Towards Efficient and Effective Self-Supervised Learning of Visual Representations

Sravanti Addepalli*, Kaushal Santosh Bhogale*, Priyam Dey, R.Venkatesh Babu
Video Analytics Lab, Indian Institute of Science, Bangalore, India

Contributions

- We empirically show that a key reason for the slow convergence of instance similarity based approaches is the presence of noise in the training objective.
- We propose to strengthen the instance similarity based SSL algorithms using a noise free auxiliary training objective such as rotation prediction in a multi-task framework.
- We demonstrate significant gains in performance across CIFAR-10, CIFAR-100, ImageNet datasets.

Motivation

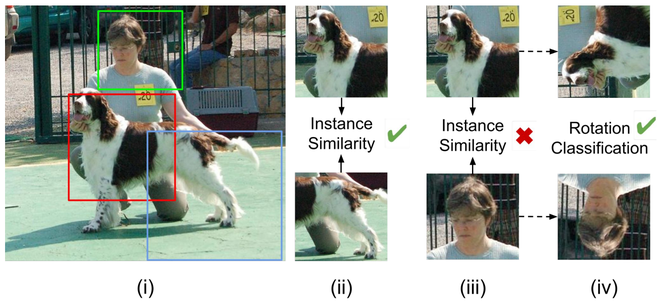

Fig.1: We demonstrate noise in the training objectives of instance-similarity based learning tasks. Consider the three random crops of the input image (i). The two crops in (ii) are desirable, while the crops shown in (iii) give an incorrect signal to the network. In (iv), we show that pretext tasks like rotation prediction can provide a noise-free training objective.

Impact of False Negatives and False Positives

Table 1: **Eliminating False Negatives** in contrastive learning across varying levels of supervision (% Labels). Elimination of noise in the training objective leads to higher linear evaluation accuracy (%) within a fixed training budget.

% Labels	SimCLR	Ours	Gain (%)
0	88.77	90.91	2.14
30	92.26 +3.49	93.94	1.68
50	92.93 +0.67	94.02	1.09
100	93.27 +0.34	94.15	0.88

Table 2: **Eliminating False Positives** in BYOL[1] across varying levels of supervision (%Good Crops). Elimination of noise in the training objective leads to higher linear evaluation accuracy (%) within a fixed training budget.

% Good Crops	BYOL	Ours	Gain (%)
0	63.64	68.62	4.98
25	64.50+0.86	68.30	3.80
50	66.30+1.80	68.90	2.60
100	66.72+0.42	70.26	3.54

Proposed Approach

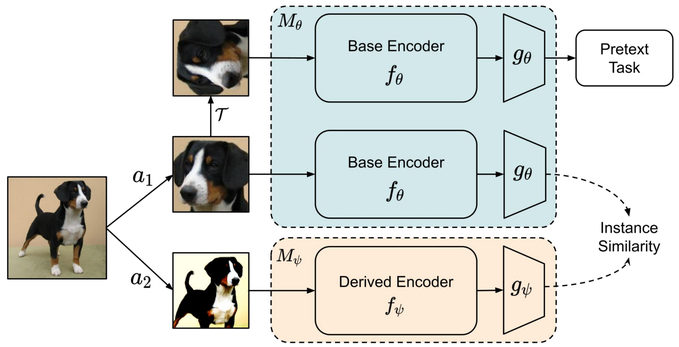

Fig.2: Schematic diagram illustrating the proposed approach. A pretext task such as rotation prediction is combined with base methods like BYOL and SimCLR. For methods like BYOL and MoCo, the derived network M_ψ is a momentum-averaged version of M_θ, and for methods like SimCLR and SimSiam, M_θ and M_ψ share the same parameters.

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda \cdot \frac{1}{2B} \sum_{i=0}^{B-1} \sum_{m=1}^{2} \ell_{CE}(h_\theta(M_\theta(x_i^{a_m,t_k}), t_k))$$

Efficiency and Effectiveness

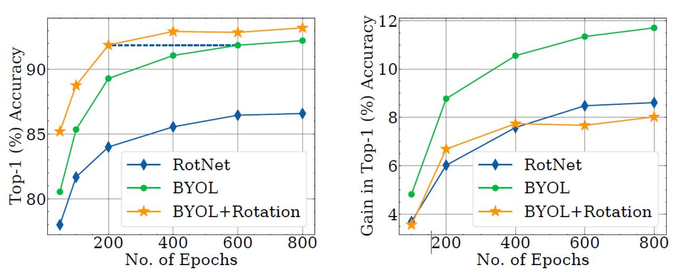

Fig. 3: Top-1 Accuracy (%, left plot) after self-supervised pretraining and Linear layer supervised training on CIFAR-10. The proposed method (BYOL+Rotation) achieves the same accuracy as the baseline in one-third the training time (blue dotted line). Gain in Top-1 Accuracy (%, right plot), is the difference between accuracy of the current epoch and epoch-50. The plots show improvements in efficiency and effectiveness of the proposed approach.

Experiments: Transfer Learning

Table 3: **Transfer Learning (Classification):** Performance (%) after linear evaluation on different datasets with a ResNet-50 backbone trained on ImageNet-1K for 30 epochs.

	ImageNet	CIFAR-10	CIFAR-100	Flowers	Caltech
SwAV	54.90	86.22	64.18	83.53	80.91
SwAV + Ours	**57.30**	**87.85**	**66.94**	**85.78**	**84.18**

	Aircraft	DTD	Cars	Food	Pets	SUN	VOC
SwAV	38.78	**69.79**	31.65	59.41	70.73	52.48	76.33
SwAV + Ours	**42.09**	69.68	**32.52**	**59.46**	**71.27**	**53.25**	**76.70**

Experiments: Linear Evaluation and Semi Supervised Learning

Table 4: Accuracy (%) of the proposed method under two evaluation settings - K Nearest Neighbor (KNN) classification with K=200 and Linear classifier training on CIFAR-10 and CIFAR-100; and under three evaluation settings - Linear classifier training and Semi-Supervised Learning with 1% and 10% labels on ImageNet-100.

Method	CIFAR-10 (200 epochs) KNN	CIFAR-10 (200 epochs) Linear	CIFAR-100 (200 epochs) KNN	CIFAR-100 (200 epochs) Linear
Rotation Pred.	78.01	84.00	36.25	50.87
SimCLR	86.37	88.77	55.10	62.96
SimCLR + Ours	88.69	90.91	57.09	65.40
BYOL	86.56	89.30	54.37	60.67
BYOL + Ours	89.80	91.89	58.41	67.03
SwAV	80.65	83.60	40.35	51.50
SwAV + Ours	85.26	87.20	50.09	58.60
SimSiam	87.05	89.77	56.90	64.27
SimSiam + Ours	**90.35**	**91.91**	**58.92**	**67.38**

Method	Linear Top-1	Linear Top-5	Semi-Supervised 1% labels Top-1	Semi-Supervised 1% labels Top-5	Semi-Supervised 10% labels Top-1	Semi-Supervised 10% labels Top-5
ImageNet-100 (100 epochs pretraining)						
Rotation Prediction	53.86	81.26	34.72	65.70	51.18	81.38
BYOL	71.02	91.78	46.60	75.50	68.00	89.80
BYOL + Ours	73.60	92.98	56.40	83.50	72.30	91.40
SimCLR	72.02	91.56	57.28	83.69	71.44	91.72
SimCLR + Ours	73.24	92.28	**57.80**	**83.84**	**72.52**	**92.10**
SwAV	72.20	92.96	49.38	78.41	67.56	90.78
SwAV + Ours	**74.40**	**93.33**	52.02	80.01	69.68	91.43

References: [1] Grill, Jean-Bastien, et al. "Bootstrap your own latent: A new approach to self-supervised learning." arXiv preprint arXiv:2006.07733 (2020).

Acknowledgements: This work was supported by the Qualcomm Innovation Fellowship. We are thankful for the support.