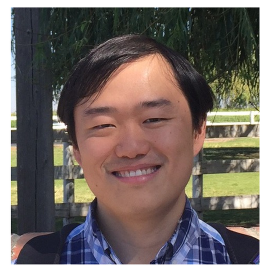
Understanding Contrastive Learning via Coordinate-wise Optimization

Yuandong Tian

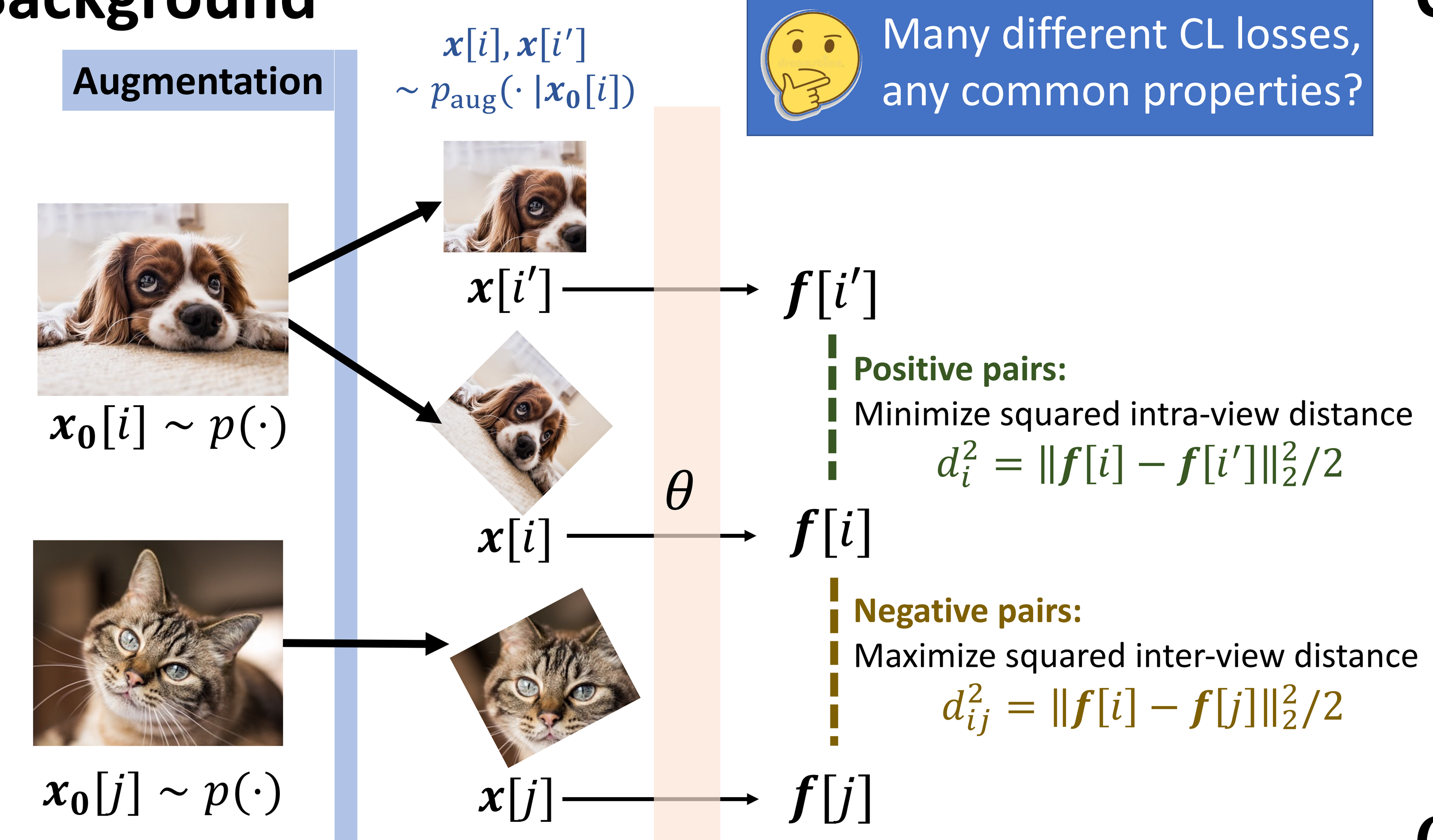
yuandong@meta.com



Meta AI

Selected as Oral

Background



Many different CL losses, any common properties?

Proposed Unified Framework

General CL loss (ϕ, ψ are monotonous increasing functions)

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

For infoNCE:

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

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

Common piece of various CL loss functions

First we can prove $\dfrac{\partial \mathcal{L}_{\phi,\psi}}{\partial \boldsymbol{\theta}} = -\dfrac{\partial \mathcal{E}_\alpha}{\partial \boldsymbol{\theta}}\Big|_{\alpha=\alpha(\boldsymbol{\theta})}$

for the energy $\mathcal{E}_\alpha$ defined as the *trace* of ***contrastive covariance*** $\mathbb{C}_\alpha$:

$$\mathcal{E}_\alpha(\boldsymbol{\theta}) := \frac{1}{2} \operatorname{tr} \mathbb{C}_\alpha[\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})]$$

where the **contrastive covariance** is defined as

$$\mathbb{C}_\alpha[\boldsymbol{f}] := \sum_{i,j} \alpha_{ij}[(\boldsymbol{f}[i] - \boldsymbol{f}[j])(\boldsymbol{f}[i] - \boldsymbol{f}[j])^T - (\boldsymbol{f}[i] - \boldsymbol{f}[i'])(\boldsymbol{f}[i] - \boldsymbol{f}[i'])^T]$$

Here the **pairwise importance** $\alpha_{ij} := \phi'(\xi_i)\psi'(d_i^2 - d_{ij}^2) \geq 0$,
where $\xi_i := \sum_{j\neq i} \psi(d_i^2 - d_{ij}^2)$

α as an adversarial player

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

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

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

Example For infoNCE:

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

Larger α_{ij} on **small d_{ij}** → distinct samples with similar representations

[Theorem] 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 constrativeness.

Min-player $\boldsymbol{\alpha}$

Find distinct sample pairs that share similar representation (i.e., **hard negative pairs**)

The pairwise importance α incorporates the effects of ϕ and ψ.

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)	$[\log(x)]_+^2$	$e^{x+\epsilon}$
Modified Triplet (Eqn. 10 in Coria et al., 2020))	x	$\text{sigmoid}(cx)$
Triplet Contrastive (Eqn. 2 in Ji et al., 2021)	Linear	Linear

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

How the min player $\alpha = \alpha(\boldsymbol{\theta})$ operates is different.

Proposed: α-CL

Why we are stuck with coordinate-wise optimization?

Optimize network parameter $\boldsymbol{\theta}$ using gradient ascent of the energy function $\mathcal{E}$:

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

Pairwise importance $\alpha_t = \alpha(\boldsymbol{\theta}_t)$

The pairwise importance α can be

1. optimized by a separate loss function, or
2. ***directly*** specified (**α-CL-direct**)

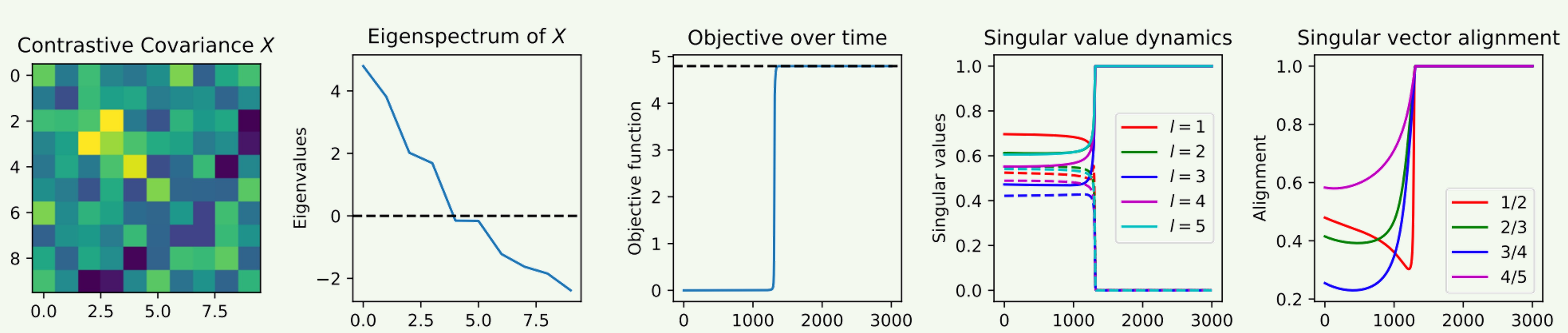
Theoretical Properties when α is fixed

Deep linear network

If $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = W_L W_{L-1} \dots W_1 \boldsymbol{x}$, then almost all local optima are global, and CL becomes Principal Component Analysis (PCA).

[Theorem] Let $X_\alpha := \mathbb{C}_\alpha[\boldsymbol{x}]$. If $\lambda_{\max}(X_\alpha) > 0$, then for any local maximum $\boldsymbol{\theta} = \{W_L, W_{L-1}, \dots, W_1\}$ whose $W_{>1}^T W_{>1}$ has distinct maximal eigenvalue, then

- $\boldsymbol{\theta}$ is aligned rank-1 (i.e., $W_l = \boldsymbol{v}_l \boldsymbol{v}_{l-1}^T$), $\boldsymbol{v}_0$ is the unit eigenvector for $\lambda_{\max}(X_\alpha)$.
- $\boldsymbol{\theta}$ is globally optimal with objective $2\mathcal{E}^* = \lambda_{\max}(X_\alpha)$.



Nonlinear network

Many interesting properties. Detailed in the paper and follow-up works (Please check Workshop on SSL, Theory and Practice on Dec. 3)

Experimental Results: α-CL

Use ResNet18 backbone, and set different α

- α-CL-r_H: Entropy regularizer
- α-CL-r_γ: Inverse regularizer
- α-CL-r_s: Square regularizer
- α-CL-direct: Directly setting α.

	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	$\mathbf{85.09 \pm 0.13}$	$\mathbf{88.00 \pm 0.12}$	$\mathbf{88.16 \pm 0.12}$	$\mathbf{79.38 \pm 0.16}$	$\mathbf{82.99 \pm 0.15}$	$\mathbf{84.06 \pm 0.24}$

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	$\mathbf{57.144 \pm 0.150}$	$\mathbf{60.110 \pm 0.187}$	$\mathbf{60.330 \pm 0.194}$

α-CL-direct:

$$\alpha_{ij} := \frac{\exp\left(-\frac{d_{ij}^p}{\tau}\right)}{\sum_{i\neq j} \exp\left(-\frac{d_{ij}^p}{\tau}\right)}$$

$(p = 4)$

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