BMWReg: Brownian-diffusive, Multiview, Whitening Regularizations for Self-supervised Learning

Suhong Moon Domas Buracas Jinkyu Kim John Canny

University of California, Berkeley

Background: BYOL

BYOL learns class-discriminative representations by training an online network to predict the features of a target network, whose weights are a moving average of the online network's.

Each pair of embeddings are produced by feeding differently augmented views of the same image to the networks.

$$\mathcal{L}_{byol}(\theta, \xi; X) \coloneqq \|\hat{p} - \hat{z}`\|_2^2 = 2 - 2\frac{\langle p, z` \rangle}{\|p\|_2 \cdot \|z`\|_2}$$

Where $\hat{p} = p/\|p\|_2$, and $\hat{z}` = z`/\|z`\|_2$ are the normalized predictions from the online network and features from the target network.

Multiview, Brownian, and Whitening Losses

Multiview Centroid Loss

Minimizes the distance between each predicted feature $\hat{p}_j$ and the center of the target features $\hat{z}`_l$

$$\mathcal{L}_c(\theta; X) = \frac{1}{K}\sum_{j=1}^{K}\left\|\hat{p}_j - \frac{1}{K}\sum_{l=1}^{K}\hat{z}`_l\right\|_2^2$$

Brownian Diffusion Loss

Applies a different random gradient to each cluster of views, inducing Brownian motion, pushing the clusters apart

$$\mathcal{L}_b(\theta; X) = \frac{1}{K}\sum_{j=1}^{K}\langle \hat{n}, \hat{p}_j \rangle$$

$$n \sim \mathcal{N}(0, I_d)$$

Whitening Loss

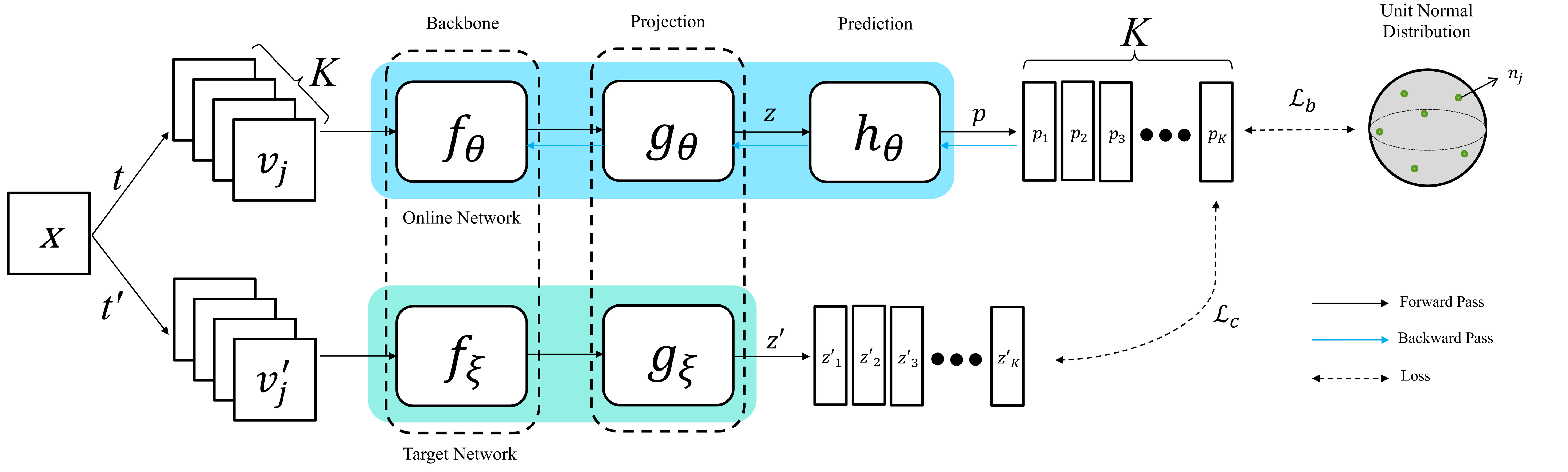
Normalizes the covariance matrix S_j of the predicted feature vectors to be orthogonal with unit norm I_d

$$\mathcal{L}_w(\theta; X) = \frac{1}{K}\sum_{j=1}^{K}\|S_j - I_d\|_F^2 = \frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{d}(\sigma_{ij} - 1)^2$$

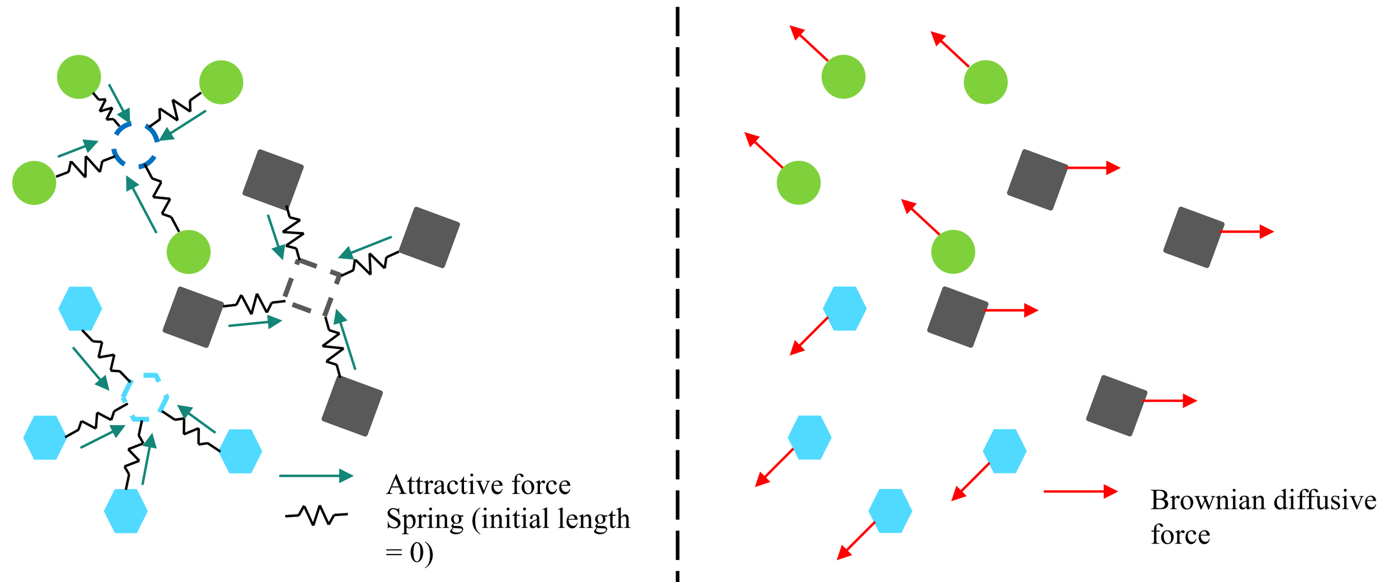
Combined Losses

$$\mathcal{L}_{ours}(\theta, \xi; X) = \mathcal{L}_c(\theta, \xi; X) + \lambda_b\mathcal{L}_b(\theta; X) + \lambda_w\mathcal{L}_w(\theta; X)$$

BMWReg Architecture



Visual Intuition



(A) Attractive force (B) Diffusive force

A. Multiview centroid loss applies an attractive force to representations of an object from multiple views, promoting semantically equivalent images to be represented similarly

B. Brownian Diffusion Loss repels representation clusters from different images

Experiments

Method	Architecture	ImageNet-100		
		Top-1 (%)	Top-5 (%)	5-NN (%)
BYOL† (Grill et al., 2020)	ResNet-18	71.56	91.18	63.18
MoCo† (He et al., 2020)	ResNet-50	72.80	91.64	-
Wang and Isola (Wang and Isola, 2020)	ResNet-50	74.60	92.74	-
W-MSE 4 (Ermolov et al., 2020)	ResNet-18	79.02	94.46	71.32
Ours ($K = 4$)	ResNet-18	80.38	94.92	74.3
Ours ($K = 8$)	ResNet-18	81.56	95.2	75.24

All our experiments are done with the ImageNet-100 dataset

In this table, we referenced the official implementation of MoCo and implemented BYOL ourselves.

This table shows our proposed method outperforming state-of-the-art baselines on ImageNet-100.

We observe that using a larger K, i.e. more augmented views is better than using less.

Also note that MoCo and Wang & Isola use ResNet-50, which is a more powerful feature extractor than ResNet-18.

Training Efficiency

Method	300	600	1200
BYOL (BS=1024)	74.48	74.98	N/A
BYOL (BS=2048)	75.44	78.1	79.08
Ours ($K = 4$)	80.38	N/A	N/A
Ours ($K = 8$)	81.56	N/A	N/A

Our method achieves better results with the same amount of compute. (Rows 1 & 3 and 2 & 4 have about the same cost per epoch)

Ablation Study

Method	Multiview	Brownian	Whitening	top-1 (%)
BYOL	✗	✗	✗	71.92
BYOL	✗	✗	✓	72.84
BYOL	✗	✓	✗	72.84
BYOL	✗	✓	✓	72.41
Ours	K=4	✗	✗	78.24
Ours	K=4	✗	✓	79.68
Ours	K=4	✓	✗	79.74
Ours	K=4	✓	✓	80.38
Ours	K=8	✗	✗	79.54
Ours	K=8	✗	✓	79.96
Ours	K=8	✓	✗	80.28
Ours	K=8	✓	✓	81.56

Each of our loss additions improve performance, with Multiview being the most significant

References

1. Jean-Bastien Grill, Florian Strub, Florent Altche, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya,Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Boot-strap your own latent: A new approach to self-supervised learning, 2020.
2. 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(CVPR), June 2020.
3. Tongzhou Wang and Phillip Isola. Understanding con-trastive representation learning through alignment and uniformity on the hypersphere, 2020.
4. Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto,and Nicu Sebe. Whitening for self-supervised representation learning.CoRR, abs/2007.06346, 2020. URL: `https://arxiv.org/abs/2007.06346`.

Berkeley UNIVERSITY OF CALIFORNIA