Recent Progress on Self-Supervised Representation Learning
Self-Supervised Representation Learning
• Learn image features without human labels
• Map similar semantics closer
• Transferrable to downstream tasks

Instance discrimination:
• Augmentations of the same image has similar features
• Augmentations of different images has distinct features

Recent papers:
• SimCLR
• SwAV
• BYOL
• SimSiam
A simple framework for contrastive learning of visual representations (SimCLR)
Chen, Ting, et al. ICML 2020

Positive pair: \((i, j)\)
Augmentations of the same image

Negative pair:
Augmentations of different images

Contrastive loss (cosine similarity):
\[
\ell_{i,j} = -\log \frac{\exp({\text{sim}(z_i, z_j)/\tau})}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp({\text{sim}(z_i, z_k)/\tau})},
\]
A simple framework for contrastive learning of visual representations (SimCLR)
Chen, Ting, et al. ICML 2020

**Positive pair: (i, j)**
Augmentations of the same image

**Negative pair:**
Augmentations of different images

**Contrastive loss** (cosine similarity):
\[
\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)},
\]

Negative pairs prevent features **collapse** to be constant but require expensive large batch size.
Online clustering and predict the codes of an augmentation using the features of the other augmentation of the same image

\[ L(z_t; z_s) = \ell(z_t; q_s) + \ell(z_s; q_t), \]

\[ \ell(z_t; q_s) = - \sum_k q_s^{(k)} \log p_t^{(k)}, \]

where

\[ p_t^{(k)} = \frac{\exp \left( \frac{1}{\tau} z_t^\top c_k \right)}{\sum_{k'} \exp \left( \frac{1}{\tau} z_t^\top c_{k'} \right)}. \]

In short, augmentations of the same image should belong to the same cluster.
Unsupervised learning of visual features by contrasting cluster assignments (SwAV)
Caron, Mathilde, et al. NIPS 2020

Online clustering and predict the codes of an augmentation using the features of the other augmentation of the same image

\[ L(z_t, z_s) = \ell(z_t, q_s) + \ell(z_s, q_t), \]

\[ \ell(z_t, q_s) = -\sum_k q_s^{(k)} \log p_t^{(k)}, \]

where

\[ p_t^{(k)} = \frac{\exp \left( \frac{1}{\tau} z_t^\top c_k \right)}{\sum_{k'} \exp \left( \frac{1}{\tau} z_t^\top c_{k'} \right)}. \]

SwAV has no negative pairs and prevent collapsing by constraining that the samples in a batch should be equally partitioned by the clusters

In short, augmentations of the same image should belong to the same cluster
Bootstrap your own latent: A new approach to self-supervised learning (BYOL)
Grill, Jean-Bastien, et al. Arxiv 2020

**Encoder:** CNN

**Projector:** MLP

**Predictor:** MLP

**Update target model using mean teacher**

\[
\xi \leftarrow \tau \xi + (1 - \tau) \theta.
\]

**Stop-gradient**

**MSE Loss between final normalized features**

\[
\mathcal{L}_\theta^{\text{BYOL}} \triangleq \| \bar{q}_\theta(z_\theta) - \bar{z}_\xi' \|_2^2
\]
Bootstrap your own latent: A new approach to self-supervised learning (BYOL)
Grill, Jean-Bastien, et al. Arxiv 2020

**BYOL has no negative pairs and prevents collapsing by something unknown (probably the delicate balance provided by the mean teacher).**

MSE Loss between final normalized features:

\[
\mathcal{L}_{\theta}^{\text{BYOL}} \triangleq \| \overline{q_{\theta}}(z_{\theta}) - z'_{\xi} \|_2^2
\]
Maximize negative cosine similarities \( (\mathcal{D}) \) with stop gradient:

\[
\mathcal{L} = \frac{1}{2} \mathcal{D}(p_1, \text{stopgrad}(z_2)) + \frac{1}{2} \mathcal{D}(p_2, \text{stopgrad}(z_1)).
\]

*Mean teacher with zero momentum*
Exploring Simple Siamese Representation Learning (SimSiam)
Chen, Xinlei, and Kaiming He. Arxiv 2020

Maximize negative cosine similarities ($D$) with stop gradient:
\[
\mathcal{L} = \frac{1}{2} D(p_1, \text{stopgrad}(z_2)) + \frac{1}{2} D(p_2, \text{stopgrad}(z_1)).
\]

**Very simple**

SimSiam has no negative pairs and prevent collapsing by something more unknown (empirically only the stop-gradient operation)
Comparison

SimCLR

BYOL

SwAV

SimSiam
## Experiments

<table>
<thead>
<tr>
<th>method</th>
<th>batch size</th>
<th>negative pairs</th>
<th>momentum encoder</th>
<th>100 ep</th>
<th>200 ep</th>
<th>400 ep</th>
<th>800 ep</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimCLR (repro.)</td>
<td>4096</td>
<td>✓</td>
<td></td>
<td>66.5</td>
<td>68.3</td>
<td>69.8</td>
<td>70.4</td>
</tr>
<tr>
<td>MoCo v2 (repro.)</td>
<td><strong>256</strong></td>
<td>✓</td>
<td>✓</td>
<td>67.4</td>
<td>69.9</td>
<td>71.0</td>
<td>72.2</td>
</tr>
<tr>
<td>BYOL (repro.)</td>
<td>4096</td>
<td></td>
<td>✓</td>
<td>66.5</td>
<td><strong>70.6</strong></td>
<td><strong>73.2</strong></td>
<td><strong>74.3</strong></td>
</tr>
<tr>
<td>SwAV (repro.)</td>
<td>4096</td>
<td></td>
<td></td>
<td>66.5</td>
<td>69.1</td>
<td>70.7</td>
<td>71.8</td>
</tr>
<tr>
<td>SimSiam</td>
<td><strong>256</strong></td>
<td></td>
<td></td>
<td><strong>68.1</strong></td>
<td>70.0</td>
<td>70.8</td>
<td>71.3</td>
</tr>
</tbody>
</table>

Table 4. **Comparisons on ImageNet linear classification.** All are based on ResNet-50 pre-trained with two $224 \times 224$ views. Evaluation is on a single crop. All competitors are from our reproduction, and “+” denotes improved reproduction vs. original papers (see supplement).
Leaning signal for self-supervised learning:
- Invariant to augmentations
- Dataset intrinsic structure, e.g., ImageNet is intrinsically for clustering
- Model inductive bias and prior
- Method inductive bias and prior

Future directions for self-supervised learning:
- Augmentation >> From video, From 3D world
- Data >> Active video collection
- Model >> Better model dedicated for self-supervised learning
- Method >> Underlying mechanism and math