Understanding Deep Contrastive Learning via Coordinate-wise Optimization

Yuandong Tian

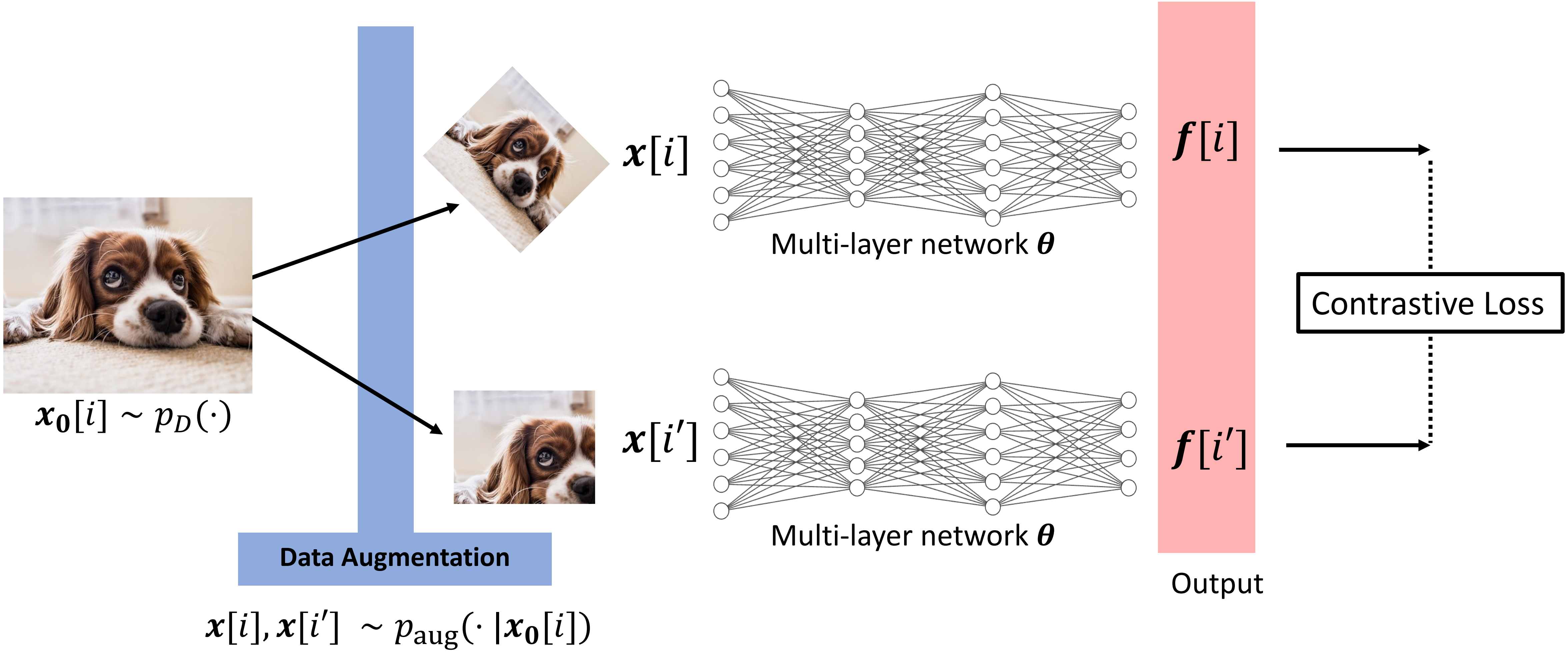
Research Scientist and Senior Manager

Meta AI (FAIR)

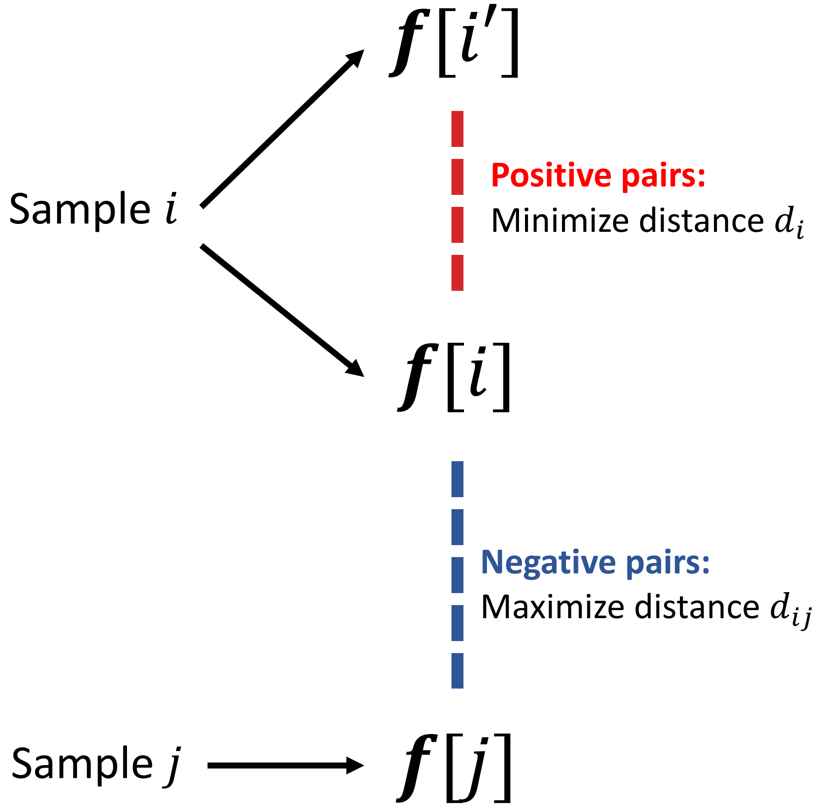
[Published in NeurIPS'22]

Great Empirical Success of Deep Models

Contrastive Learning (CL)



Formulation of Contrastive Learning



InfoNCE loss:

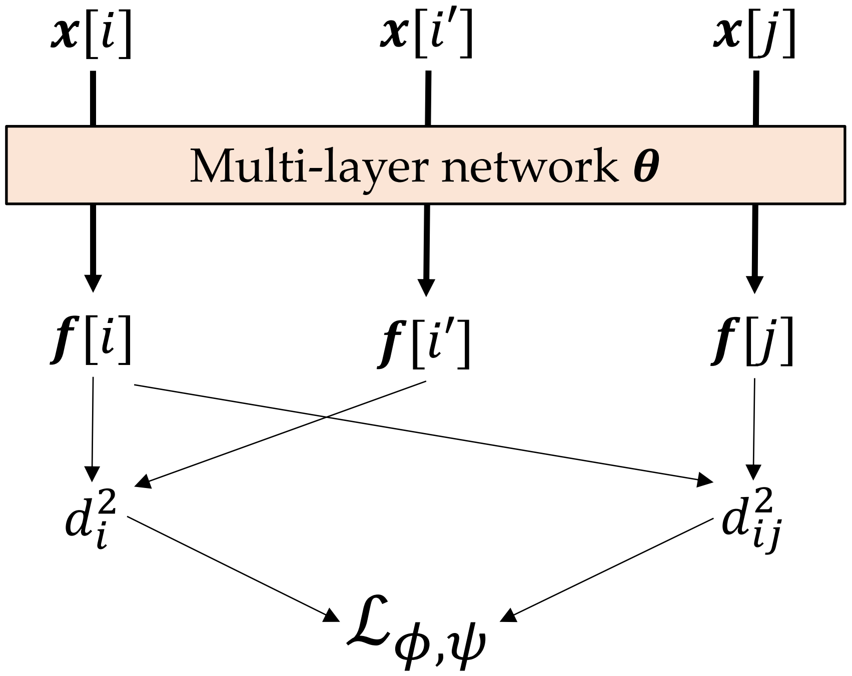
$$\mathcal{L}_{nce} := -\tau \sum_{i=1}^{N} \log \frac{\exp(-d_i^2/\tau)}{\epsilon \exp\left(-d_i^2/\tau\right) + \sum_{j\neq i} \exp(-d_{ij}^2/\tau)}$$

Intra-view distance $d_i^2 = \|\boldsymbol{f}[i] - \boldsymbol{f}[i']\|_2^2/2$

Inter-view distance $d_{ij}^2 = \|\boldsymbol{f}[i] - \boldsymbol{f}[j]\|_2^2/2$

A family of contrastive losses

General Loss function we consider (ϕ, ψ are monotonous increasing functions)



$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\phi,\psi}(\boldsymbol{\theta}) := \sum_{i=1}^{N} \phi\left(\sum_{j \neq i} \psi(d_i^2 - d_{ij}^2)\right)$$

Intra-view distance $d_i^2 = \|\boldsymbol{f}[i] - \boldsymbol{f}[i']\|_2^2/2$

Inter-view distance $d_{ij}^2 = \|\boldsymbol{f}[i] - \boldsymbol{f}[j]\|_2^2/2$

A general family

Contrastive Loss	$\phi(x)$	$\psi(x)$
InfoNCE (Oord et al., 2018)	$\tau \log(\epsilon + x)$	$e^{x/\tau}$
MINE (Belghazi et al., 2018)	$\log(x)$	e^x
Triplet (Schroff et al., 2015)	x	$[x + \epsilon]_+$
Soft Triplet (Tian et al., 2020c)	$\tau \log(1 + x)$	$e^{x/\tau+\epsilon}$
N+1 Tuplet (Sohn, 2016)	$\log(1 + x)$	e^x
Lifted Structured (Oh Song et al., 2016)	$[\log(x)]_+^2$	$e^{x+\epsilon}$
(Coria et al., 2020)	x	$\mathrm{sigmoid}(cx)$
(Ji et al., 2021)	linear	linear

Example: InfoNCE

$$\mathcal{L}_{nce} := -\tau \sum_{i=1}^{N} \log \frac{\exp(-d_i^2/\tau)}{\epsilon \exp\left(-d_i^2/\tau\right) + \sum_{j\neq i} \exp(-d_{ij}^2/\tau)}$$

$$= \tau \sum_{i=1}^{N} \log\left(\epsilon + \sum_{j\neq i} \exp\left(\frac{d_i^2 - d_{ij}^2}{\tau}\right)\right)$$

$$\phi(x) = \tau \log(\epsilon + x) \qquad \psi(x) = \exp(x/\tau)$$

Coordinate-wise Optimization

Claim: if $\psi(x) = e^{x/\tau}$, minimizing $\mathcal{L}_{\phi,\psi}$ ⇔ Coordinate-wise optimization:

$$\alpha_t := \arg\min_{\alpha \in \mathcal{A}} \mathcal{E}_\alpha(\boldsymbol{\theta}_t) - \mathcal{R}(\alpha)$$

$$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \eta \nabla_{\boldsymbol{\theta}} \mathcal{E}_{\alpha_t}(\boldsymbol{\theta}_t)$$

Max-player $\boldsymbol{\theta}$

Learns the representation to maximize contrastiveness.

Min-player α

Emphasize distinct sample pairs that share similar representation (**hard negative pairs**)

Different Losses, Same Energy Function

Contrastive Loss	$\phi(x)$	$\psi(x)$
InfoNCE (Oord et al., 2018)	$\tau \log(\epsilon + x)$	$e^{x/\tau}$
MINE (Belghazi et al., 2018)	$\log(x)$	e^x
Triplet (Schroff et al., 2015)	x	$[x + \epsilon]_+$
Soft Triplet (Tian et al., 2020c)	$\tau \log(1 + x)$	$e^{x/\tau + \epsilon}$
N+1 Tuplet (Sohn, 2016)	$\log(1 + x)$	e^x
Lifted Structured (Oh Song et al., 2016)	$[\log(x)]_+^2$	$e^{x+\epsilon}$
(Coria et al., 2020)	x	$\text{sigmoid}(cx)$
(Ji et al., 2021)	linear	linear

Different loss functions (ϕ, ψ) corresponds to the **same energy function $\mathcal{E}$**

How the min player α operates are different.

How min player $\boldsymbol{\alpha}$ is determined?

If $\psi(x) = e^{x/\tau}$, then we have $\alpha(\boldsymbol{\theta}) \coloneqq \arg\min_{\alpha \in \mathcal{A}} \mathcal{E}_\alpha(\boldsymbol{\theta}) - \mathcal{R}(\alpha)$

where the feasible set $\mathcal{A} \coloneqq \left\{\alpha\colon \ \forall i, \sum_{j \neq i} \alpha_{ij} = \tau^{-1}\xi_i \phi'(\xi_i), \alpha_{ij} \geq 0 \right\}$

and entropy regularization term $\mathcal{R}(\alpha) \coloneqq 2\tau \sum_{i=1}^{N} H(\alpha_{i\cdot})$

$$\xi_i \coloneqq \sum_{j \neq i} \psi(d_i^2 - d_{ij}^2)$$

For infoNCE with $\epsilon = 0$, solving the optimization problem yields:

$$\alpha_{ij}(\boldsymbol{\theta}) = \frac{\exp\left(-d_{ij}^2/\tau\right)}{\sum_{j \neq i} \exp\left(-d_{ij}^2/\tau\right)}$$

We put more weights on **small $\boldsymbol{d_{ij}}$**, i.e., distinct samples with similar representations

Coordinate-wise Optimization

Minimizing $\mathcal{L}_{\phi,\psi}$ ⇔ Coordinate-wise optimization:

$$\alpha_t \coloneqq \arg\min_{\alpha \in \mathcal{A}} \mathcal{E}_\alpha(\boldsymbol{\theta}_t) - \mathcal{R}(\alpha)$$

$$\boldsymbol{\theta}_{t+1} \coloneqq \boldsymbol{\theta}_t + \eta \nabla_{\boldsymbol{\theta}} \mathcal{E}_{\alpha_t}(\boldsymbol{\theta}_t)$$

Coordinate-wise Optimization

Minimizing $\mathcal{L}_{\phi,\psi}$ ⇔ Coordinate-wise optimization:

$$\alpha_t \coloneqq \arg\min_{\alpha \in \mathcal{A}} \mathcal{E}_\alpha(\boldsymbol{\theta}_t) - \mathcal{R}(\alpha)$$

$$\boldsymbol{\theta}_{t+1} \coloneqq \boldsymbol{\theta}_t + \eta \nabla_{\boldsymbol{\theta}} \mathcal{E}_{\alpha_t}(\boldsymbol{\theta}_t)$$

Proposed: Pair-weighed CL ($\boldsymbol{\alpha}$-CL)

The min player α can be optimized by a loss function, or ***directly*** specified:

Pairwise importance

$$\alpha_t = \text{sg}(\alpha(\boldsymbol{\theta}_t))$$

$$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \eta \nabla_{\boldsymbol{\theta}} \mathcal{E}_{\alpha_t}(\boldsymbol{\theta}_t)$$

Experimental Results

	CIFAR-10			*STL-10*		
	100 epochs	300 epochs	500 epochs	100 epochs	300 epochs	500 epochs
$\mathcal{L}_{quadratic}$	63.59 ± 2.53	73.02 ± 0.80	73.58 ± 0.82	55.59 ± 4.00	64.97 ± 1.45	67.28 ± 1.21
$\mathcal{L}_{nce}$	84.06 ± 0.30	87.63 ± 0.13	87.86 ± 0.12	78.46 ± 0.24	82.49 ± 0.26	83.70 ± 0.12
backprop $\alpha(\boldsymbol{\theta})$	83.42 ± 0.25	87.18 ± 0.19	87.48 ± 0.21	77.88 ± 0.17	81.86 ± 0.30	83.19 ± 0.16
α-CL-r_H	84.27 ± 0.24	87.75 ± 0.25	87.92 ± 0.24	78.53 ± 0.35	82.62 ± 0.15	83.74 ± 0.18
α-CL-r_γ	83.72 ± 0.19	87.51 ± 0.11	87.69 ± 0.09	78.22 ± 0.28	82.19 ± 0.52	83.47 ± 0.34
α-CL-r_s	84.72 ± 0.10	86.62 ± 0.17	86.74 ± 0.15	76.95 ± 1.06	80.64 ± 0.77	81.65 ± 0.59
α-CL-direct	$\mathbf{85.09 \pm 0.13}$	$\mathbf{88.00 \pm 0.12}$	$\mathbf{88.16 \pm 0.12}$	$\mathbf{79.38 \pm 0.16}$	$\mathbf{82.99 \pm 0.15}$	$\mathbf{84.06 \pm 0.24}$

- (α-CL-r_H) Entropy regularizer $r_H(\alpha_{ij}) = -2\tau\alpha_{ij}\log\alpha_{ij}$;
- (α-CL-r_γ) Inverse regularizers $r_\gamma(\alpha_{ij}) = \frac{2\tau}{1-\gamma}\alpha_{ij}^{1-\gamma}$ ($\gamma > 1$).
- (α-CL-r_s) Square regularizer $r_s(\alpha_{ij}) = -\frac{\tau}{2}\alpha_{ij}^2$.
- (α-CL-direct) Directly setting α: $\alpha_{ij} = \exp(-d_{ij}^p/\tau)$ ($p > 1$).

Experimental Results

More datasets

	CIFAR-100		
	100 epochs	300 epochs	500 epochs
$\mathcal{L}_{nce}$	55.696 ± 0.368	59.706 ± 0.360	59.892 ± 0.340
α-CL-direct	$\mathbf{57.144 \pm 0.150}$	$\mathbf{60.110 \pm 0.187}$	$\mathbf{60.330 \pm 0.194}$

Backbone = ResNet50

Dataset	Method	100 epochs	300 epochs	500 epochs
CIFAR-10	$\mathcal{L}_{nce}$	86.388 ± 0.157	89.974 ± 0.138	90.194 ± 0.232
	α-CL-direct	$\mathbf{87.406 \pm 0.227}$	$\mathbf{90.228 \pm 0.185}$	$\mathbf{90.366 \pm 0.209}$
CIFAR-100	$\mathcal{L}_{nce}$	60.162 ± 0.482	65.400 ± 0.310	65.532 ± 0.297
	α-CL-direct	$\mathbf{62.650 \pm 0.181}$	$\mathbf{65.630 \pm 0.263}$	$\mathbf{65.636 \pm 0.269}$
STL-10	$\mathcal{L}_{nce}$	81.635 ± 0.244	86.570 ± 0.174	$\mathbf{87.900 \pm 0.222}$
	α-CL-direct	$\mathbf{82.850 \pm 0.171}$	$\mathbf{86.870 \pm 0.178}$	87.653 ± 0.175

Roadmap of α-CL

$$\mathcal{E}_\alpha(\boldsymbol{\theta}) := \operatorname{tr} \mathbb{C}_\alpha[\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})]$$

α-CL

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\phi,\psi}(\boldsymbol{\theta})$$

Minimization of various CL losses

Applications →

- Finding the best $\alpha = \alpha(\boldsymbol{\theta})$ for performance gain
- Receptive-field specific α
- More applications (e.g., CL in GNN)

Understanding →

- Dynamics of $\boldsymbol{\theta}$ with fixed α in the linear setting
- Dynamics of $\boldsymbol{\theta}$ in the nonlinear setting
- Hierarchical representation learning

Roadmap of $\boldsymbol{\alpha}$-CL

$$\mathcal{E}_\alpha(\boldsymbol{\theta}) \coloneqq \mathrm{tr}\, \mathbb{C}_\alpha[\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})]$$

α-CL

Applications →

- Finding the best $\alpha = \alpha(\boldsymbol{\theta})$ for performance gain
- Receptive-field specific α
- More applications (e.g., CL in GNN)

Understanding →

- Dynamics of $\boldsymbol{\theta}$ with fixed α in the linear setting
- Dynamics of $\boldsymbol{\theta}$ in the nonlinear setting
- Hierarchical representation learning

↑

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\phi,\psi}(\boldsymbol{\theta})$$

Minimization of various CL losses

Deep linear case with fixed $\boldsymbol{\alpha}$

If $\boldsymbol{f_\theta}(\boldsymbol{x}) = W(\boldsymbol{\theta})\boldsymbol{x}$, then Contrastive Learning reduces to PCA objective

Corollary 2 (Representation learning in Deep Linear CL reparameterizes Principal Component Analysis (PCA))**.** *When $\boldsymbol{z} = W(\boldsymbol{\theta})\boldsymbol{x}$ with a constraint $WW^\top = I$, $\mathcal{E}_\alpha$ is the objective of Principal Component Analysis (PCA) with reparameterization $W = W(\boldsymbol{\theta})$:*

$$\max_{\boldsymbol{\theta}} \mathcal{E}_\alpha(\boldsymbol{\theta}) = \mathrm{tr}(W(\boldsymbol{\theta})X_\alpha W^\top(\boldsymbol{\theta})) \quad \text{s.t. } WW^\top = I \tag{9}$$

here $X_\alpha := \mathbb{C}_\alpha[\boldsymbol{x}]$ is the contrastive covariance of input $\boldsymbol{x}$.

Deep linear case with fixed $\boldsymbol{\alpha}$

If $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = W_L W_{L-1} \ldots W_1 x$, then almost all local optima are global and it is PCA

Theorem 3 (Representation Learning with `DeepLin` is PCA). *If $\lambda_{\max}(X_\alpha) > 0$, then for any local maximum $\boldsymbol{\theta} \in \Theta$ of Eqn. 11 whose $W_{>1}^\top W_{>1}$ has distinct maximal eigenvalue:*

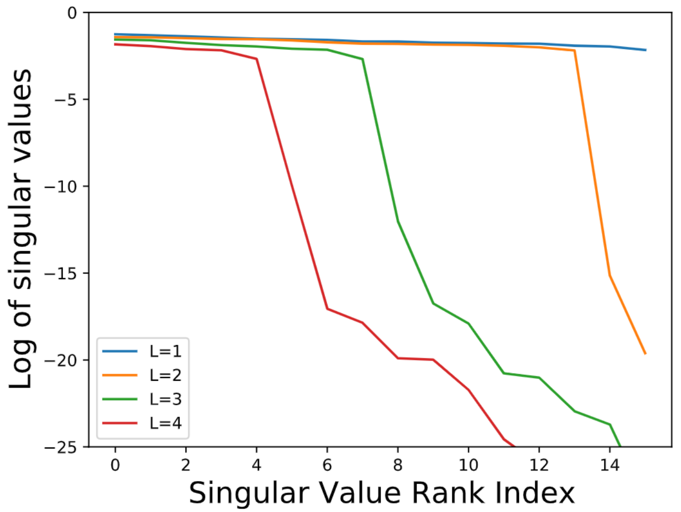
- *there exists a set of unit vectors $\{\boldsymbol{v}_l\}_{l=0}^{L}$ so that $W_l = \boldsymbol{v}_l \boldsymbol{v}_{l-1}^\top$ for $1 \le l \le L$, in particular, $\boldsymbol{v}_0$ is the unit eigenvector corresponding to $\lambda_{\max}(X_\alpha)$,*
- *$\boldsymbol{\theta}$ is global optimal with objective $\mathcal{E}^* = \lambda_{\max}(X_\alpha)$.*

1. Nearby weights align
2. All W_l has rank-1 structure

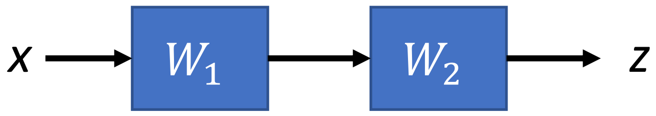
Corollary 3. *If we additionally use per-filter normalization (i.e., $\|\boldsymbol{w}_{lk}\|_2 = 1/\sqrt{n_l}$), then Thm. 3 holds and $\boldsymbol{v}_l$ is more constrained: $[\boldsymbol{v}_l]_k = \pm 1/\sqrt{n_l}$ for $1 \le l \le L-1$.*

Dimensional Collapsing in CL

Shouldn't contrastive SSL make full use of all dimensions? The answer is **No…**



(a) multiple layers



W_1 and W_2 will align with each other

If things are aligned, why not let them align directly?

Loss function	Projector	Top-1 Accuracy
SimCLR	2-layer nonlinear projector	66.5
SimCLR	1-layer linear projector	61.1
SimCLR	no projector	51.5
DirectCLR	no projector	62.7

DirectCLR *[L. Jing, P. Vincent, Y. LeCun, **Y. Tian**, Understanding Dimensional Collapse in Contrastive Self-supervised Learning, ICLR'22]*

Roadmap of $\boldsymbol{\alpha}$-CL

$$\mathcal{E}_\alpha(\boldsymbol{\theta}) \coloneqq \mathrm{tr}\, \mathbb{C}_\alpha[\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})]$$

α-CL

Applications

- Finding the best $\alpha = \alpha(\boldsymbol{\theta})$ for performance gain
- Receptive-field specific α
- More applications (e.g., CL in GNN)

Understanding

- Dynamics of $\boldsymbol{\theta}$ with fixed α in the linear setting
- Dynamics of $\boldsymbol{\theta}$ in the nonlinear setting

  NeurIPS 2022 Workshop:
  Self-Supervised Learning - Theory and Practice
- Hierarchical representation learning

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\phi,\psi}(\boldsymbol{\theta})$$

Minimization of various CL losses

Thanks!