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 whose weights are a moving average of the online network’s. By feeding differently.

Multiview, Brownian, and Whitening Losses

Multiview Centroid Loss

Minimizes the distance between each predicted feature $\hat{y}_i$ and the center of the target features $\bar{y}$.

$$L_c(\theta, X) = \frac{1}{N} \sum_{i=1}^N \| \hat{y}_i - \bar{y} \|^2$$

Brownian Diffusion Loss

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

$$L_b(\theta, X) = \frac{1}{N} \sum_{i=1}^N \sum_{j \neq i} \| \hat{y}_i - \hat{y}_j \|^2$$

Whitening Loss

Normalizes the covariance matrix $S$ of the predicted feature vectors to be orthogonal with unit norm $I$.

$$L_w(\theta, X) = \frac{1}{N} \sum_{i=1}^N \| \hat{y}_i - \bar{y} \|^2$$

Combined Losses

$$L_{ours}(\theta, X) = L_c(\theta, X) + \lambda_w L_w(\theta, X)$$

Training Efficiency

<table>
<thead>
<tr>
<th>Method</th>
<th>300</th>
<th>600</th>
<th>1200</th>
</tr>
</thead>
<tbody>
<tr>
<td>BYOL (BS=1024)</td>
<td>74.48</td>
<td>74.98</td>
<td>N/A</td>
</tr>
<tr>
<td>BYOL (BS=2048)</td>
<td>75.44</td>
<td>78.1</td>
<td>79.08</td>
</tr>
<tr>
<td>Ours (K = 4)</td>
<td>80.38</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Ours (K = 6)</td>
<td>81.56</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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

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

References