Robust Similarity Learning with Difference Alignment Regularization

Shuo Chen[1], Gang Niu[1], Chen Gong[2], Okan Koc[1], Jian Yang[2], Masashi Sugiyama[1,3]

1. RIKEN Center for Advanced Intelligence Project 2. Nanjing University of Science and Technology

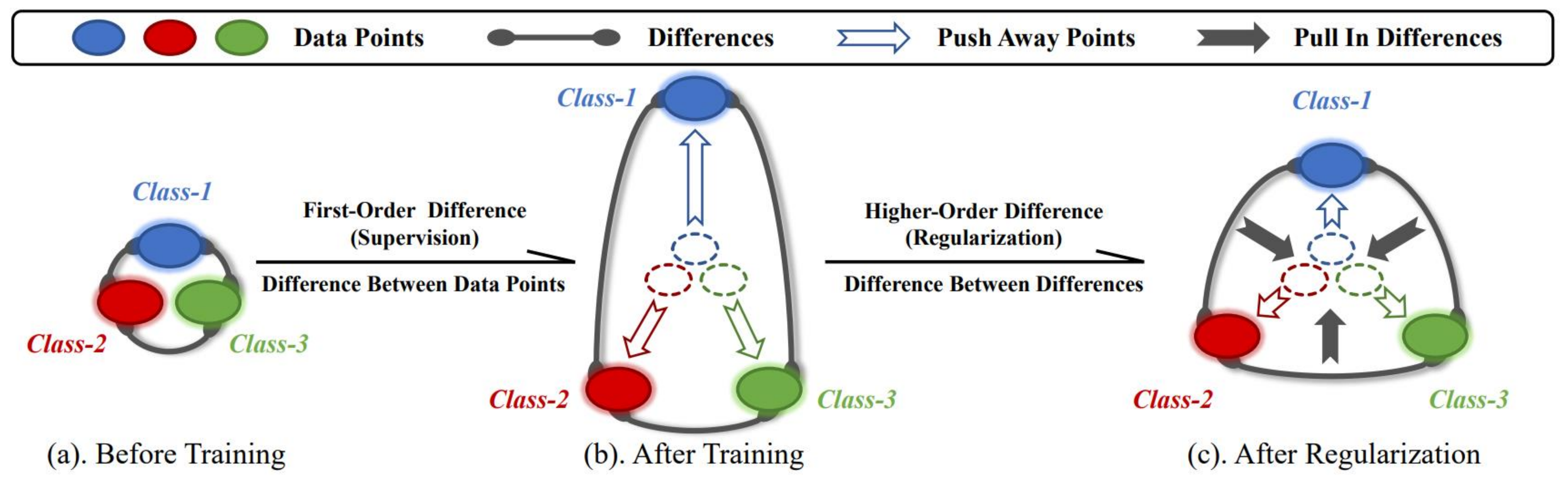
3. The University of Tokyo

Abstract

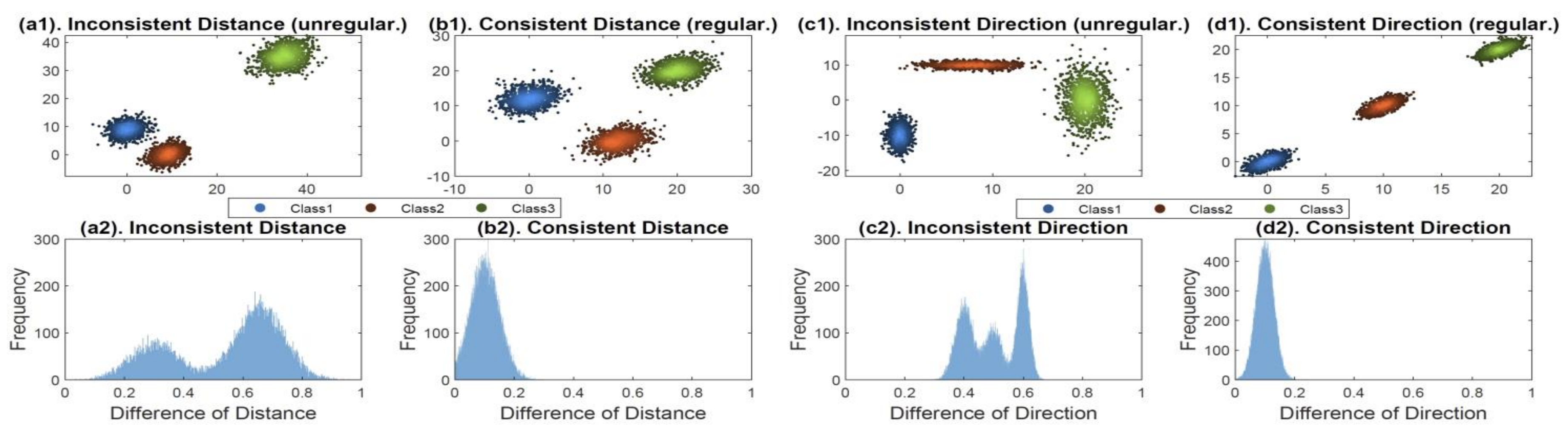
Similarity-based representation learning has shown impressive capabilities in both supervised (e.g., metric learning) and unsupervised (e.g., contrastive learning) scenarios. Existing approaches effectively constrained the representation difference (i.e., the disagreement between the embeddings of two instances) to fit the corresponding (pseudo) similarity supervision. However, most of them can hardly restrict the variation of representation difference, sometimes leading to overfitting results where the clusters are disordered by drastically changed differences. We propose a novel *difference alignment regularization* (DAR) to encourage all representation differences between inter-class instances to be as close as possible, so that the learning algorithm can produce consistent differences to distinguish data points from each other. To this end, we construct a new cross-total-variation (CTV) norm to measure the divergence among representation differences. Experiments on multi-domain data demonstrate the superiority of DASL in both supervised metric learning and unsupervised contrastive learning tasks.

Motivation

Minimizing higher-order difference to pull in first-order difference



The consistent/inconsistent differences (e.g., values and directions)



Formulation and Algorithm

Higher-Order Difference

$$\nabla_{\varphi}^{(2)}(\boldsymbol{x},\widehat{\boldsymbol{x}},\boldsymbol{z},\widehat{\boldsymbol{z}}) = \nabla_{\varphi}^{(1)}(\boldsymbol{x},\widehat{\boldsymbol{x}}) - \nabla_{\varphi}^{(1)}(\boldsymbol{z},\widehat{\boldsymbol{z}}),$$ *(the difference between differences)*

$$\sum\nolimits_{1\leq i<j\leq N,\, 1\leq k<l\leq N,\, (i,j)\neq(k,l), y_i\neq y_j, y_k\neq y_l} \mathcal{G}\left(\nabla_{\varphi}^{(2)}(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k,\boldsymbol{x}_l)\right)$$
$$= 2\left\|\left[\nabla_{\varphi}^{(1)}(\boldsymbol{x}_1,\boldsymbol{x}_2),\ldots,\nabla_{\varphi}^{(1)}(\boldsymbol{x}_i,\boldsymbol{x}_j),\ldots,\nabla_{\varphi}^{(1)}(\boldsymbol{x}_{N-1},\boldsymbol{x}_N)\right]_{1\leq i<j\leq N, y_i\neq y_j}\right\|_{\mathrm{ctv}}$$

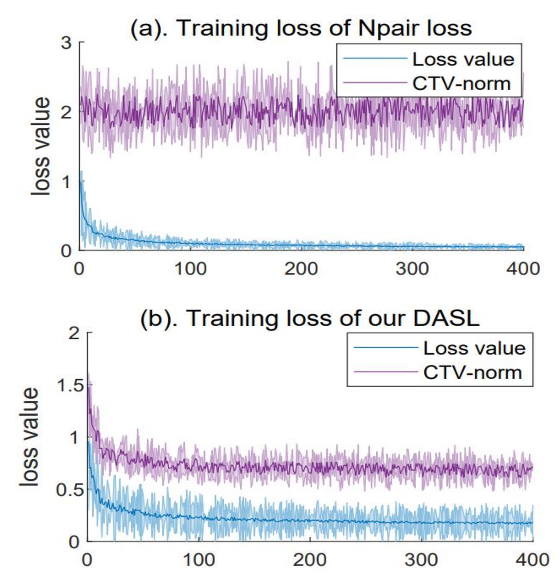
A CTV-Norm based Regularizer

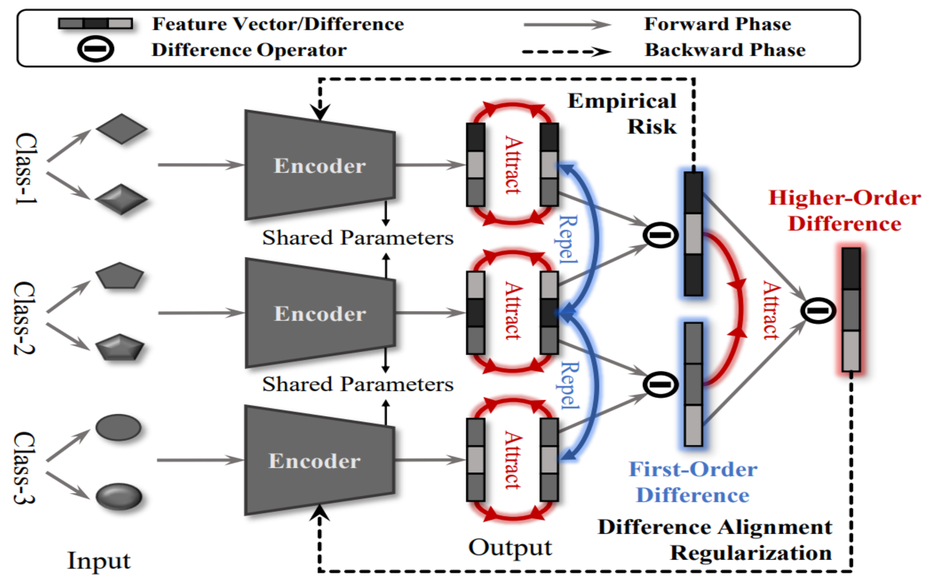
$$\min_{\varphi\in\mathcal{H}} \{\mathcal{F}(\varphi) = \mathcal{L}_{\mathrm{emp}}(\varphi;\mathscr{X}) + \lambda\mathcal{R}_{\mathrm{align}}(\varphi;\mathscr{X})\}$$ *(the regularizer is independent of L)*

$$\mathbb{E}_{\{b_j\}_{j=0}^{n}}\left\{\left\|\left[\nabla_{\varphi}^{(1)}(\boldsymbol{x}_{b_0},\boldsymbol{x}_{b_1}),\ldots,\nabla_{\varphi}^{(1)}(\boldsymbol{x}_{b_i},\boldsymbol{x}_{b_{i+1}}),\ldots,\nabla_{\varphi}^{(1)}(\boldsymbol{x}_{b_{n-1}},\boldsymbol{x}_{b_n})\right]\right\|_{\mathrm{ctv}}\right\}$$

Theorem 1. *The function $\|\cdot\|_{\mathrm{ctv}}:\mathbb{R}^{h\times H}\to\mathbb{R}^{+}$ is a strictly defined norm if and only if the measure function $\mathcal{G}(\cdot):\mathbb{R}^{h}\to\mathbb{R}^{+}$ is a strictly defined norm.*

Learning Framework and Training Curves





Theoretical Results

Distance Difference Bound

Theorem 2. *Suppose that the instances $\boldsymbol{x}$, $\widehat{\boldsymbol{x}}$, $\boldsymbol{z}$, and $\widehat{\boldsymbol{z}}$ are independently sampled from the same distribution as the training set $\mathscr{X}$. Then, for any feature representation φ learned from the objective $\mathcal{F}(\varphi)$, we have that with probability at least $1-\delta$,*

$$|d_{\varphi}(\boldsymbol{x},\widehat{\boldsymbol{x}}) - d_{\varphi}(\boldsymbol{z},\widehat{\boldsymbol{z}})| \leq \xi(\lambda)(\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|_2 + \|\boldsymbol{z}-\widehat{\boldsymbol{z}}\|_2)\max\{d_{\varphi}(\boldsymbol{t},\widehat{\boldsymbol{t}})|\boldsymbol{t},\widehat{\boldsymbol{t}}\in\mathscr{X}\}\sqrt{[\ln(2/\delta)]/(2N)}, \quad (9)$$

where $\xi(\lambda) = L\frac{\mathcal{L}_{emp}(\varphi^{(0)};\mathscr{X})}{\lambda}$ is monotonically decreasing w.r.t. the regularization parameter λ and the constant $L>0$ is independent of φ and $\mathscr{X}$.

Generalization Error Bound

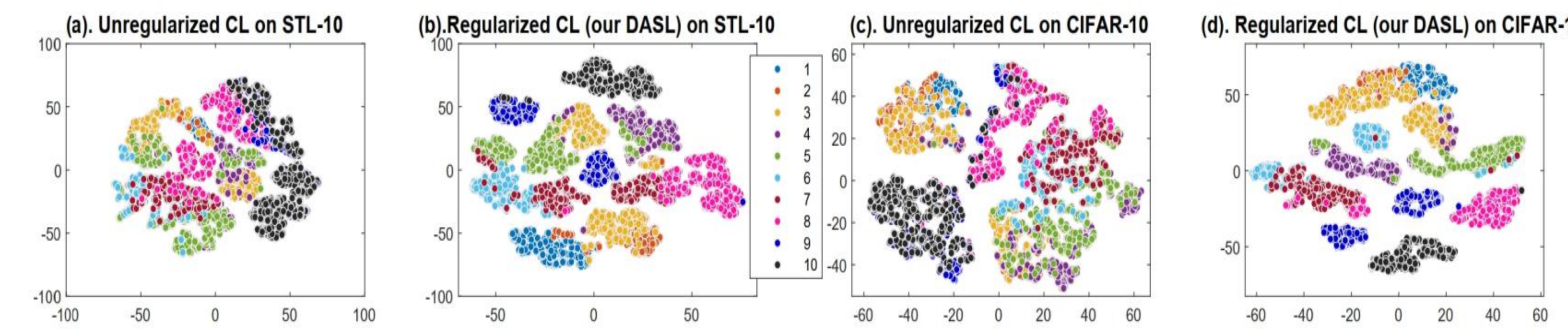
Theorem 3. *For any φ learned from the objective $\mathcal{F}(\varphi)$ and any given constant $\delta\in(0,1)$, we have that with probability at least $1-\delta$,*

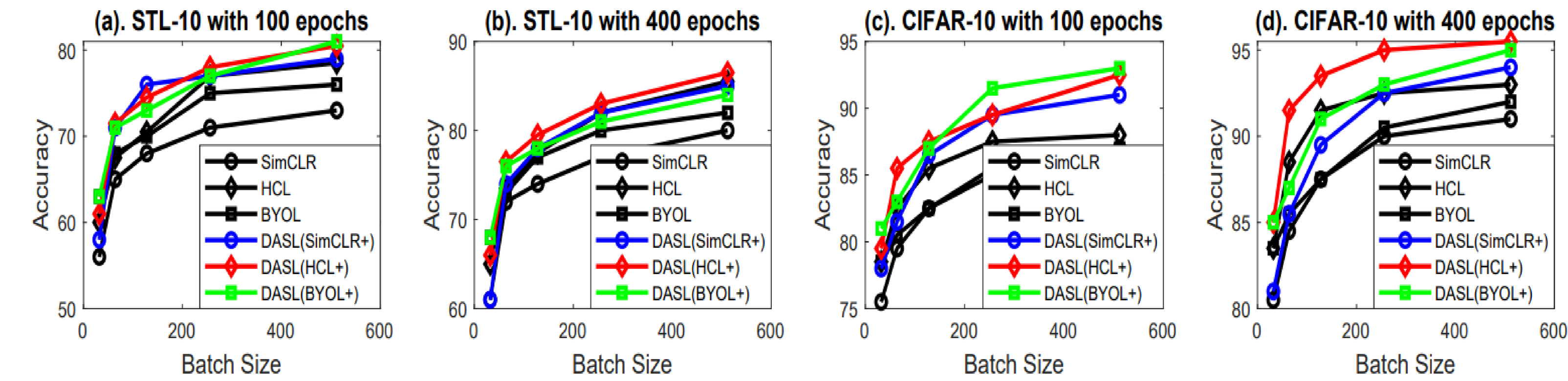
$$|\mathcal{L}(\varphi) - \widetilde{\mathcal{L}}(\varphi;\mathscr{D})| \leq \beta(\lambda)\omega(n)\log(1+\max\{d_{\varphi}(\boldsymbol{t},\widehat{\boldsymbol{t}})|\boldsymbol{t},\widehat{\boldsymbol{t}}\in\mathscr{X}\})\sqrt{[\ln(2/\delta)]/(2N)}, \quad (10)$$

where $\beta(\lambda) = (C+2/C)/\lambda$ is monotonically decreasing w.r.t. λ and $\omega(n) = \log\left(\frac{e^2}{n}+1\right)$ is monotonically decreasing w.r.t. n. Here the constant $C>0$ is independent of φ and $\mathscr{X}$.

Experimental Results

Visualization and Ablation Study





Experiments on Supervised Metric Learning

METHOD	CAR-196				CUB-200				SOP			
	NMI	R@1	R@4	R@8	NMI	R@1	R@4	R@8	NMI	R@1	R@10	R@100
Npair(Sohn, 2016)	69.50	82.57	94.97	95.92	69.53	64.52	85.63	91.15	91.11	76.21	88.43	92.08
ProxyA.(Kim et al., 2020b)	75.72	87.71	95.76	97.86	72.31	69.72	87.01	92.41	91.02	78.39	90.48	96.16
JDR(Chu et al., 2020)	70.56	84.86	94.56	97.21	70.32	69.44	87.01	91.33	92.21	79.21	90.53	96.01
IBC(Seidenschwarz et al., 2021)	74.82	88.11	96.21	98.21	74.01	70.32	87.61	92.72	92.61	81.42	91.32	95.89
AVSL(Zhang et al., 2022)	75.86	91.51	97.02	98.41	73.21	71.91	88.11	93.21	91.21	79.61	91.40	96.40
MetricF.(Yan et al., 2022)	76.23	91.76	96.31	97.21	75.41	**74.42**	85.75	92.53	92.71	82.23	92.62	96.33
ContextS.(Liao et al., 2023)	76.32	91.80	97.14	98.41	74.01	71.91	88.82	93.42	92.61	82.63	92.56	96.74
DASL-NP (ours)	75.96↑	86.34↑	97.56↑	98.87↑	73.52↑	69.63↑	89.62↑	93.61↑	92.85↑	79.21↑	93.21↑	97.86↑
DASL-PA (ours)	**77.32**↑	**92.31**↑	**97.82**↑	**98.90**↑	**76.50**↑	73.96↑	**90.54**↑	**94.21**↑	**93.86**↑	**83.32**↑	**93.86**↑	**97.95**↑

Experiments on Self-Supervised Contrastive Learning

METHOD	ImageNet-100						ImageNet-1K						#Arch.
	100 epochs			400 epochs			300 epochs			800 epochs			
	k-NN	Top-1	Top-5	k-NN	Top-1	Top-5	k-NN	Top-1	Top-5	k-NN	Top-1	Top-5	
SimCLR(Chen et al., 2020)	55.9	61.3	78.6	70.6	75.2	92.1	64.2	67.4	87.9	66.1	69.3	89.6	Res.50
BYOL(Grill et al., 2020)	56.3	65.5	77.8	69.2	73.2	90.1	66.9	71.2	90.5	67.2	73.2	91.5	Res.50
CMC(Tian et al., 2020a)	57.7	60.2	79.2	71.6	73.6	92.1	63.2	68.2	87.2	67.2	71.2	89.9	Res.50
PCL(Li et al., 2021)	55.9	60.2	77.2	71.5	76.1	93.2	59.5	66.5	86.7	62.2	70.5	90.5	Res.50
SwAV(Caron et al., 2020)	58.2	61.0	79.4	72.1	75.8	92.9	65.4	73.1	91.2	65.7	75.3	91.5	Res.50
HCL(Robinson et al., 2021)	55.9	60.8	79.3	70.2	74.6	92.3	64.2	71.2	91.2	67.2	71.7	90.7	Res.50
MetAug(Li et al., 2022)	59.2	61.1	79.4	69.8	75.6	93.2	65.4	74.2	91.1	67.8	76.0	92.9	Res.50
DASL (cluster-free)	60.5	65.2	79.8	73.5	76.6	93.9	68.3	72.7	91.9	68.2	**76.7**	92.9	Res.50
DASL (cluster-used)	**61.5**	**67.3**	**80.1**	**74.2**	**77.5**	**94.5**	**69.1**	**74.8**	**92.4**	**69.1**	76.6	**93.2**	Res.50
BYOL(Grill et al., 2020)	57.2	62.8	77.9	72.1	76.9	93.8	66.6	71.4	91.2	68.2	74.2	92.8	ViT-B/16
SwAV(Caron et al., 2020)	60.1	62.5	80.5	74.2	77.8	94.2	64.7	71.8	91.1	69.2	75.6	91.8	ViT-B/16
DINO(Caron et al., 2021)	61.5	67.5	81.8	78.2	79.2	95.5	72.3	76.1	92.4	76.2	78.2	94.2	ViT-B/16
iBOT(Zhou et al., 2022b)	61.5	68.2	82.2	77.5	78.5	95.2	71.5	75.0	91.9	75.2	76.0	92.6	ViT-B/16
PQCL(Zhang et al., 2023)	62.3	66.7	82.5	78.5	79.5	94.8	70.8	76.5	91.9	**78.3**	76.9	93.0	ViT-B/16
DASL (cluster-free)	62.5	69.5	81.1	78.4	80.1	96.1	71.5	**77.8**	92.7	76.2	79.2	94.5	ViT-B/16
DASL (cluster-used)	**63.4**	**69.7**	**82.8**	**79.3**	**82.3**	**96.8**	**72.9**	76.8	**93.5**	77.9	**79.9**	**96.3**	ViT-B/16

Acknowledgement

M.S. was supported by JST CREST Grant Number JPMJCR18A2 and a grant from Apple, Inc. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple Inc. J.Y. was supported by the National Science Fund of China under Grant No. 62361166670.