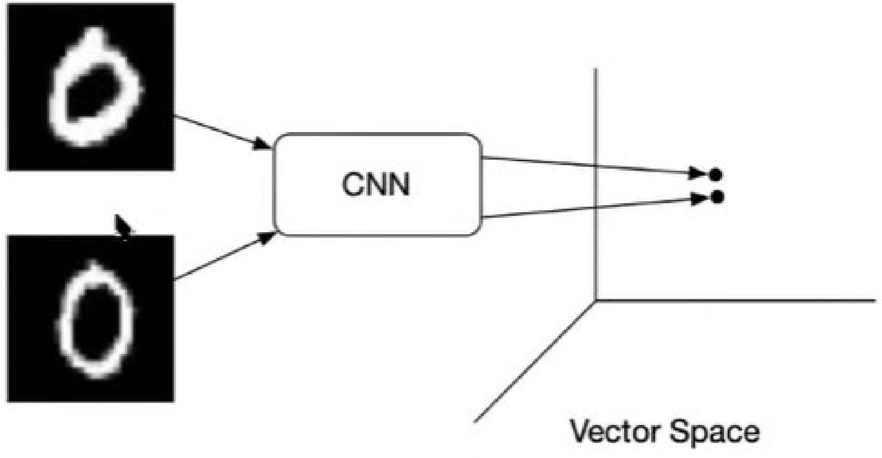
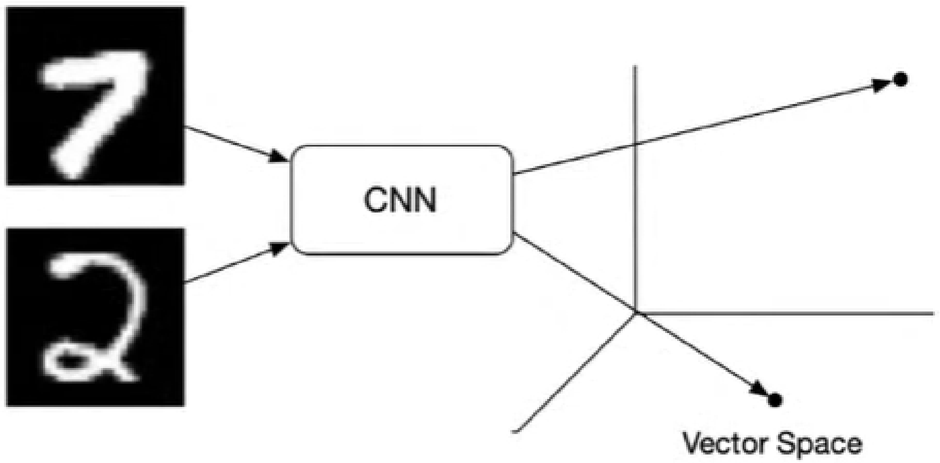
Large-Margin Contrastive Learning with Distance Polarization Regularizer

（ICML 2021）

Presenter: Minjie Cheng

Motivation





$$\mathcal{L}_{\mathrm{NCE}}(\varphi)$$

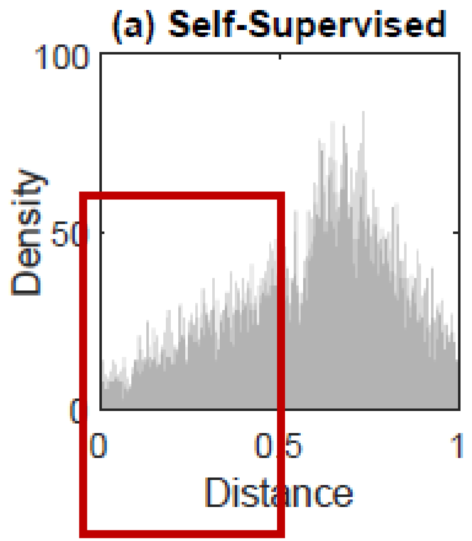
$$= \mathbb{E}_{\boldsymbol{x},\, \boldsymbol{x}_j^- \in \mathcal{X}} \left[-\log \frac{e^{\varphi(\boldsymbol{x})^\top \varphi(\boldsymbol{x}^+)}}{e^{\varphi(\boldsymbol{x})^\top \varphi(\boldsymbol{x}^+)} + \sum_{j=1}^{n} e^{\varphi(\boldsymbol{x})^\top \varphi(\boldsymbol{x}_j^-)}} \right]$$

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^m \mid i = 1, 2, \ldots, N\}$$

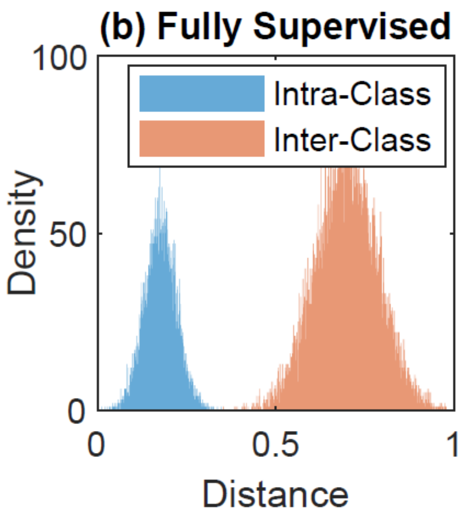
Motivation

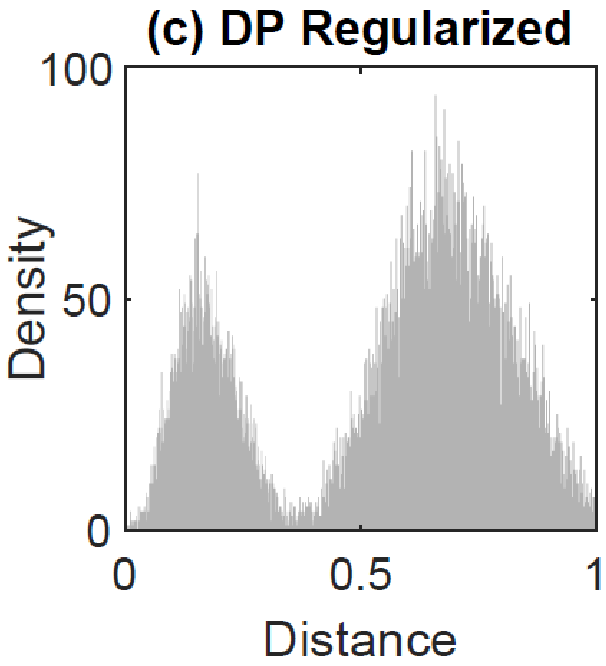
$$\mathcal{D}_{ij}^{\varphi} = (1 - \varphi(x_i)^{\top}\varphi(x_j))/2,$$

$$\mathcal{L}_{\mathrm{TUP}}(\varphi) = \mathbb{E}_{y_i = y_k \neq y_{b_j}}\left[-\log \frac{e^{\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\top}\boldsymbol{\varphi}(\boldsymbol{x}_k)}}{e^{\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\top}\boldsymbol{\varphi}(\boldsymbol{x}_k)} + \sum_{j=1}^{n} e^{\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\top}\boldsymbol{\varphi}(\boldsymbol{x}_{b_j})}}\right]$$

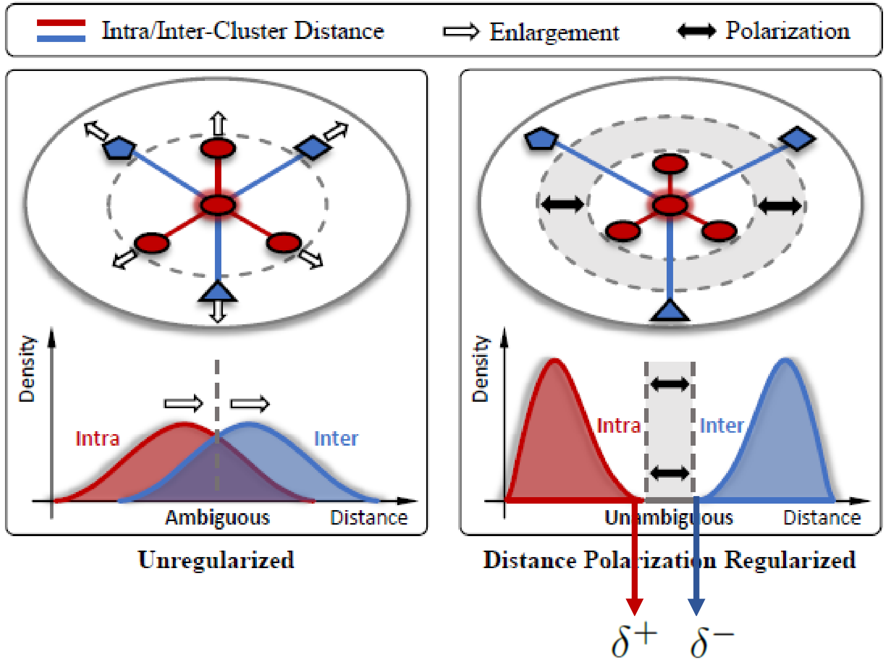


SimCLR-CIFAR-10





Distance Polarization Regularizer



$$\mathcal{D}^{\varphi} = [\mathcal{D}^{\varphi}_{ij}] \in \mathbb{R}^{N\times N}$$

$$\mathcal{R}_0(\varphi) = \|\min((\mathcal{D}^{\varphi} - \boldsymbol{\Delta}^+)\odot(\mathcal{D}^{\varphi} - \boldsymbol{\Delta}^-), 0)\|_0$$

$$\boldsymbol{\Delta}^+ = \delta^+ \cdot \mathbf{1}_{N\times N} \qquad \boldsymbol{\Delta}^- = \delta^- \cdot \mathbf{1}_{N\times N}$$

$$\boxed{\min_{\varphi\in\mathcal{H}} \mathcal{L}_{\mathrm{NCE}}(\varphi)} + \boxed{\lambda\mathcal{R}_0(\varphi)}$$

$$\mathcal{L}_{\mathrm{NCE}}(\varphi) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}_j^- \in \mathcal{X}}\left[-\log \frac{e^{\varphi(\boldsymbol{x})^\top \varphi(\boldsymbol{x}^+)}}{e^{\varphi(\boldsymbol{x})^\top \varphi(\boldsymbol{x}^+)} + \sum_{j=1}^{n} e^{\varphi(\boldsymbol{x})^\top \varphi(\boldsymbol{x}_j^-)}}\right]$$

Distance Polarization Regularizer： Optimization

$$\mathcal{R}_0(\boldsymbol{\varphi}) = \|\min((\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^{+}) \odot (\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^{-}), 0)\|_0$$

non-continuous non-convex

$$\downarrow$$

$$\mathcal{R}_1(\boldsymbol{\varphi}) = \|\min((\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^{+}) \odot (\boldsymbol{\mathcal{D}}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^{-}), 0)\|_1$$

$$\downarrow$$

$$\begin{aligned}
&\mathcal{R}_1(\boldsymbol{\varphi}) \\
&= \frac{2}{\binom{N}{n}} \sum\nolimits_{\boldsymbol{b} \in B} \sum\nolimits_{j=1}^{n+1} |\min((\mathcal{D}^{\boldsymbol{\varphi}}_{b_i b_j} - \delta^{+}) \odot (\mathcal{D}^{\boldsymbol{\varphi}}_{b_i b_j} - \delta^{-}), 0)| \\
&= \frac{1}{\binom{N}{n+1}} \sum\nolimits_{\boldsymbol{b} \in B} r(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1}),
\end{aligned}$$

$$\{\boldsymbol{x}_{b_j} | \boldsymbol{x}_{b_j} \in \mathcal{X}, b_j \in B\}_{j=1}^{n+1}$$

$$\ell(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1}) = -\log(\exp(\boldsymbol{\varphi}(\boldsymbol{x}_{b_{n+1}})^{\top} \boldsymbol{\varphi}(\boldsymbol{x}^{+}_{b_{n+1}})) / (\exp(\boldsymbol{\varphi}(\boldsymbol{x}_{b_{n+1}}))^{\top} \boldsymbol{\varphi}(\boldsymbol{x}^{+}_{b_{n+1}})) + \sum\nolimits_{j=1}^{n} \exp(\boldsymbol{\varphi}(\boldsymbol{x}_{b_j}))^{\top} \boldsymbol{\varphi}(\boldsymbol{x}^{-}_{b_j}))))$$

$$+$$

$$f(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1}) = \ell(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1}) + \lambda r(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1})$$

Distance Polarization Regularizer： Optimization

$$f(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1}) = \ell(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1}) + \lambda r(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1})$$

$$\downarrow$$

Algorithm 1 Solving Eq. (9) via Adam.

Input: Training Data $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$; Step Size $\eta > 0$; Regularization Parameter $\lambda > 0$; Batch Size $n \in \mathbb{N}_+$.

Initialize: Momentum Vectors $\boldsymbol{m}_{(0)} = \boldsymbol{v}_{(0)} = 0$; Decay Rates $\alpha_1, \alpha_2 \in (0, 1)$; Iteration Number $t = 0$.

For t **from** 1 **to** T**:**

1). Uniformly pick $(n+1)$ data points $\{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1}$ from $\mathcal{X}$;

2). Compute the stochastic gradient via Eq. (10):

$$\boldsymbol{g}_{(t)} \leftarrow \nabla_{\boldsymbol{\varphi}}(\ell(\boldsymbol{\varphi};\{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1}) + \lambda r(\boldsymbol{\varphi};\{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1})); \quad (11)$$

3). Compute moment vectors: $\boldsymbol{m}_{(t+1)} \leftarrow \alpha_1 \boldsymbol{m}_t + (1 - \alpha_1)\boldsymbol{g}_{(t)}$, and $\boldsymbol{v}_{(t+1)} \leftarrow \alpha_2 \boldsymbol{v}_t + (1 - \alpha_2)\boldsymbol{g}_{(t)} \odot \boldsymbol{g}_{(t)}$;

4). Update the learning parameter:

$$\boldsymbol{\varphi}_{(t+1)} \leftarrow \boldsymbol{\varphi}_{(t)} - \eta \frac{\boldsymbol{m}_{(t+1)}/(1 - \alpha_1^{t+1})}{\sqrt{\boldsymbol{v}_{(t+1)}/(1 - \alpha_2^{t+1})} + \epsilon}; \quad (12)$$

End.

Output: The converged $\tilde{\boldsymbol{\varphi}}$.

$$\begin{aligned}&\mathcal{R}_1(\boldsymbol{\varphi})\\ &= \frac{2}{\binom{N}{n}} \sum_{\boldsymbol{b} \in B} \sum_{j=1}^{n+1} |\min((\mathcal{D}^{\boldsymbol{\varphi}}_{b_i b_j} - \delta^+) \odot (\mathcal{D}^{\boldsymbol{\varphi}}_{b_i b_j} - \delta^-), 0)|\\ &= \frac{1}{\binom{N}{n+1}} \sum_{\boldsymbol{b} \in B} r(\boldsymbol{\varphi}; \{\boldsymbol{x}_{b_j}\}_{j=1}^{n+1}),\end{aligned} \quad (10)$$

Distance Polarization Regularizer： Error Bound for Downstream Classification

Theorem 4. *Let* $\varphi^* \in \arg\min_{\varphi \in \mathcal{H}} \mathcal{L}_{\text{NCE}}(\varphi) + \lambda \mathcal{R}_1(\varphi)$. *Then with probability at least* $1-\delta$*, we have that*

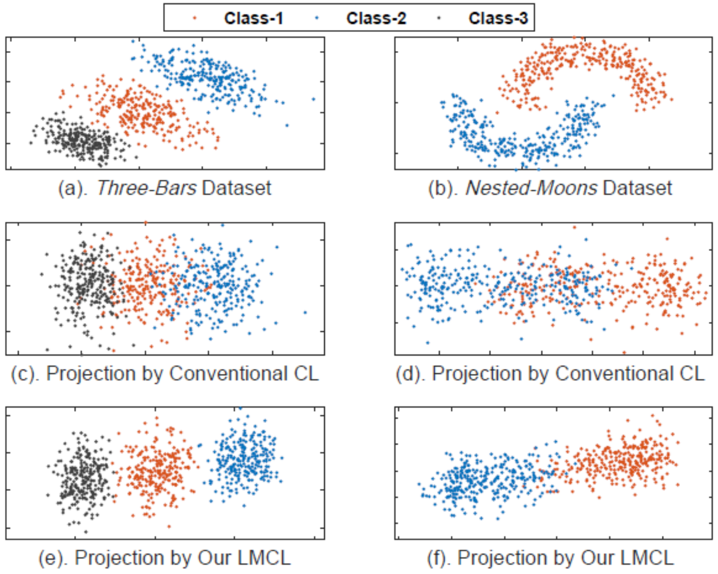
$$\left|\mathcal{L}_{\text{SM}}^{\mathcal{T}}(\varphi^*) - \mathcal{L}_{\text{NCE}}(\varphi^*)\right| \leq \mathcal{O}\left(\frac{Q_1 \mathfrak{R}_{\mathcal{H}}(\lambda)}{N} + \sqrt{\frac{Q_2}{N}}\right), \quad (14)$$

where $Q_1 = \sqrt{1+1/n}$, $Q_2 = \log(1/\delta) \cdot \log^2(n)$, *and*[5] $\mathfrak{R}_{\mathcal{H}}(\lambda)$ *is monotonically decreasing w.r.t.* λ.

Lemma 3. (Saunshi et al., 2019) *Assume that* $\varphi^* \in \arg\min_{\varphi \in \mathcal{H}} \mathcal{L}_{\text{NCE}}(\varphi) + \lambda \mathcal{R}_1(\varphi)$. *Then with probability at least* $1-\delta$ *over the training data* $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, *for any* $\varphi \in \mathcal{H}$

$$\mathcal{L}_{\text{NCE}}(\varphi^*) \leq \mathcal{L}_{\text{NCE}}(\varphi) + \mathcal{O}\left(\frac{Q_1 \mathfrak{R}_{\mathcal{H}}(\lambda)}{N} + \sqrt{\frac{Q_2}{N}}\right), \quad (19)$$

Experiments



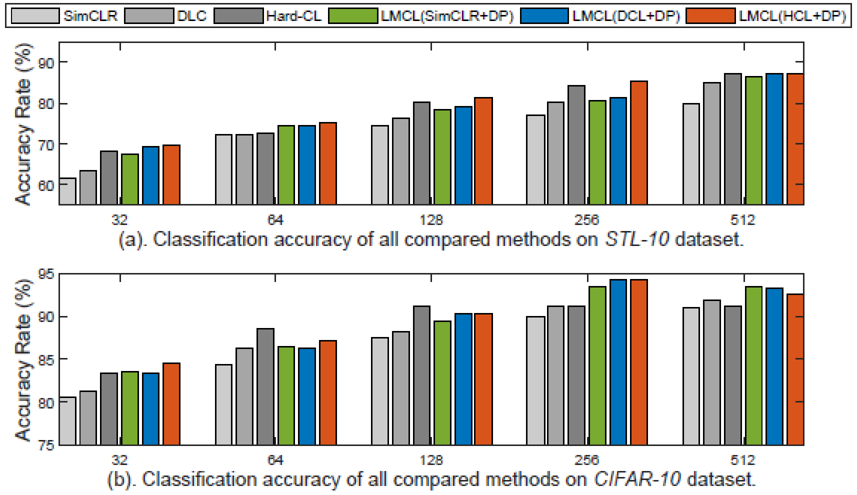


Table 4. Classification accuracy (%) of all methods on *BookCorpus* dataset including six text classification tasks.

METHOD	*MR*	*CR*	*SUBJ*	*MPQA*	*TREC*	*MSRP*
QT	76.8	81.3	86.6	93.4	89.8	73.6
DCL	76.2	82.9	86.9	93.7	89.1	74.7
HCL	77.4	83.6	86.8	93.4	88.7	73.5
LMCL(QT+DP)	77.3	82.3	86.9	93.7	**90.2**	74.1
LMCL(DCL+DP)	77.2	**83.7**	**87.2**	93.8	90.1	**75.1**
LMCL(HCL+DP)	**78.1**	83.5	**87.2**	**94.0**	89.1	74.2

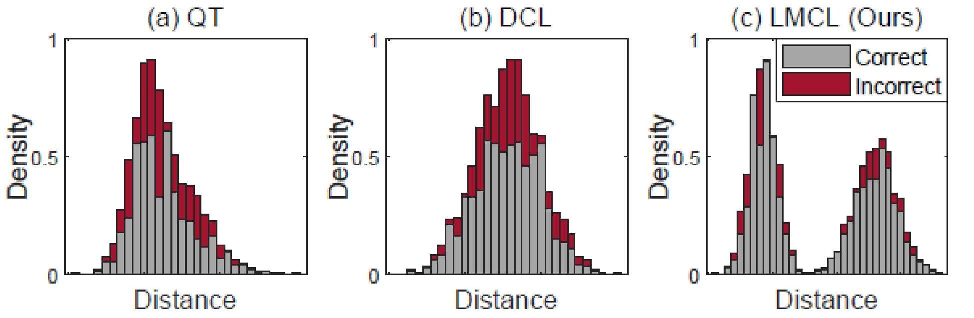


Figure 5. Distance histograms obtained by different methods (QT, DCL, and our proposed LMCL) on *BookCorpus* dataset.

Summary

1. A regularizer for contrastive learning
2. Theoretical analyses and enough experiments