Open the black-box of self-supervised learning.

Yuandong Tian
Research Scientist and Manager
Facebook AI Research
Great Empirical Success
Self-supervised Learning (SSL)

Reinforcement Learning (sparse reward signals)

Self-supervised Learning (dense signals)

Learning Representation without Human Label!

Why they work and achieve good performance? Can we do better?
Self-supervised Learning (SSL)

Dataset

$\mathbf{x} \sim p(\cdot)$

Data Augmentation

$\mathbf{x}_1, \mathbf{x}_2 \sim p_{\text{aug}}(\cdot | \mathbf{x})$

Network 1

Network 2

Feature for downstream tasks

Loss

facebook Artificial Intelligence
Similarity with Teacher Student Setting

\[
\mathbf{x} \sim p(\cdot) \\

\text{Student } \mathcal{W}_1 \\

\text{Teacher } \mathcal{W}_2 = \mathcal{W}^* \\

r := \|\mathbf{f}_{1,L} - \mathbf{f}_{2,L}\|^2
\]

The mathematical framework is similar!

[Y. Tian, Student Specialization in Deep ReLU Networks With Finite Width and Input Dimension, ICML 2020]
Contrastive versus Non-contrastive SSL

**Contrastive SSL**

Minimize distance $d_i$

Maximize distance $d_{ij}$

$L := - \sum_{i=1}^{N} \log \frac{\exp(-d_{ij}^2)}{\exp(-d_i^2) + \sum_{j \neq i} \exp(-d_{ij}^2)}$

**Non-contrastive SSL**

Minimize distance $d_i$

$L := \sum_{i=1}^{N} d_i^2$
Non-contrastive SSL (BYOL/SimSiam)

**BYOL**: [J. Grill, Bootstrap your own latent: A new approach to self-supervised Learning, NeurIPS 2020]

**SimSiam**: [X. Chen and K. He, Exploring Simple Siamese Representation Learning, CVPR 2021]
Non-contrastive SSL (BYOL/SimSiam)?

Why do they not collapse to trivial solutions?

Dataset

\(X \sim p(\cdot)\)

Data Augmentation

\(x_1, x_2 \sim p_{aug}(\cdot|X)\)

No Negative Pairs !!!

BYOL: [J. Grill, Bootstrap your own latent: A new approach to self-supervised Learning, NeurIPS 2020]

SimSiam: [X. Chen and K. He, Exploring Simple Siamese Representation Learning, CVPR 2021]
A simple model

Objective:

\[ J(W, W_p) := \frac{1}{2} \mathbb{E}_{x_1, x_2} \left[ \left\| W_p f_1 - \text{StopGrad}(f_{2a}) \right\|_2^2 \right] \]

Linear online network \( W \)

Linear target network \( W_a \)

Linear predictor \( W_p \)

\textbf{DirectPred} [Y. Tian et al, Understanding Self-Supervised Learning Dynamics without Contrastive Pairs, ICML'21 Outstanding Paper Honorable Mentions]
The Dynamics of Training Procedure

Lemma 1. **BYOL learning dynamics following Eqn. 1:**

\[
\begin{align*}
\dot{W}_p &= \alpha_p (-W_p W (X + X') + W_a X) W^\top - \eta W_p \\
\dot{W} &= W_p^\top (-W_p W (X + X') + W_a X) - \eta W \\
\dot{W}_a &= \beta (-W_a + W)
\end{align*}
\]

**Part I** Why we need (1) an **extra predictor** and (2) **stop-gradient**?

**Part II** Why the system doesn’t **collapse** to trivial solutions?

**Part III** The role played by different hyperparameters

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha_p)</td>
<td>Relative learning rate of the predictor</td>
</tr>
<tr>
<td>(\eta)</td>
<td>Weight decay</td>
</tr>
<tr>
<td>(\beta)</td>
<td>The rate of Exponential Moving Average (EMA)</td>
</tr>
</tbody>
</table>

**Part IV** Novel non-contrastive SSL algorithm **DirectPred**

\[
\bar{x}(x) := \mathbb{E}_{x' \sim \text{aug}(\cdot|x)} [x'] \\
X = \mathbb{E} [\bar{x} \bar{x}^\top] \quad \text{Covariance of the data} \\
X' = \mathbb{E}_x [\nabla_{x'} |_x [x']] \quad \text{Covariance of the augmentation}
\]
Part I  No Predictor / No Stop-Gradient do not work

If there is no EMA ($W = W_a$), then the dynamics becomes:

No Predictor

$$\dot{W} = -(X' + \eta I)W$$

PSD matrix

No Stop-Gradient (Here $\tilde{W}_p := W_p - I$)

$$\frac{d}{dt} \text{vec}(W) = - \left[ X' \otimes (W_p^T W_p + I) + X \otimes \tilde{W}_p^T \tilde{W}_p + \eta I_{n_1 n_2} \right] \text{vec}(W)$$

PSD matrix

In both cases, $W \rightarrow 0$
**Part II Assumptions**

*Assumption 1* (Isotropic Data and Augmentation): $X = I$ and $X' = \sigma^2 I$

*Assumption 2*: the EMA weight $W_a(t) = \tau(t)W(t)$ is a linear function of $W(t)$
Symmetrization of the dynamics

**Assumption 3** (Symmetric predictor $W_p$): $W_p(t) = W_p^T(t)$

$W_p$ becomes increasingly symmetric over training

Perfect symmetric $W_p$ might hurt training
Symmetrized Dynamics

Under the three assumptions, the dynamics becomes:

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

\{A, B\} := AB + BA is the anti-commutator.

Here \( F := E[ff^T] = WXW^T \) is the correlation matrix of the input of the predictor \( W_p \). \( F \) is well-defined even with nonlinear network.
Eigenspace Alignment

**Theorem 3**: Under certain conditions,

\[ FW_p - W_p F \to 0 \]

and the eigenspace of \( W_p \) and \( F \) gradually aligns.
Why non-contrastive SSL doesn’t collapse?

When eigenspace aligns, the dynamics becomes decoupled:

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

Where \(p_j\) and \(s_j\) are eigenvalues of \(W_p\) and \(F\)

Invariance holds: \(s_j(t) = \alpha_p^{-1} p_j^2(t) + e^{-2\eta t} c_j\)
Why non-contrastive SSL doesn’t 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**
Why non-contrastive SSL doesn’t 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

Stable stationary point
Unstable stationary point

Stable Trivial
Stable Nontrivial
Why non-contrastive SSL doesn’t collapse?

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

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

Variance due to data augmentation

Weight Decay

EMA

Stable
Trivial

Non-trivial Basin

Trivial Basin

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

Stable stationary point

Unstable stationary point
Part III The Effect of Weight Decay $\eta$

(a) No Weight Decay

(b) Weak Weight Decay

(c) Strong Weight Decay

$\eta = 0$

$p^*_j \sim \frac{\eta}{\tau}$

$\eta < \frac{\tau^2}{4(1+\sigma^2)}$

$\eta > \frac{\tau^2}{4(1+\sigma^2)}$

Saddle Point

Stable Trivial

Stable Nontrivial

Stable stationary point

Unstable stationary point
The Benefit of Weight Decay

Eigenspace alignment condition

\[ p_j [\tau - (1 + \sigma^2)p_j] < \frac{1}{2} [\alpha_p (1 + \sigma^2)s_j + 3\eta] \]

Higher weight decay \(\rightarrow\) alignment condition is more likely to satisfy!
Relative learning rate of the predictor $\alpha_p$

**Positive 😊**
1. Large $\alpha_p$ shrinks the size of trivial basin
2. Relax the condition of eigenspace alignment

**Negative 😞** With very large $\alpha_p$, eigenvalue of $F$ won’t grow (and no feature learning)
Exponential Moving Average rate $\beta$

$\beta$ large $\rightarrow W_a(t)$ catches $W(t)$ faster $\rightarrow \tau$ grows faster to 1

**Positive 🌼**: Slower rate (small $\beta$) relaxes the condition of eigenspace alignment

**Negative 🙁**: Slower rate makes the training slow and expands the size of trivial basin
Part IV DirectPred

• Directly setting linear $W_p$ rather than relying on gradient update.

1. Estimate $\hat{F} = \rho \hat{F} + (1 - \rho)E[f f^T]$
2. Eigen-decompose $\hat{F} = \hat{U} \Lambda_F \hat{U}^T$, $\Lambda_F = \text{diag} [s_1, s_2, ..., s_d]$
3. Set $W_p$ following the invariance:

$$p_j = \sqrt{s_j} + \epsilon \max_j s_j, \quad W_p = \hat{U} \text{diag}[p_j] \hat{U}^T$$

Guaranteed Eigenspace Alignment 😊
Performance of **DirectPred** on STL-10/CIFAR-10

<table>
<thead>
<tr>
<th>Downstream Classification Top-1</th>
<th>100</th>
<th>300</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>STL-10</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DirectPred</td>
<td>77.86 ± 0.16</td>
<td>78.77 ± 0.97</td>
<td>78.86 ± 1.15</td>
</tr>
<tr>
<td>DirectPred (freq=5)</td>
<td>77.54 ± 0.11</td>
<td><strong>79.90 ± 0.66</strong></td>
<td><strong>80.28 ± 0.62</strong></td>
</tr>
<tr>
<td>SGD baseline</td>
<td>75.06 ± 0.52</td>
<td>75.25 ± 0.74</td>
<td>75.25 ± 0.74</td>
</tr>
<tr>
<td></td>
<td><strong>CIFAR-10</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DirectPred</td>
<td><strong>85.21 ± 0.23</strong></td>
<td><strong>88.88 ± 0.15</strong></td>
<td><strong>89.52 ± 0.04</strong></td>
</tr>
<tr>
<td>DirectPred (freq=5)</td>
<td>84.93 ± 0.29</td>
<td>88.83 ± 0.10</td>
<td><strong>89.56 ± 0.13</strong></td>
</tr>
<tr>
<td>SGD baseline</td>
<td>84.49 ± 0.20</td>
<td>88.57 ± 0.15</td>
<td>89.33 ± 0.27</td>
</tr>
</tbody>
</table>
## Performance of DirectPred on ImageNet

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.

DirectPred using linear predictor is better than SGD with linear predictor, and is comparable with 2-layer predictor.
Summary

• A systematic analysis on the dynamics of non-contrastive self-supervised learning (SSL) methods
  • **Part I** Why we need (1) an extra predictor and (2) stop-gradient?
  • **Part II** Why training doesn’t collapse to trivial solutions?
  • **Part III** The role played by different hyperparameters

• Propose **DirectPred**, a novel non-contrastive SSL method
  • Directly align the eigenspace of the predictor $W_p$ with the correlation matrix $F$
  • Comparable performance in downstream classification tasks, compared to vanilla BYOL
    • CIFAR-10/STL-10
    • ImageNet (60 epochs / 300 epochs)

**Code:** [https://github.com/facebookresearch/luckmatters/tree/master/ssl](https://github.com/facebookresearch/luckmatters/tree/master/ssl)
Can we get rid of eigen-decomposition?

Propose \textbf{DirectSet}(\alpha):

\[ W_p = \frac{F^{\alpha}}{\|F^{\alpha}\|} + \epsilon I \]

\begin{align*}
\text{DirectPred} & \quad \alpha = 1/2 \\
\text{DirectCopy} & \quad \alpha = 1 \\
& \quad \text{(no eigen-decomp)}
\end{align*}

\begin{itemize}
\item \textbf{DirectCopy} [X. Wang, X. Chen, S. Du, Y. Tian, Towards Demystifying Representation Learning with Non-contrastive Self-supervision]
\end{itemize}
How DirectSet($\alpha$) learns the feature?

**Assumption 1** (Isotropic Data and Augmentation):
\( X = I \) and \( X' = \sigma^2 I \)

**Relaxed Assumption** \( X' = \sigma^2 P_B \)

\( P_B \): Nuisance Subspace

---

**Diagram:**
- Graph showing augmentation variance vs. feature index.
- Feature classification:
  - Nuisance feature: \( \sigma^2 > \frac{1}{4\eta} - 1 \)
  - Feature \( \to 0 \)
  - Invariant feature: \( \sigma^2 < \frac{1}{4\eta} - 1 \)
  - Feature \( \to \) positive value

---
Effect of Weight Decay $\eta$

Performance Peaked at $\eta = 4 \times 10^{-4}$

<table>
<thead>
<tr>
<th></th>
<th>Number of epochs</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>$\eta = 0$</td>
<td>71.94±0.93</td>
<td>78.53±0.40</td>
<td></td>
</tr>
<tr>
<td>$\eta = 0.0004$</td>
<td><strong>77.83±0.56</strong></td>
<td><strong>82.01±0.28</strong></td>
<td></td>
</tr>
<tr>
<td>$\eta = 0.001$</td>
<td>77.65±0.16</td>
<td>80.28±0.16</td>
<td></td>
</tr>
<tr>
<td>$\eta = 0.01$</td>
<td>58.12±0.94</td>
<td>58.53±0.76</td>
<td></td>
</tr>
<tr>
<td>$\eta = 0$</td>
<td>79.15±0.08</td>
<td>85.35±0.31</td>
<td></td>
</tr>
<tr>
<td>$\eta = 0.0004$</td>
<td><strong>84.02±0.37</strong></td>
<td><strong>89.17±0.12</strong></td>
<td></td>
</tr>
<tr>
<td>$\eta = 0.001$</td>
<td>83.91±0.33</td>
<td>87.75±0.16</td>
<td></td>
</tr>
<tr>
<td>$\eta = 0.01$</td>
<td>65.31±1.19</td>
<td>65.63±1.30</td>
<td></td>
</tr>
</tbody>
</table>
The role played by $\alpha$ in DirectSet($\alpha$)

$W \rightarrow \left(\frac{1+\sqrt{1-4\eta}}{2}\right)^{\frac{1}{2\alpha}} P_S$

The larger the $\alpha$, the larger the signal-noise ratio

Why not use $\alpha = 1$? No eigen-decomposition!

$P_S$: Invariant Subspace
### Experimental Result of DirectSet(\(\alpha\))

#### STL-10

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td>DirectCopy</td>
<td>77.83±0.56</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>
</tbody>
</table>

#### CIFAR-10

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td>DirectCopy</td>
<td>84.02±0.37</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>

#### CIFAR-100

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td>DirectCopy</td>
<td>55.40±0.19</td>
</tr>
<tr>
<td>DirectPred</td>
<td>56.60±0.27</td>
</tr>
<tr>
<td>DirectPred (freq=5)</td>
<td>56.43±0.21</td>
</tr>
<tr>
<td>SGD baseline</td>
<td>54.94±0.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ImageNet (100 epoch)</th>
<th>Report 2-layer baseline</th>
<th>DirectPred</th>
<th>DirectCopy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1 downstream accuracy</td>
<td>66.5</td>
<td>68.5</td>
<td>68.8</td>
</tr>
</tbody>
</table>
Beyond Linear Models

“Bad” eigenvalues bounce back later
Top-1 accuracy 89.62%

BYOL + linear predictor
“Bad” eigenvalues do not bounce back later
Top-1 accuracy 88.83%

BYOL + non-linear predictor
“Bad” eigenvalues bounce back later
Top-1 accuracy 90.25%
Contrastive Self-supervised Learning

Data Augmentation

$x_i, x'_i \sim p_{\text{aug}}(\cdot | \tilde{x}_i)$

Contrastive Loss

$\mathcal{W}$

Minimize distance $d_i$

Maximize distance $d_{ij}$

$\mathcal{X}_i$

Positive Sample

Current Sample

Negative Sample

$\mathcal{X}_j$


SimCLR

$\tilde{x}_i \sim p(\cdot)$

$x_i \sim p(\cdot)$

$x'_i \sim p(\cdot)$

$\mathcal{X}_i$

$\mathcal{X}_j$

$\mathcal{X}_i'$

$L := - \sum_{i=1}^{N} \log \frac{\exp(-d_i^2)}{\epsilon \exp(-d_i^2) + \sum_{j \neq i} \exp(-d_{ij}^2)}$
Contrastive SSL: Dimensional Collapsing

Shouldn’t contrastive SSL make full use of all dimensions? The answer is **No**...

Two puzzling questions:
1. Why contrastive SSL still has collapsing issues?
2. Why $L = 1$ doesn’t have collapsing, but $L \geq 2$ has the issue?

---

DirectCLR [L. Jing, P. Vincent, Y. LeCun, Y. Tian, Understanding Dimensional Collapse in Contrastive Self-supervised Learning]
The dynamics can be written down as follows:

\[
\frac{dW}{dt} = W (\Sigma_0 - \Sigma_{\text{Aug}})
\]

Inter-class covariance

\[
\Sigma_0 := \sum_{i,j} \alpha_{ij} (x_i - x_j)(x_i - x_j)^T
\]

Augmentation covariance

\[
\Sigma_{\text{Aug}} := \sum_i \left( \sum_{j \neq i} \alpha_{ij} \right) (x_i - x_i')(x_i - x_i')^T
\]

If \(\Sigma_0 - \Sigma_{\text{Aug}}\) has negative eigenvalues, then \(W\) will be low-rank.
Deep Model leads to Dimensional Collapsing

- What if $\Sigma_0 - \Sigma_{\text{Aug}}$ is PSD?
- Still dimensional collapsing for deep models.

1. $W_1$ and $W_2$ will align with each other.
2. The dynamics of their singular values satisfy

$$\dot{\sigma}_1^k = \sigma_1^k (\sigma_2^k)^2 (\nu_1^k)^T X \nu_1^k, \quad \dot{\sigma}_2^k = \sigma_2^k (\sigma_1^k)^2 (\nu_1^k)^T X \nu_1^k$$

$\sigma_1^k$ and $\sigma_2^k$ grow much faster for $k$ if $(\nu_1^k)^T X \nu_1^k$ is large.
DirectCLR

• If things are aligned, why not let them align directly?

<table>
<thead>
<tr>
<th>Loss function</th>
<th>Projector</th>
<th>Top-1 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimCLR</td>
<td>2-layer nonlinear projector</td>
<td>66.5</td>
</tr>
<tr>
<td>SimCLR</td>
<td>1-layer linear projector</td>
<td>61.1</td>
</tr>
<tr>
<td>SimCLR</td>
<td>no projector</td>
<td>51.5</td>
</tr>
<tr>
<td>DirectCLR</td>
<td>no projector</td>
<td>62.7</td>
</tr>
</tbody>
</table>
Thanks!