Understanding Contrastive Learning via Coordinate-wise Optimization

Yudong Tian
yuandong@meta.com

Meta AI

Selected as Oral

Background

Many different CL losses, any common properties?

Augmentation

Proposed Unified Framework

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

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

For InfoNCE:

$$\mathcal{L}_{{\text{InfoNCE}}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{-\alpha d_{ii}^2 / \tau}}{e^{-\alpha d_{ii}^2 / \tau} + \sum_{j \neq i} e^{-\alpha d_{ij}^2 / \tau}} = \frac{1}{N} \sum_{i=1}^{N} \log \left( 1 + \frac{e^{\alpha d_{ii}^2 / \tau}}{\sum_{j \neq i} e^{\alpha d_{ij}^2 / \tau}} \right)$$

Here φ(x) = x log(1 + x) and ψ(y) = exp(x/y)

Common piece of various CL loss functions

First we can prove

$$\frac{\partial \mathcal{L}_{\phi, \psi}}{\partial \theta} = -\frac{\partial \mathcal{E}_\alpha}{\partial \theta} \bigg| \alpha = \alpha(\theta)$$

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

$$\mathcal{E}_\alpha(\theta) = \frac{1}{2} \text{tr} \mathbb{C}_\alpha (\mathbf{f}(x))$$

where the contrastive covariance is defined as

$$\mathbb{C}_\alpha [\mathbf{f}](\theta) = \frac{1}{T} \sum_{t \in \mathcal{T}} a_{ij}(t)(\mathbf{f}^t(\theta) - \mathbf{f}(\theta))^2 - (\mathbf{f}^t(\theta) - \mathbf{f}(\theta))^2$$

Here the pairwise importance $a_{ij}(t) = \phi'(d_{ij})(d_{ij}^2 - d_{ji}^2) \geq 0$

where $\xi_i = \sum_{ij} (d_{ij}^2 - d_{ji}^2)$

$\alpha$ as an adversarial player

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

where $\mathfrak{A} = \{ \alpha : \forall i, \xi_i \alpha(i) \geq 0 \}$

and entropy regularization term $\mathcal{R}(\alpha) = \tau \sum_i H(a_i)$

Example For infoNCE:

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

Larger $a_{ij}$ on small $d_{ij}$ \rightarrow distinct samples with similar representations

Proposed: $\alpha$-CL

Why are we stuck with coordinate-wise optimization?

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

$$\theta_{t+1} = \theta_t + \eta \nabla_{\theta} \mathcal{E}(\theta_t)$$

Pairwise importance $a_{ij} = \alpha(\theta_t)$

The pairwise importance $\alpha_e$ can be

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

Theoretical Properties when $\alpha$ is fixed

Deep linear network

If $f_i(x) = W_i W_{i-1} \ldots W_1 x$, then almost all local optima are global, and CL becomes Principal Component Analysis (PCA).

[Theorem] Let $X_\alpha := C_{\alpha}[x]$, if $\lambda_{\max}(X_\alpha) > 0$, then for any local maximum $\theta = (W_1 W_2 \ldots W_{N-1})$, whose $W_{N-1}$ $W_1$ has distinct maximal eigenvalue, then

* The $\alpha$ is aligned rank-1 (i.e., $W_i = \phi(x_{ij}) \nu_i$, $\nu_i$ is the unit eigenvector for $\lambda_{\max}(X_\alpha)$.
* $\alpha$ is globally optimal with objective $2 \mathcal{E}^* = \lambda_{\max}(X_\alpha)$.

Contrastive Loss

InfoNCE (Oord et al, 2018)

$$\phi(x) = \tau \log(e^x + 1)$$

MINE (Belghazi et al, 2018)

$$\phi(x) = \log(x)$$

Triplet (Schroff et al, 2015)

$$\phi(x) = |x + e|$$

Soft Triplet (Tian et al., 2020)

$$\phi(x) = \tau \log(1 + x)$$

$N+1$ Triplet (Soehn, 2016)

$$\phi(x) = \log(1 + x)$$

Lifted Structured (Gh Song et al., 2016)

$$\phi(x) = \log(x)^2$$

Modified Triplet (Eqn. 10 in Cora et al., 2020)

$$\phi(x) = \text{sigmoid}(x)$$

Triplet Contrastive (Eqn. 2 in Ji et al., 2021)

Linear

Experimental Results: $\alpha$-CL

Use ResNet18 backbone, and set different $\alpha$

$$\alpha_{\text{CL-direct}}$$: Entropy regularizer
$$\alpha_{\text{CL-sigmoid}}$$: Inverse regularizer
$$\alpha_{\text{CL-square}}$$: Square regularizer

Deep linear network

$\mathcal{L}_\alpha(\theta)$ = $\frac{1}{N} \sum_{i=1}^{N} \phi(a_{ij}(\theta))$

More datasets

CIFAR-10

<table>
<thead>
<tr>
<th>Method</th>
<th>$\mathcal{L}_\alpha$</th>
<th>CIFAR-10</th>
<th>STL-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{\text{CL-direct}}$</td>
<td>57.144 ± 0.150</td>
<td>60.110 ± 0.187</td>
<td>60.530 ± 0.194</td>
</tr>
<tr>
<td>$\alpha_{\text{CL-square}}$</td>
<td>60.996 ± 0.160</td>
<td>65.400 ± 0.310</td>
<td>65.330 ± 0.340</td>
</tr>
<tr>
<td>$\alpha_{\text{CL-sigmoid}}$</td>
<td>62.650 ± 0.181</td>
<td>65.630 ± 0.263</td>
<td>65.630 ± 0.269</td>
</tr>
</tbody>
</table>

Backbone = ResNet50

<table>
<thead>
<tr>
<th>$\mathcal{L}_\alpha$</th>
<th>CIFAR-10</th>
<th>STL-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>87.426 ± 0.227</td>
<td>90.234 ± 0.185</td>
<td>87.850 ± 0.175</td>
</tr>
<tr>
<td>88.040 ± 0.209</td>
<td>90.569 ± 0.194</td>
<td>87.850 ± 0.222</td>
</tr>
</tbody>
</table>

Many interesting properties. Detailed in the paper and follow-up works.

(Please check Workshop on SSL, Theory and Practice on Dec. 3)