Understanding Self-Supervised Learning Dynamics without Contrastive Pairs

Mehmet F. Demirel

July 29, 2021

Contrastive vs Non-contrastive Learning

- **Contrastive**: Minimize the difference between positive pairs, and contrast negative pairs.
  - Former encourages **modeling invariances** while the latter prevents collapse.

- **Non-contrastive**: No negative pair contrasting, e.g., SimSiam and BYOL
  - Pair of Siamese networks: Want the views from the online + predictor network and the target network to match.
Non-contrastive Learning

• **Non-contrastive:** No negative pair contrasting, e.g., SimSiam and BYOL.

  • Pair of Siamese networks: Want the views from the online + predictor network and the target network to match.

  • Target network is not trained with GD, i.e., a direct copy or a momentum encoder.

  • Don’t require large batch size, memory queue, or negative pairs.
Why do these models not collapse?

- Predictor and stop-gradient is essential in non-contrastive SSL. BYOL and SimSiam collapse without either of these.
- EMA or momentum encoder is not necessary in BYOL and SimSiam.
- Both BYOL and SimSiam say that the predictor must be optimal in achieving minimal error when predicting the target network’s outputs from the online network’s.
- BYOL suggests that no weight decay leads to unstable results.
A simple model

Simple, bias-free, linear BYOL model.

Minimize \( J(W, W_p) = \frac{1}{2} \mathbb{E}_{x_1, x_2} \left[ \left\| W_p f_1 - \text{StopGrad}(f_{2a}) \right\|^2 \right] \)

\( W \): Linear online network

\( W_a \): Linear target network

\( W_p \): Linear predictor network

Figure 1. Two-layer setting with a linear, bias-free predictor.
Training Dynamics in Closed Form

Lemma 1. BYOL learning dynamics following Eqn. 1:

\begin{align*}
\dot{W}_p &= \alpha_p (-W_pW(X + X') + W_aX) W^\top - \eta W_p \quad (2) \\
\dot{W} &= W_p^\top (-W_pW(X + X') + W_aX) - \eta W \quad (3) \\
\dot{W}_a &= \beta (-W_a + W) \quad (4)
\end{align*}

Here, \( X := \mathbb{E} [\bar{x}\bar{x}^\top] \) where \( \bar{x}(\bar{x}) := \mathbb{E}_{x' \sim \text{aug}(\cdot \mid x)} [x'] \) is the average augmented view of a data point \( \bar{x} \) and \( X' := \mathbb{E}_x [\nabla_{x'}|x [x']] \) is the covariance matrix \( \nabla_{x'}|x [x'] \) of augmented views \( x' \) conditioned on \( x \), subsequently averaged over the data \( x \). Note that \( \alpha_p \) and \( \beta \) reflect multiplicative learning rate ratios between the predictor and target networks relative to the online network. Finally, the terms involving \( \eta \) reflect weight decay.
Exploration 1
Balancing of $W$ and $W_p$ that comes from weight decay

- In BYOL and SimSiam, the match between the representations produced by online and target networks cannot be explained solely by the predictor weights.

- A non-zero weight decay, parametrized by $\eta$, will remove the second term on the RHS $\implies$ more balance between online and predictor networks.

**Theorem 1** (Weight decay promotes balancing of the predictor and online networks.). Completely independent of the particular dynamics of $W_a$ in Eqn. 4, the update rules (Eqn. 2 and Eqn. 3) possess the invariance

$$W(t)W^\top(t) = \alpha_p^{-1}W_p(t)W_p(t) + e^{-2\eta t}C,$$

where $C$ is a symmetric matrix that depends only on the initialization of $W$ and $W_p$. 
Exploration 2

No predictor or no stop-gradient = collapse

- Shown to be true in empirical studies various times, but there is no theoretical explanation.

- When there is no EMA, meaning that $W_a = W$,

$$\frac{d}{dt} \text{vec}(W) = -H(t) \cdot \text{vec}(W)$$

where $H(t)$ is a PSD matrix. This implies that if the minimal eigenvalue of $H(t)$ is bounded below, $W(t) \to 0$, i.e., collapse.

Similar case for no predictor, i.e. $W_p = I$.

**Theorem 2** (The stop-gradient signal is essential for success.). With $W_a = W$ (SimSiam case), removing the stop-gradient signal yields a gradient update for $W$ given by positive semi-definite (PSD) matrix $H(t) := X' \otimes (W_p^T W_p + I_{n_2}) + X \otimes \dot{W}_p^T \dot{W}_p + \eta I_{n_1 n_2}$ (here $\dot{W}_p := W_p - I_{n_2}$ and $\otimes$ is the Kronecker product): 

$$\frac{d}{dt} \text{vec}(W) = -H(t) \text{vec}(W).$$

If the minimal eigenvalue $\lambda_{\min}(H(t))$ over time is bounded below, $\inf_{t \geq 0} \lambda_{\min}(H(t)) \geq \lambda_0 > 0$, then $W(t) \to 0$. 


**Assumptions**

- **Assumption (Isotropic Data and Augmentation):** $X = I$ and $X' = \sigma^2 I$, which comes from the assumptions that that data distribution $p(x)$ has zero mean and identity covariance whereas the augmentation distribution $p_{aug}(\cdot | x)$ has $x$ mean and $\sigma^2 I$ covariance.

- **Assumption (Proportional EMA):** EMA weight $W_a(t) = \tau(t)W(t)$ is a linear function of $W(t)$, i.e. $W_a$ points to the same direction as the online weight.

- **Assumption (Symmetric Predictor):** $W_p(t) = W_p^T(t)$
New Dynamics

- Under these assumptions, we have the following dynamics

\[
\begin{align*}
\dot{W}_p &= -\frac{\alpha_p}{2}(1 + \sigma^2)\{W_p, F\} + \alpha_p \tau F - \eta W_p \\
\dot{F} &= -(1 + \sigma^2)\{W_p^2, F\} + \tau\{W_p, F\} - 2\eta F
\end{align*}
\]

where \(\{A, B\} := AB + BA\) and \(F = \mathbb{E}[f_1f_1^\top] = WXW^\top\) is the correlation matrix of the input of the predictor \(W_p\).
Exploration 3
Eigenspace of $W_p$ aligns with $F$

**Theorem 3**: Under certain conditions,

$$FW_p - W_p F \to 0$$

and the eigenspace of $W_p$ and $F$ gradually aligns.
Decoupled Dynamics

Let columns of $U$ be the common eigenvectors of $W_p$ and $F$ so that $W_p = U\Lambda_{W_p}U^\top$ where $\Lambda_{W_p} = \text{diag}[p_1, \ldots, p_d]$, and $F = U\Lambda_FU^\top$ where $\Lambda_F = \text{diag}[s_1, \ldots, s_d]$.

\[
\begin{align*}
\dot{p}_j &= \alpha_p s_j \left[ \tau - (1 + \sigma^2)p_j \right] - \eta p_j \\
\dot{s}_j &= 2p_j s_j \left[ \tau - (1 + \sigma^2)p_j \right] - 2\eta s_j \\
s_j \dot{\tau} &= \beta(1 - \tau)s_j - \tau \dot{s}_j / 2.
\end{align*}
\]

Invariance holds: $s_j(t) = \alpha_p^{-1} p_j^2(t) + e^{-2\eta t} c_j$
No collapse

1D dynamics of the eigenvalue $p_j$ of $W_p$:

$$\dot{p}_j = p_j^2 \left[ \tau(t) - (1 + \sigma^2)p_j \right] - \eta p_j$$

- EMA
- Variance due to data augmentation
- Weight Decay

$$p_j = \tau - \frac{\sqrt{\tau^2 - 4\eta(1 + \sigma^2)}}{2(1 + \sigma^2)} \approx \frac{\eta}{\tau}$$

- Trivial Basin
- Non-trivial Basin
- Stable Trivial
- Stable Nontrivial

- Stable stationary point
- Unstable stationary point
**Effect of Weight Decay**

BUT, if weight decay is large, then the eigenspace alignment condition is more likely to satisfy!
Effect of Other Hyperparameters

\[ s_j(t) = \alpha_p^{-1} p_j^2(t) + e^{-2\eta t} c_j \]

<table>
<thead>
<tr>
<th></th>
<th>(\alpha_p) ↑</th>
<th>(\beta) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of trivial basin</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>Condition of eigenspace alignment</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>Training speed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenvalue of F won’t grow (no feature learning)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DirectPred

- Hmm, it looks like it is very important that eigenspaces of $W_p$ and $F$ align well. **Why not do this directly without relying on gradient descent?**

- Directly set linear $W_p$, so no optimization.

1. Estimate $\hat{F} = \rho \hat{F} + (1 - \rho) \mathbb{E}[ff^T]$

2. Eigen-decompose $\hat{F} = \hat{U} \Lambda_F \hat{U}^\top$, $\Lambda_F = \text{diag}[s_1, \ldots, s_d]$

3. Set $W_p = \hat{U} \text{diag}[p_j] \hat{U}^\top$ following the invariance $p_j = \sqrt{s_j} + \epsilon \max_j s_j$

**Eigenspaces are always and automatically aligned!**
## DirectPred

<table>
<thead>
<tr>
<th>Downstream Classification Top-1</th>
<th>Number of epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td><strong>STL-10</strong></td>
<td></td>
</tr>
<tr>
<td>DirectPred</td>
<td>77.86 ± 0.16</td>
</tr>
<tr>
<td>DirectPred (freq=5)</td>
<td>77.54 ± 0.11</td>
</tr>
<tr>
<td>SGD baseline</td>
<td>75.06 ± 0.52</td>
</tr>
<tr>
<td><strong>CIFAR-10</strong></td>
<td></td>
</tr>
<tr>
<td>DirectPred</td>
<td>85.21 ± 0.23</td>
</tr>
<tr>
<td>DirectPred (freq=5)</td>
<td>84.93 ± 0.29</td>
</tr>
<tr>
<td>SGD baseline</td>
<td>84.49 ± 0.20</td>
</tr>
</tbody>
</table>
**DirectPred**

Downstream classification (ImageNet):

<table>
<thead>
<tr>
<th>BYOL variants</th>
<th>Accuracy (60 ep)</th>
<th></th>
<th>Accuracy (300 ep)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-1</td>
<td>Top-5</td>
</tr>
<tr>
<td>2-layer predictor*</td>
<td>64.7</td>
<td>85.8</td>
<td>72.5</td>
<td>90.8</td>
</tr>
<tr>
<td>linear predictor</td>
<td>59.4</td>
<td>82.3</td>
<td>69.9</td>
<td>89.6</td>
</tr>
<tr>
<td>DirectPred</td>
<td>64.4</td>
<td>85.8</td>
<td>72.4</td>
<td>91.0</td>
</tr>
</tbody>
</table>

* 2-layer predictor is BYOL default setting.