A Unified Mixture-View Framework for Unsupervised Representation Learning

Xiangxiang Chu[1] Xiaohang Zhan[2] Bo Zhang[1]

[1]Meituan [2]The Chinese University of Hong Kong

Summary

We propose an effective approach called Beyond Single Instance Multi-view (BSIM). Specifically, we impose more accurate instance discrimination capability by *measuring the joint similarity between two randomly sampled instances and their mixture*, namely **spurious-positive pairs**.

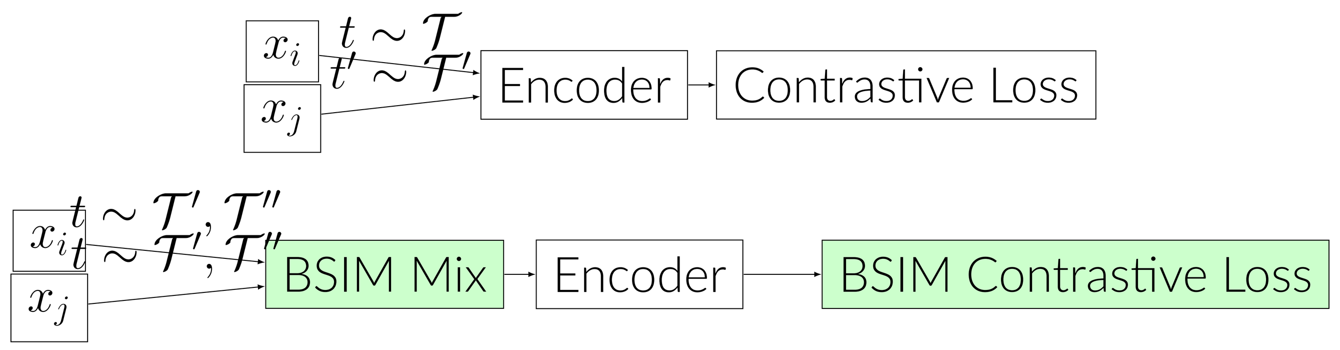

Figure 1. Our generic BSIM framework (b) serves as a plug-and-play adds-on for current contrastive learning paradigm (a). Note $\mathcal{T}$ and $\mathcal{T}'$ are augmentation policy distributions.

We apply it as an orthogonal improvement for unsupervised contrastive representation learning, including current outstanding methods SimCLR [2], MoCo [7], BYOL [6] and SimSiam [4]. We evaluate our learned representations on many downstream benchmarks like linear classification on ImageNet-1k and PASCAL VOC 2007, object detection on MS COCO 2017 and VOC, etc. We obtain substantial gains with a large margin almost on all these tasks compared with prior arts.

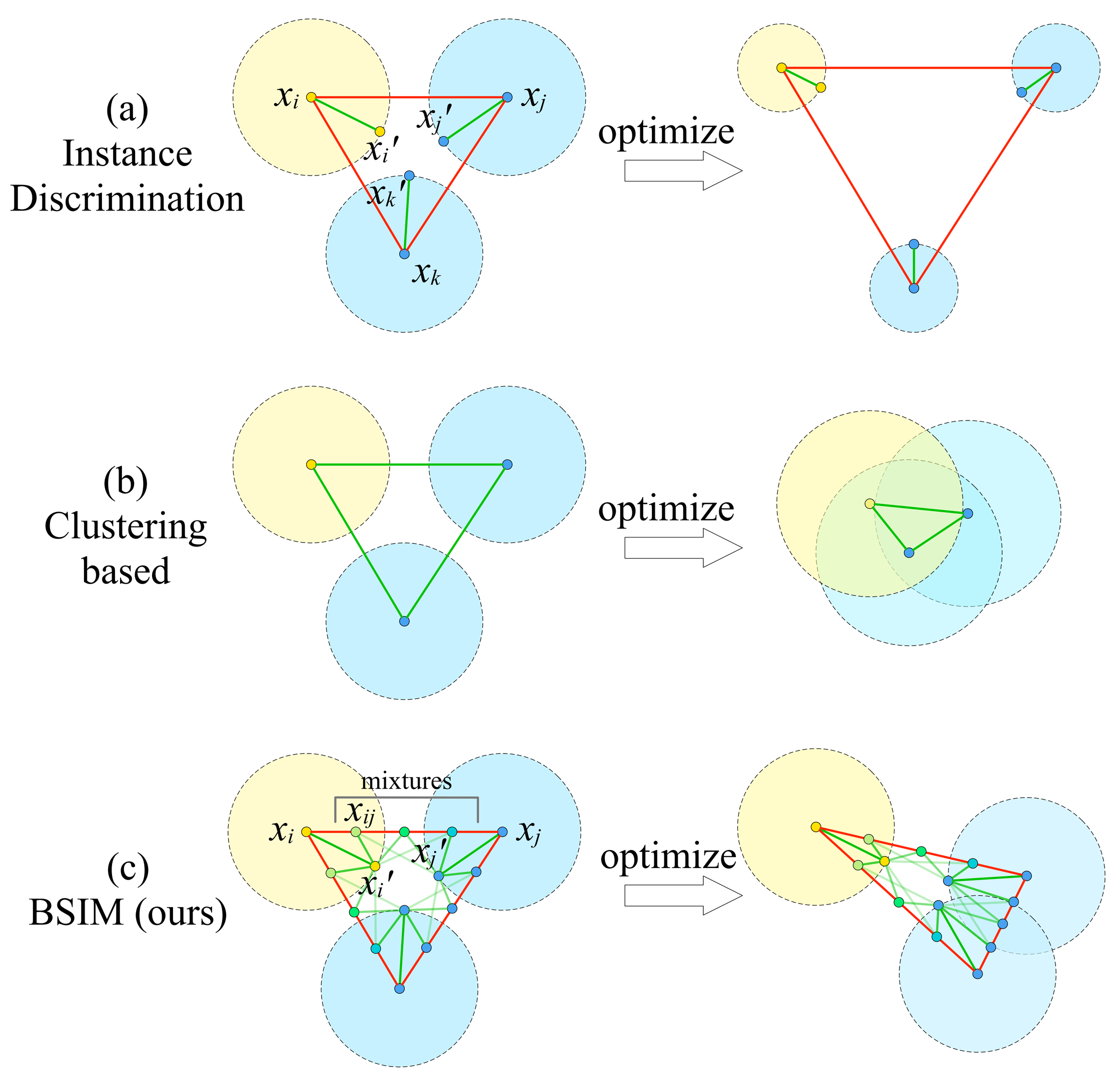

Figure 2. A schematic view of three self-supervised paradigms.

Related Work

- SimCLR [2] produces positive and negative pairs within a mini-batch of training data and chooses InfoNCE [8] loss to train the feature extraction backbone. It requires a large batch-size to effectively balance the positive and negative ones.
- MoCo [7] makes use of a feature queue to store negative samples, which greatly reduces high memory cost in [2]. Moreover, it proposes a momentum network to boost the consistency of features.
- BYOL [6] challenges the indispensability of negative examples and achieves impressive performance by only using positive ones. A mean square error loss is applied to make sure that positive pairs can predict each other.
- SimSiam [4] utilizes stop-gradient as an alternative method to avoid mode collapse, simplifying the design compared to prior arts.

Method

SimCLR-BSIM. SimCLR uses a single augmentation distribution, i.e. $\mathcal{T}'$ and $\mathcal{T}''$ are identical herein. The encoder network f encodes $x'_{1,2}$ as $f(x'_{1,2})$. Note $x'_{1,2}$ should show similarities with x''_1 as well as x''_2, which is measured by the sim function in the projected z space. We follow the definition in [2] for the similarity function as $\mathrm{sim}(z_i, z_j) = z_i^\top z_j/(\|z_i\|\|z_j\|)$. We use λ to regularize these similarities and the matching loss can be formulated as,

$$\begin{aligned} \ell'_i(\lambda) = & -\lambda \log \frac{e^{\mathrm{sim}(z'_{i,j}, z''_i)/\tau}}{\sum_{k=1}^{N}[e^{\mathrm{sim}(z'_{i,j}, z''_k)/\tau} + \cdot\, e^{\mathrm{sim}(z'_{i,j}, z'_{i,k})/\tau}]} \\ & -(1-\lambda) \log \frac{e^{\mathrm{sim}(z'_{i,j}, z''_j)/\tau}}{\sum_{k=1}^{N}[e^{\mathrm{sim}(z'_{i,j}, z''_k)/\tau} + \cdot\, e^{\mathrm{sim}(z'_{i,j}, z'_{i,k})/\tau}]}. \\ \text{where} \quad & = \begin{cases} 1 & k \notin \{i, j\} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

Similarly, we can formulate ℓ''_i if we use $x''_{1,2}$ as the anchor. Hence, the NT-Xent [2] loss is defined by the summation of each individual loss within the mini-batch data of size N as,

$$L_{\texttt{NT-Xent}}(\lambda) = \frac{1}{2N}\sum_{k=1}^{N} \ell'_i(\lambda) + \ell''_i(\lambda), \lambda \sim \beta(\alpha, \alpha). \tag{2}$$

SimCLR [2] has $2N$ positive pairs and $2N(N-1)$ negative ones in total at each iteration. Whereas, our method includes $4N$ spurious-positive pairs, i.e., $(x'_{i,j}, x''_i)$, $(x'_{i,j}, x''_j)$, $(x''_{i,j}, x'_i)$, $(x''_{i,j}, x'_j)$, and $2N(N-2)$ negative ones.

MoCo-BSIM. We produce the query q of MoCo by forwarding the mixed image controlled by λ.

$$\mathcal{L}_q = -\lambda \log \frac{\exp(q \cdot k_+^{\lambda}/\tau)}{\sum_{i=1}^{N}\exp(q \cdot k_i/\tau)} - (1-\lambda)\log \frac{\exp(q \cdot k_+^{1-\lambda}/\tau)}{\sum_{i=1}^{N}\exp(q \cdot k_i/\tau)} \tag{3}$$

where k_+^{λ} and $k_+^{1-\lambda}$ represent the corresponding key of images that produced the mixture respectively, and k_i are the keys in the current queue. τ is the softmax temperature.

BYOL-BSIM. BYOL-BSIM generates two s $x'_1 t'(x_1)$ and $x''_1 t''(x_1)$ from x_1 by applying respectively s $t' \sim \mathcal{T}'$ and $t'' \sim \mathcal{T}''$. Following the same procedure, we produce x'_2 and x''_2. Then we produce a new image $x'_{1,2}$ by λ-based mixture x'_1 and x'_2 through cutmix. The online network outputs $y'_{f}(x'_{1,2})$ and the projection $z'_{g}(y')$. The target network yields two ℓ_2-normalized projections $\bar{z}''_1$, $\bar{z}''_2$ from x''_1 and x''_2.

We sum up the MSE loss between the projection of the mixed image and its parents by the mixture coefficient λ. Formally, the loss is:

$$\mathcal{L}'_{,} = -2[\lambda \frac{\langle q'_{(}z'_{)}, z''_{i,}\rangle}{\|q'_{(}z'_{)}\|_2 \cdot \|z''_{i,}\|_2} + (1-\lambda)\frac{\langle q'_{(}z'_{)}, z''_{j,}\rangle}{\|q'_{(}z'_{)}\|_2 \cdot \|z''_{j,}\|_2}] \tag{4}$$

Note $z''_{i,}$ and $z''_{j,}$ mean the projection of the representation of x''_i and x''_j generated by the target network.

Experimental Results

Method	Epoch	SVM %mAP	SVM Low-Shot (%mAP) 1	2	4	8	16	32	64	96
Supervised	-	87.2	53.0	63.6	73.7	78.8	81.8	83.8	85.2	86.0
SimCLR [2]	200	79.0	32.5	40.8	50.4	59.1	65.5	70.1	73.6	75.4
SimCLR-BSIM	200	80.0	33.9	44.7	50.9	60.5	67.8	72.0	75.4	77.2
MoCo [7]	200	79.2	30.0	37.7	47.6	58.8	66.0	70.6	74.6	76.1
MoCoV2 [3]	200	83.8	43.7	55.2	63.2	71.5	75.4	79.1	81.2	82.0
MoCoV2-BSIM	200	84.8	**50.0**	53.9	**65.3**	72.4	76.3	79.3	81.7	82.8
MoCoV2-WBSIM	200	85.4	46.5	56.9	64.6	**74.7**	78.2	80.6	82.8	83.7
BYOL [6]	200	85.1	44.5	52.1	62.9	70.9	76.2	79.5	81.9	83.1
BYOL-BSIM	200	**86.5**	42.6	**55.9**	64.6	72.7	**78.8**	**81.9**	**83.6**	**84.6**
BYOL300 [6]	300	86.6	42.5	56.1	64.7	73.0	77.7	82.2	83.7	84.7
BYOL-BSIM300	300	87.6	**45.7**	54.5	66.4	75.0	79.8	83.2	**85.2**	86.0
BYOL-WBSIM300	300	**87.7**	44.1	**60.7**	**68.1**	**76.0**	**81.0**	**83.6**	**85.2**	**86.3**
SwAV [1]*	400	85.4	-	-	-	-	-	-	-	-

Table 1. ResNet-50 linear SVMs mAP on VOC07 [5] classification using two 224× 224 views. BYOL variants with "300" are trained for 300 epochs as [6]. *: SwAV is trained for 400 epochs.

Method	Epoch	Backbone							Top-1 Accuracy
InfoMin Aug [9]	200	R50	-	-	-	-	70.1	-	70.1
MoCo [7]	200	R50	15.3	33.1	44.7	57.3	60.6	61.0	61.0
SimCLR[2]	200	R50	17.1	31.4	41.4	54.4	61.6	60.1	61.6
SimCLR-BSIM	200	R50	**18.0**	32.5	42.7	55.3	62.3 (+0.7↑)	60.7	62.3 (+0.7↑)
MoCoV2 [3]	200	R50	14.7	32.8	45.0	61.6	66.7	67.5	67.5
MoCoV2-BSIM	200	R50	15.7	34.2	46.8	63.1	67.6	68.0 (+0.5↑)	68.0 (+0.5↑)
MoCoV2-WBSIM	200	R50	16.0	35.0	48.1	64.7	68.2	68.4 (+0.9↑)	68.4 (+0.9↑)
BYOL [6]	200	R50	16.7	34.2	46.6	60.8	69.1	67.1	69.1
BYOL-BSIM	200	R50	17.5	35.1	47.4	62.0	69.8 (+0.7↑)	67.9	69.8 (+0.7↑)
BYOL [6]†	300	R50	14.1	34.4	47.2	63.1	72.3	70.3	72.3
BYOL-BSIM	300	R50	16.4	**35.3**	48.5	65.1	72.7 (+0.4↑)	70.7	72.7 (+0.4↑)
BYOL-WBSIM	300	R50	15.4	**35.3**	**48.7**	**65.7**	**73.0** (**+0.7**↑)	**71.1**	**73.0** (**+0.7**↑)
SimSiam [4]	200	R50	-	-	-	-	70.0	-	70.0
SimSiam-BSIM [4]	200	R50	-	-	-	-	70.4(+0.4↑)	-	70.4(+0.4↑)
SimSiam-WBSIM [4]	200	R50	-	-	-	-	70.8(+0.8↑)	-	70.8(+0.8↑)
SwAV [1]	200	R50	-	-	-	-	69.1	-	69.1
SwAV [1]	400	R50	-	-	-	-	70.7	-	70.7

Table 2. Linear classification on ImageNet (top-1 center-crop accuracy on the validation set). All models are trained with two 224×224 views. †: reproduced. SwAV result is from SimSiam [4].

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33, 2020.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[8] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[9] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.