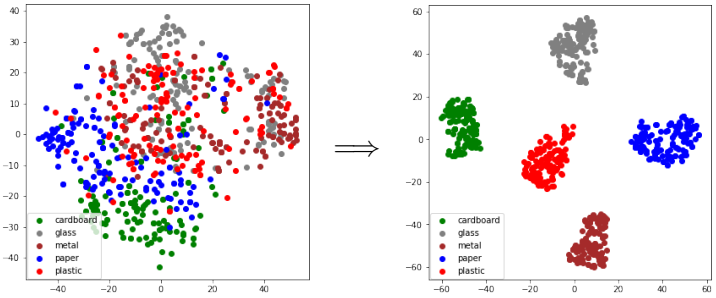
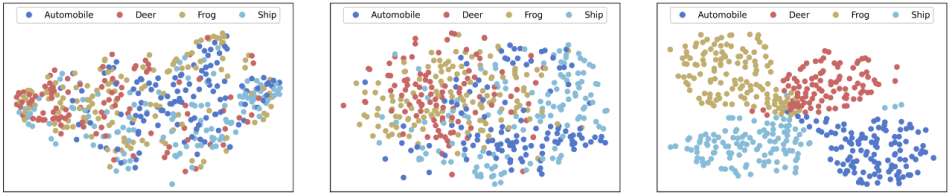


Figure: Embedding Space (https://github.com/mwdhont/SimCLRv1-keras-tensorflow).

Observations in Self-Supervised Contrastive Learning

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.

Observations in Self-Supervised Contrastive Learning

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

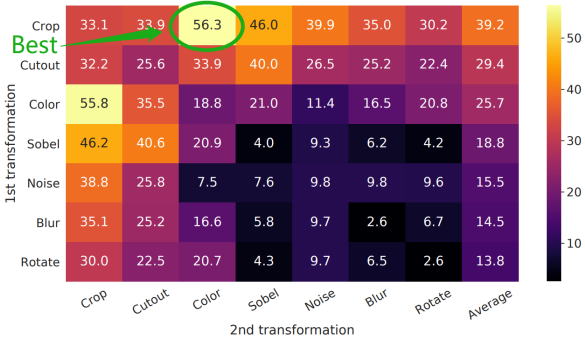


Figure: Experimental results reported in SimCLR paper.

Observations in Self-Supervised Contrastive Learning

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

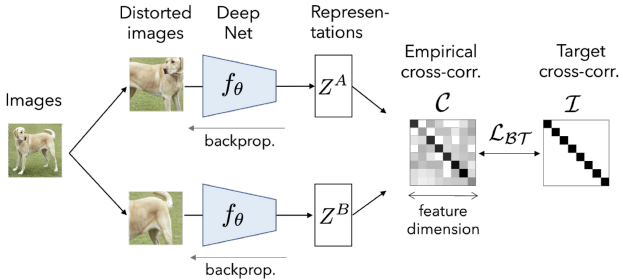


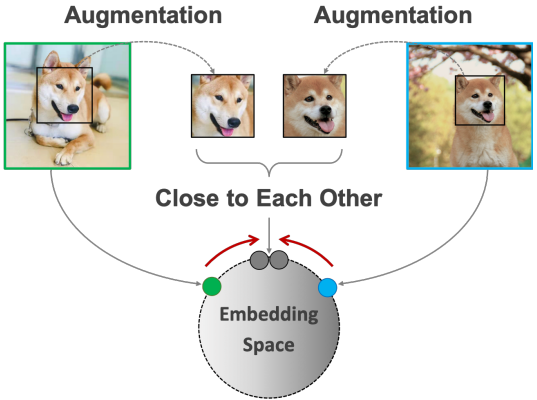
Figure: Barlow Twins aims to decorrelate the components of representation.

Theoretical Analysis of Self-Supervised Contrastive Learning

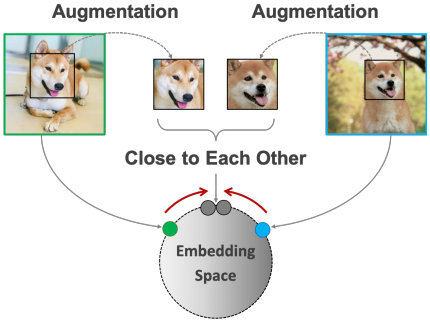
"Towards the Generalization of Contrastive Self-Supervised Learning."
Huang[*,†], Yi[*], Zhao[*], Jiang. ICLR 2023.

Intuition

Why does contrastive learning work?



Intuition



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

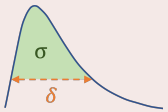
$$d_A(\mathbf{x}_1, \mathbf{x}_2) = \min_{\mathbf{x}_1' \in A(\mathbf{x}_1), \mathbf{x}_2' \in A(\mathbf{x}_2)} \left\| \mathbf{x}_1' - \mathbf{x}_2' \right\|.$$

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



The sharpness of concentration:

- Larger σ and smaller δ indicate the sharper concentration of augmented data.
- Richer data augmentation leads to sharper concentration.

Performance Guarantee of Self-Supervised Contrastive Learning

Theorem 1

Under mild assumptions, 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 $\Delta_\mu = 1 - \min_{k \in [K]} \frac{\|\mu_k\|^2}{r^2}$.

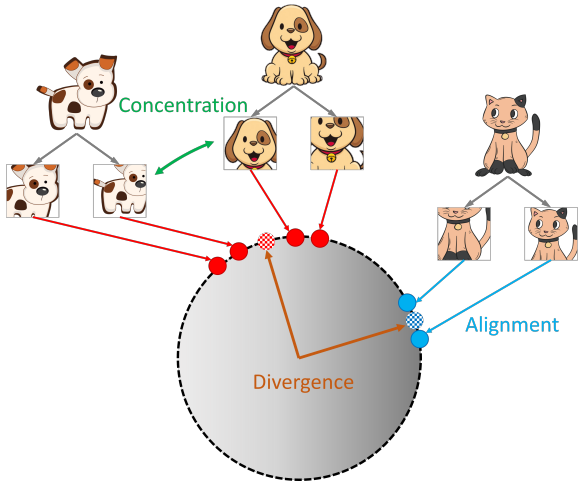
Messages From Theorem 1

The generalization ability depends on three key factors:

1. (**Alignment of positive samples**) How close positive samples are located to each other in the embedding space;
2. (**Divergence of class centers**) How far apart class centers are located from each other in the embedding space;
3. (**Concentration of augmented data**) How sharp the concentration of augmented data is.

Only the first two factors can be optimized during the learning process. In contrast, the third factor is priorly decided by the pre-defined data augmentation and is independent of the learning process.

Messages From Theorem 1



Loss Functions

- InfoNCE (e.g., SimCLR): 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): 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.

Loss Functions

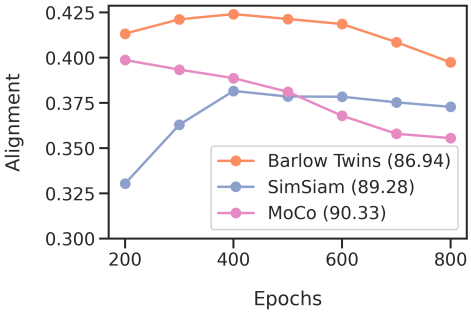
The above two losses can be split into two parts:

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

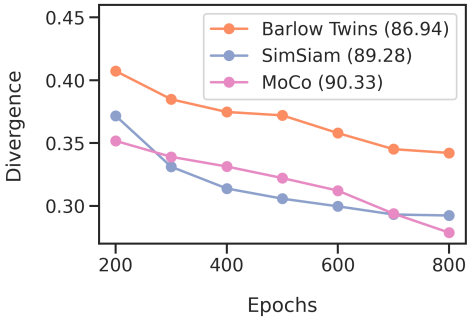
- For InfoNCE, we prove that $\mu_k^\top \mu_\ell \lesssim \mathcal{L}_{\text{regularization}}(f)$;
- For Cross-Correlation, we prove that $\mu_k^\top \mu_\ell \lesssim \sqrt{\mathcal{L}_{\text{regularization}}(f)}$.

Therefore, minimizing these two losses can achieve good alignment and large divergence.

How Alignment and Divergence Change During Training Process



(a) Alignment change of different algorithms



(b) Divergence change of different algorithms

Follow-Up

A Cookbook of Self-Supervised Learning

Randall Balestriero*, Mark Ibrahim*, *Vlad Sobal*, Ari Morcos*, Shashank Shekhar*, Tom Goldstein†, Florian Bordes*†, Adrien Bardes*, Gregoire Mialon*, Yuandong Tian*, Avi Schwarzschild†, Andrew Gordon Wilson**, Jonas Geiping†, Quentin Garrido*‡, Pierre Fernandez*§, Amir Bar*, Hamed Pirsiavash+, Yann LeCun* and Micah Goldblum**

*Meta AI, FAIR
**New York University
†University of Maryland
+University of California, Davis
‡Universite de Montreal, Mila
§Univ Gustave Eiffel, CNRS, LIGM
*Univ. Rennes, Inria, CNRS, IRISA
italicEqual contributions, randomized ordering

Contents

2.6.1 Theoretical Study of SSL

Numerous works have attempted to unify various SSL methods. In Huang et al. [2021], Barlow Twins' criterion is shown to be linked to an upper bound of a contrastive loss. This suggests a link exists between contrastive and covariance-based methods. This direction was further pursued in Garrido et al. [2022b], where a covariance-based and contrastive criterion are shown to be equivalent up to normalization by deriving the precise gap between the two approaches. These results were further validated empirically as methods were shown to exhibit similar performance and representation properties at ImageNet's scale (1.2 million samples). The similarities among methods was also studied in Tao et al. [2021] where this unification was tackled from a study of the losses' gradients.

The Effect of Concentration Factor

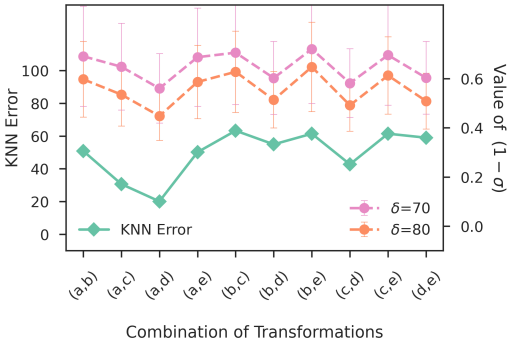
Dataset	Transformations (a)	(b)	(c)	(d)	(e)	Accuracy SimCLR	Barlow Twins	MoCo	SimSiam
CIFAR-10	✓	✓	✓	✓	✓	**89.76 ± 0.12**	**86.91 ± 0.09**	**90.12 ± 0.12**	**90.59 ± 0.11**
	✓	✓	✓	✓		88.48 ± 0.22	85.38 ± 0.37	89.69 ± 0.11	89.34 ± 0.09
	✓	✓	✓			83.50 ± 0.14	82.00 ± 0.59	86.78 ± 0.07	85.38 ± 0.09
	✓	✓				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.

The Effect of Concentration Factor

Dataset	Color Distortion Strength	Accuracy			
		SimCLR	Barlow Twins	MoCo	SimSiam
CIFAR-10	1	**82.75 ± 0.24**	**82.58 ± 0.25**	**86.68 ± 0.05**	**82.50 ± 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	**46.67 ± 0.42**	**50.39 ± 1.09**	**58.50 ± 0.51**	**49.94 ± 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

The Effect of Concentration Factor



(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.

Short Summary

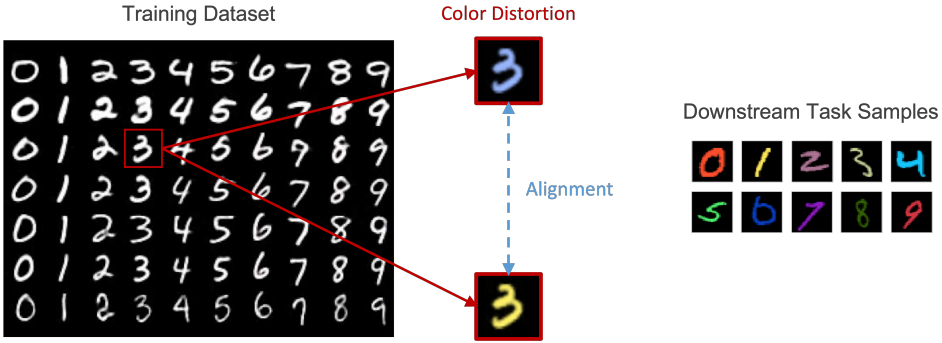
- We provide a mathematical formulation to model the data augmentation.
- We show that alignment of positive samples, divergence of class centers and concentration of augmented data are three key factors of self-supervised contrastive learning.
- We prove that SimCLR and Barlow Twins implicitly optimize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

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

Transferability of Self-Supervised Contrastive Learning

"ArCL: Enhancing Contrastive Learning with Augmentation-Robust Representations."
Zhao*, Du*, Wang, Yao, **Huang**†. ICLR 2023.

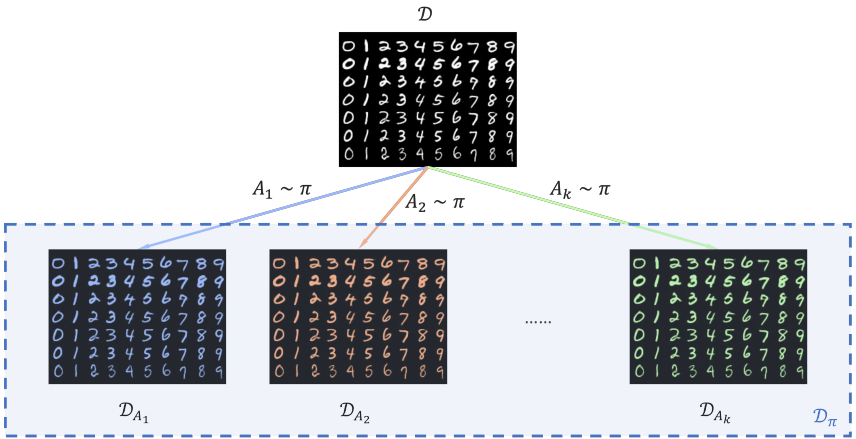
Paradox



Question: Can contrastive learning extract augmentation-invariant features?

Notations for Distributions

Let $\mathcal{D}$ denote the original upstream distribution, and $\mathcal{D}_A$ denote the augmented distribution after applying some random transformation A which follows distribution π.



Transferiability Evaluation

We first define "risk" of representation f over downstream distribution $\mathcal{D}^{\text{tar}}$ as

$$\mathcal{R}(f; \mathcal{D}^{\text{tar}}) := \min_{h} \mathbb{E}_{(X,Y)\sim\mathcal{D}^{\text{tar}}} \ell(h \cdot f(X), Y),$$

where h is a linear classifier and ℓ is the loss function.

For an augmentation-invariant representation, the risk should be at about the same level on two different downstream datasets $\mathcal{D}_{A_1}, \mathcal{D}_{A_2}$ augmented from $\mathcal{D}$.

In other words, $|\mathcal{R}(f; \mathcal{D}_{A_1}) - \mathcal{R}(f; \mathcal{D}_{A_2})|$ should be small.

A Counter-Example

Example 1

Data distribution $(X_1, X_2) \sim \mathcal{N}(0, I)$, $Y = \mathbb{1}(X_1 \geq 0)$.
Augmentation distribution $A_\theta(X_1, X_2) = (X_1, \theta \cdot X_2)$ where $\theta \sim \mathcal{N}(0, 1)$.
In this case, X_1 is the augmentation-invariant feature.

In fact, we can prove that

$$\forall \varepsilon > 0, \exists f, \mathcal{D}_{A_1}, \mathcal{D}_{A_2}, \text{ s.t. } \mathcal{L}_{\text{align}}(f; \mathcal{D}) < \varepsilon \text{ and } |\mathcal{R}(f; \mathcal{D}_{A_1}) - \mathcal{R}(f; \mathcal{D}_{A_2})| > \text{const.}$$

Proof (1/2)

For any $\varepsilon > 0$, let $f(x_1, x_2) = x_1 + \frac{\sqrt{\varepsilon}}{2} \cdot x_2$.

$$\begin{aligned}
\mathcal{L}_{\text{align}}(f; \mathcal{D}, \pi) &= \mathbb{E}_{X \sim \mathcal{D}} \mathbb{E}_{(A_1, A_2) \sim \pi^2} \|f(A_1(X)) - f(A_2(X))\|^2 \\
&= \mathbb{E}_{(X_1, X_2) \sim \mathcal{N}(0, I)} \mathbb{E}_{(\theta_1, \theta_2) \sim \mathcal{N}(0, I)} \left| (X_1 + \frac{\sqrt{\varepsilon}}{2} \theta_1 X_2) - (X_1 + \frac{\sqrt{\varepsilon}}{2} \theta_2 X_2) \right|^2 \\
&= \frac{\varepsilon}{4} \mathbb{E}_{X_2 \sim \mathcal{N}(0,1)} X_2^2 \mathbb{E}_{(\theta_1, \theta_2) \sim \mathcal{N}(0, I)} (\theta_1 - \theta_2)^2 = 2 \left(\frac{\sqrt{\varepsilon}}{2} \right)^2 < \varepsilon.
\end{aligned}$$

Proof (2/2)

Let $c = 0$ and $c' = 2/\sqrt{\varepsilon}$. Then we have two domains

$$\mathcal{D}_c = \{(X_1, 0) \colon X_1 \sim \mathcal{N}(0, 1)\}$$

$$\mathcal{D}_{c'} = \{(X_1, 2X_2/\sqrt{\varepsilon}) \colon X_1 \sim \mathcal{N}(0, 1), X_2 \sim \mathcal{N}(0, 1)\}$$

Therefore, we can get $\mathcal{R}(f; \mathcal{D}_c) = 0$, but

$$\begin{aligned}
\mathcal{R}(f; \mathcal{D}_{c'}) &= P(Y = 0, hf(X) \geq 0) + P(Y = 1, hf(X) < 0) \\
&\qquad\qquad \text{(suppose } h \in \mathbb{R}^+ \text{ without loss of generality)} \\
&= P(X_1 < 0, f(X) \geq 0) + P(X_1 \geq 0, f(X) < 0) \\
&= P(X_1 < 0, X_1 + X_2 \geq 0) + P(X_1 \geq 0, X_1 + X_2 \leq 0) \\
&= \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.
\end{aligned}$$

Augmentation-Robust Loss

We define the Augmentation-Robust loss as

$$\mathcal{L}_{\mathrm{AR}}(f;\mathcal{D}) := \mathbb{E}_{X\in\mathcal{D}} \sup_{A_1,A_2} \|f(A_1(X)) - f(A_2(X))\|^2 \geq \mathcal{L}_{\mathrm{align}}(f;\mathcal{D}).$$

Theorem 2

For any A, let $h_A \in \arg\min_h \mathcal{R}(h\circ f, \mathcal{D}_A)$, we have

$$0 \leq \mathcal{R}(h_{A'}\circ f;\mathcal{D}_A) - \mathcal{R}(h_A\circ f;\mathcal{D}_A) \leq c\cdot(\|h_A\| + \|h_{A'}\|)\mathcal{L}_{AR}(f,\mathcal{D}).$$

Note that for any augmentation-invariant feature f, $\mathcal{R}(h_{A'}\circ f;\mathcal{D}_A) - \mathcal{R}(h_A\circ f;\mathcal{D}_A) = 0$.

The empirical version of AR loss is

$$\widehat{\mathcal{L}}_{\mathrm{AR}}(f) := \frac{1}{n}\sum_{k=1}^{n} \max_{A_i,A_j} \|f(A_i(X_k)) - f(A_j(X_k))\|^2.$$

Plug-And-Play ArCL

Algorithm 1: SimCLR + ArCL

input : Batch size N, temperature τ, augmentation π, number of views m, epoch T, encoder f, projector g.

for $t = 1, \ldots, T$ **do**
  sample minibatch $\{X_i\}_{i=1}^{N}$;
  for $i = 1, \ldots, N$ **do**
    draw m augmentations $\widehat{\mathcal{A}} = \{A_1, \ldots, A_m\} \sim \pi$;
    $z_{i,j} = g(f(A_j X_i))$ for $j \in [m]$;
    # select the worst positive samples;
    $s_i^+ = \min_{j,k \in [m]} \{z_{i,j}^\top z_{i,k} / (\|z_{i,j}\| \|z_{i,k}\|)\}$;
    # select the negative samples;
    for $j = 1, \ldots, N$ **do**
      $s_{i,j}^- = z_{i,1}^\top z_{j,1} / (\|z_{i,1}\| \|z_{j,1}\|)$;
      $s_{i,j+N}^- = z_{i,1}^\top z_{j,2} / (\|z_{i,1}\| \|z_{j,2}\|)$;
  compute $L = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s_i^+/\tau)}{\sum_{j=1, j \neq i}^{2N} \exp(s_{i,j}^-/\tau)}$;
  update f and g to minimize L;
return f

Performance on Augmented Datasets

	Method	Batch Size	Aug 1	Aug 2	Aug 3	Aug 4	Aug 5	Original
CIFAR10	SimCLR	256	86.36	83.21	86.93	86.42	86.13	86.76
	SimCLR + ArCL (views=4)	256	88.68	86.77	89.01	88.70	88.31	88.95
	SimCLR + ArCL (views=6)	256	**88.95**	**87.18**	**89.54**	**88.92**	**88.61**	**89.11**
	SimCLR	512	88.62	86.27	88.96	88.56	88.37	88.81
	SimCLR + ArCL (views=4)	512	89.97	88.06	90.48	89.91	89.59	90.20
	SimCLR + ArCL (views=6)	512	**90.24**	**89.54**	**90.69**	**90.43**	**90.07**	**90.69**
CIFAR100	SimCLR	256	51.65	47.55	53.17	52.05	51.36	52.75
	SimCLR+ArCL(views=4)	256	53.76	49.80	55.68	54.19	52.96	54.83
	SimCLR+ArCL(views=6)	256	**54.13**	**50.74**	**55.74**	**54.75**	**53.46**	**55.29**
	SimCLR	512	52.28	48.09	53.45	52.58	51.53	53.12
	SimCLR+ArCL(views=4)	512	53.40	50.16	54.92	53.77	52.61	54.20
	SimCLR+ArCL(views=6)	512	**54.00**	**50.57**	**56.24**	**55.04**	**53.77**	**55.60**

Aug 1: Grayscale; Aug 2: RandomCrop; Aug 3: HorizontalFlip; Aug 4: ColorJitter; Aug 5: Aug 1 + Aug 4.

Performance on OOD Datasets

	Epochs	Aircraft	Caltech101	Cars	CIFAR10	CIFAR100	DTD	Flowers	Food	Pets	**Avg**
Linear	MoCo	41.79	87.92	39.31	92.28	74.90	73.88	90.07	68.95	83.30	72.49
	MoCo + AAL (views=2)	40.53	87.80	38.64	92.23	75.14	**74.95**	88.64	69.24	83.17	72.26
	MoCo + ArCL (views=2)	**44.29**	**89.79**	**42.15**	**93.07**	**76.70**	74.20	**90.40**	**70.94**	**83.68**	**73.91**
	MoCo + AAL (views=3)	40.41	87.79	42.09	92.64	75.31	74.89	89.23	69.37	83.79	72.84
	MoCo + ArCL (views=3)	**44.57**	**89.48**	**42.11**	**93.29**	**77.33**	**74.63**	**91.13**	**71.16**	**84.23**	**74.21**
Finetune	MoCo	83.56	82.54	85.09	95.89	71.81	69.95	95.26	76.81	88.83	83.30
	MoCo + AAL (views=2)	83.87	82.76	85.90	**96.38**	71.43	**72.71**	95.50	76.95	**89.05**	83.84
	MoCo + ArCL (views=2)	**86.05**	**87.38**	**87.28**	96.33	**79.39**	72.18	**95.89**	**81.36**	89.03	**86.10**
	MoCo + AAL (views=3)	83.07	83.21	85.19	96.37	72.02	72.55	95.74	79.62	88.83	84.07
	MoCo + ArCL (views=3)	**84.03**	**87.64**	**86.34**	**96.88**	**80.98**	**72.87**	**96.14**	**81.90**	**89.20**	**86.22**

AAL: Average Alignment Loss.

ArCL: Augmentation-robust Contrastive Loss.

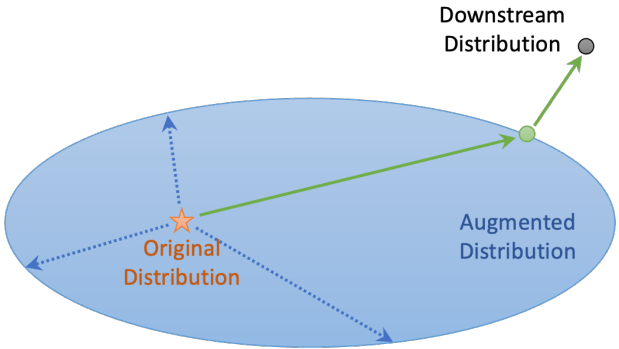
Short Summary

- We show that contrastive learning fails to learn augmentation-invariant features, which limits its transferability.
- We propose a theory-inspired loss ArCL which can be easily integrated with existing contrastive learning algorithms.
- We empirically verify that ArCL significantly improves the transferability of contrastive learning.

PS: In another ICLR'23 paper, we improve the transferability from the SNE perspective (see "Your Contrastive Learning Is Secretly Doing Stochastic Neighbor Embedding").

Conclusion

Training data distribution $\rightarrow$ Any downstream data distribution?



The Capability Boundary of Self-Supervised Contrastive Learning.

Thank you!

Interns and visitors are welcome!

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