Residual Relaxation for Multi-view Representation Learning

Yifei Wang, Zhengyang Geng, Feng Jiang, CM Li,
Yisen Wang, Jiansheng Yang, Zhouchen Lin

Peking University

NeurIPS 2021

https://yifeiwang77.github.io
Motivation

• Multi-view methods become dominant for unsupervised learning
  • SimCLR, MoCo, BYOL, SimSiam, etc
  • For each input \( x \), we get two views, \( x_1, x_2 \) by random augmentation
  • Learn to align augmented views \( x_1, x_2 \) by minimizing representation distances

Observation
• Pretext (e.g. image augmentation) has a large effect on the final performance
• Some augmentations, like rotation, are too strong to be aligned exactly
• However, rotation is known as an effective signal for self-supervised learning
Motivation

• Multi-view methods become dominant for unsupervised learning
  • SimCLR, MoCo, BYOL, SimSiam, etc
  • For each input x, we get two views, x1, x2 by random augmentation
  • Learn to align augmented views x1, x2 by minimizing representation distances

• Observation
  • Pretext (e.g. image augmentation) has a large effect on the final performance
Motivation

• Multi-view methods become dominant for unsupervised learning
  • SimCLR, MoCo, BYOL, SimSiam, etc
  • For each input $x$, we get two views, $x_1$, $x_2$ by random augmentation
  • Learn to align augmented views $x_1$, $x_2$ by minimizing representation distances

• Observation
  • Pretext (e.g. image augmentation) has a large effect on the final performance
  • Some augmentations, like rotation, are too strong to be aligned exactly

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSiam [5]</td>
<td>91.8</td>
</tr>
<tr>
<td>SimSiam + margin loss</td>
<td>91.9</td>
</tr>
<tr>
<td>Rotation [9]</td>
<td>88.3</td>
</tr>
<tr>
<td>SimSiam + rotation aug.</td>
<td>87.9</td>
</tr>
<tr>
<td>SimSiam + Rotation loss</td>
<td>91.7</td>
</tr>
</tbody>
</table>
Motivation

• Multi-view methods become dominant for unsupervised learning
  • SimCLR, MoCo, BYOL, SimSiam, etc
  • For each input $x$, we get two views, $x_1, x_2$ by random augmentation
  • Learn to align augmented views $x_1, x_2$ by minimizing representation distances

• Observation
  • Pretext (e.g. image augmentation) has a large effect on the final performance
  • Some augmentations, like rotation, are too strong to be aligned exactly
  • However, rotation is known as an effective signal for Self-supervised Learning

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSiam [5]</td>
<td>91.8</td>
</tr>
<tr>
<td>SimSiam + margin loss</td>
<td>91.9</td>
</tr>
<tr>
<td>Rotation [9]</td>
<td>88.3</td>
</tr>
<tr>
<td>SimSiam + rotation aug.</td>
<td>87.9</td>
</tr>
<tr>
<td>SimSiam + Rotation loss</td>
<td>91.7</td>
</tr>
</tbody>
</table>
Motivation

• Multi-view methods become dominant for unsupervised learning

How to cultivate stronger augmentations (like rotation) to design better multi-view methods?

• Learn to align augmented views $x_1, x_2$ by minimizing representation distances

• Observation
  • Pretext (e.g. image augmentation) has a large effect on the final performance
  • Some augmentations, like rotation, are too strong to be aligned exactly
  • However, rotation is known as an effective signal for Self-supervised Learning

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSiam [5]</td>
<td>91.8</td>
</tr>
<tr>
<td>SimSiam + margin loss</td>
<td>91.9</td>
</tr>
<tr>
<td>Rotation [9]</td>
<td>88.3</td>
</tr>
<tr>
<td>SimSiam + rotation aug.</td>
<td>87.9</td>
</tr>
<tr>
<td>SimSiam + Rotation loss</td>
<td>91.7</td>
</tr>
</tbody>
</table>
What does not work...

- Direct combination of multi-view and pretext-predictive objectives
  - Pretext-invariance and Pretext-awareness
  - Two goals are contradictory to each other

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSiam [5]</td>
<td>91.8</td>
</tr>
<tr>
<td>SimSiam + margin loss</td>
<td>91.9</td>
</tr>
<tr>
<td>Rotation [9]</td>
<td>88.3</td>
</tr>
<tr>
<td>SimSiam + rotation aug.</td>
<td>87.9</td>
</tr>
<tr>
<td>SimSiam + Rotation loss</td>
<td>91.7</td>
</tr>
</tbody>
</table>
What does not work...

• Direct combination of pretext-invariant and pretext-aware objectives
  • Pretext-awareness and Pretext-invariance
  • **Two goals are contradictory to each other**

• Use a margin loss to relax the alignment

\[ \mathcal{L}_{\text{margin}}(x', x; \theta) = \max \left( \| \mathcal{G}_\theta(\mathcal{F}_\phi(x')) - \mathcal{F}_\phi(x) \|_2^2 - \eta, 0 \right) \]

• the representation space keeps shifting
• difficult to choose a universal tolerance

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSiam [5]</td>
<td>91.8</td>
</tr>
<tr>
<td>SimSiam + margin loss</td>
<td>91.9</td>
</tr>
<tr>
<td>Rotation [9]</td>
<td>88.3</td>
</tr>
<tr>
<td>SimSiam + rotation aug.</td>
<td>87.9</td>
</tr>
<tr>
<td>SimSiam + Rotation loss</td>
<td>91.7</td>
</tr>
</tbody>
</table>
What does not work...

- Direct combination of pretext-invariant and pretext-aware objectives

  - Pretext-awareness and Pretext-invariance
  - Two goals are contradictory to each other

- Use a margin loss to relax the alignment

\[
\mathcal{L}_{\text{margin}}(x', x; \theta) = \max\left(\left\| \mathcal{G}_\theta(\mathcal{F}_\theta(x')) - \mathcal{F}_\phi(x) \right\|^2_2 - \eta, 0\right)
\]

- the representation space keeps shifting
- difficult to choose a universal tolerance

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSiam [5]</td>
<td>91.8</td>
</tr>
<tr>
<td>SimSiam + margin loss</td>
<td>91.9</td>
</tr>
<tr>
<td>Rotation [9]</td>
<td>88.3</td>
</tr>
<tr>
<td>SimSiam + rotation aug.</td>
<td>87.9</td>
</tr>
<tr>
<td>SimSiam + Rotation loss</td>
<td>91.7</td>
</tr>
</tbody>
</table>
Our Solution: Residual Relaxation

• Use residuals to account for the semantic shift brought by augmentations

• **Exact alignment** fails for strong augmentation

\[
\begin{align*}
\mathbf{z}' & \rightarrow \leftarrow \mathbf{z} \\
\end{align*}
\]

• **Identity alignment** always holds instead

\[
\begin{align*}
\mathbf{z}' & \rightarrow \leftarrow \mathbf{z} + \mathbf{r} \\
\end{align*}
\]

• where \( \mathbf{r} = \mathbf{z}' - \mathbf{z} \) encodes the semantic shift
Pretext-aware Residual Relaxation (Prelax)

• Baseline: similarity loss for $x'=t(x)$

$$L_{\text{sim}}(x', x; \theta) = \|G_\theta(F_\theta(x')) - F_\phi(x)\|_2^2$$

• $F_\theta$ online network, $F_\phi$ target network, $G_\theta$ online prediction network
Pretext-aware Residual Relaxation (Prelax)

- Baseline: similarity loss
  \[ L_{\text{sim}}(x', x; \theta) = \| G_\theta(\mathcal{F}_\theta(x')) - \mathcal{F}_\phi(x) \|^2 \]

- \( F_\theta \) online network, \( F_\phi \) target network, \( G_\theta \) online prediction network

- Prelax (ours)
  - Exact Alignment -> Identity Alignment
    \[ G_\theta(z'_\theta) \leftrightarrow z_\phi \Rightarrow G_\theta(z'_\theta) - G_\theta(r) \leftrightarrow z_\phi \]

- Residual Relaxed Similarity (R2S) loss (\( \alpha \) is the interpolating coefficient)
  \[ L_{\text{R2S}}^\alpha(x', x; \theta) = \| G_\theta(\mathcal{F}_\theta(x')) - \alpha G_\theta(r) - \mathcal{F}_\phi(x) \|^2 \]
Pretext-aware Residual Relaxation (Prelax)

- Prelax (ours)
  - Residual Relaxed Similarity loss ($\alpha$ is the interpolating coefficient)
    \[ \mathcal{L}_{R2S}^\alpha(x', x; \theta) = \| \mathcal{G}_{\theta}(\mathcal{F}_{\theta}(x')) - \alpha \mathcal{G}_{\theta}(r) - \mathcal{F}_{\phi}(x) \|_2^2. \]
  - Predictive Learning (PL) Loss
    - the residual $r$ should encode the semantic shift caused by the augmentation
    - thus, we utilize $r$ to predict the corresponding augmentations of $x'$, denoted as $t$
    \[ \mathcal{L}_{PL}(x', x, t; \theta) = \text{CE}(\mathcal{H}_{\theta}^d(r), t^d) + \| \mathcal{H}_{\theta}^c(r) - t^c \|_2^2 \]

A non-conflicting combination of multi-view methods and predictive methods
Pretext-aware Residual Relaxation (Prelax)

• Prelax (ours)
  • Residual Relaxed Similarity loss ($\alpha$ is the interpolating coefficient)
    \[
    \mathcal{L}_{R2S}^{\alpha}(x', x; \theta) = \| \mathcal{G}_\theta(\mathcal{F}_\theta(x')) - \alpha \mathcal{G}_\theta(r) - \mathcal{F}_\phi(x) \|^2_2 .
    \]
  • Predictive Learning (PL) Loss
    \[
    \mathcal{L}_{PL}(x', x, t; \theta) = CE(\mathcal{H}_\theta^d(r), t^d) + \| \mathcal{H}_\theta^c(r) - t^c \|^2_2
    \]
  • Constraint on the Similarity
    • the residual is unbounded, and the distance between views could be very large
    • enforce small distance by adding a constraint
    \[
    \mathcal{L}_{sim} = \| \mathcal{G}_\theta(\mathcal{F}_\theta(x')) - \mathcal{F}_\phi(x) \|^2_2 \leq \varepsilon
    \]
Pretext-aware Residual Relaxation (Prelax)

- Prelax (ours)
  - Residual Relaxed Similarity loss (\( \alpha \) is the interpolating coefficient)
    \[
    \mathcal{L}_{R2S}^\alpha (x', x; \theta) = \| G_\theta (F_\theta (x')) - \alpha G_\theta (r) - F_\phi (x) \|_2^2.
    \]
  - Predictive Learning (PL) Loss
    \[
    \mathcal{L}_{PL} (x', x, t; \theta) = CE (H_\theta^d (r), t^d) + \| H_\theta^c (r) - t^c \|_2^2
    \]
  - Constraint on the Similarity
    \[
    \mathcal{L}_{sim} = \| G_\theta (F_\theta (x')) - F_\phi (x) \|_2^2 \leq \varepsilon
    \]
Pretext-aware Residual Relaxation (Prelax)

• Prelax (ours)
  • Residual Relaxed Similarity loss ($\alpha$ is the interpolating coefficient)
    \[ L_{R2S}^\alpha(x', x; \theta) = \|G_\theta(F_\theta(x')) - \alpha G_\theta(r) - F_\phi(x)\|_2^2. \]
  • Predictive Learning (PL) Loss
    \[ L_{PL}(x', x, t; \theta) = CE(H^d_\theta(r), t^d) + \|H^c_\theta(r) - t^c\|_2^2 \]
  • Constraint on the Similarity
    \[ L_{sim} = \|G_\theta(F_\theta(x')) - F_\phi(x)\|_2^2 \leq \varepsilon \]
  • Combined
    \[
    \min_\theta L_{R2S}^\alpha(x', x; \theta) + \gamma L_{PL}(x', x; \theta),
    \]
    \[ s.t. \quad \|G_\theta(F_\theta(x')) - F_\phi(x)\|_2^2 \leq \varepsilon. \]
    \[
    \text{Penalized} \quad L_{R2S}^\alpha(x', x; \theta) + \gamma L_{PL}(x', x; \theta) + \beta L_{sim}(x', x; \theta),
    \]
Pretext-aware Residual Relaxation (Prelax)

• Theoretical results
  • Prelax provably enjoys better downstream performance
  • An information-theoretical characterization
    • X input, T downstream task, S self-supervised signal, Z representation
    • $S_V$: multi-view learning, $S_a$: predictive learning
    • Goal: maximize mutual information $I(Z;T)$ with downstream task
Pretext-aware Residual Relaxation (Prelax)

- Theoretical results
  - Prelax provably enjoys better downstream performance
  - An information-theoretical characterization
    - $\mathbf{X}$ input, $\mathbf{T}$ downstream task, $\mathbf{S}$ self-supervised signal, $\mathbf{Z}$ representation
    - $S_v$: multi-view learning, $S_a$: predictive learning
    - Goal: maximize mutual information $I(\mathbf{Z}; \mathbf{T})$ with downstream task
  - Prelax extracts more task-relevant information than multi-view ($\mathbf{Z}_{mv}$) and predictive ($\mathbf{Z}_{PL}$) methods

**Theorem 1.** Assume that by maximizing the mutual information, each method can retain all information in $\mathbf{X}$ about the learning signal $\mathbf{S}$ (or $\mathbf{T}$), i.e., $I(\mathbf{X}; \mathbf{S}) = \max_{\mathbf{Z}} I(\mathbf{Z}; \mathbf{S})$. Then we have the following inequalities on their task-relevant information $I(\mathbf{Z}; \mathbf{T})$:

$$ I(\mathbf{X}; \mathbf{T}) = I(\mathbf{Z}_{\text{sup}}; \mathbf{T}) \geq I(\mathbf{Z}_{\text{Prelax}}; \mathbf{T}) \geq \max(I(\mathbf{Z}_{mv}; \mathbf{T}), I(\mathbf{Z}_{PL}; \mathbf{T})). $$ (10)
Pretext-aware Residual Relaxation (Prelax)

• Theoretical results
  • Prelax provably enjoys better downstream performance
  • An information-theoretical characterization
    • X input, T downstream task, S self-supervised signal, Z representation
    • S_v: multi-view learning, S_a: predictive learning
    • Goal: maximize mutual information I(Z;T) with downstream task
  • Prelax extracts more task-relevant information than multi-view (Z_{mv}) and predictive (Z_{PL}) methods
  • As a result, Prelax has a tighter upper bound on the downstream Bayes error

Theorem 2. Further assume that T is a K-class categorical variable. In general, we have the upper bound $u^e$ on the downstream Bayes errors $P^e := \mathbb{E}_z [1 - \max_{t \in T} P (T = t | z)]$,

$$\bar{P}^e \leq u^e := \log 2 + P^e_{\sup} \cdot \log K + I(X; T|S).$$ (11)

where $\bar{P}^e = \text{Th}(P^e) = \min \{\max \{P^e, 0\}, 1 - 1/K\}$ denotes the thresholded Bayes error. Accordingly, we have the following inequalities on the upper bounds for different unsupervised methods,

$$u^e_{\sup} \leq u^e_{\text{Prelax}} \leq \min(u^e_{\text{mv}}, u^e_{\text{PL}}) \leq \max(u^e_{\text{mv}}, u^e_{\text{PL}}).$$ (12)
Practical Implementations of Prelax

• Backbone (e.g. SimSiam) between two augmented views $\mathbf{x}_1, \mathbf{x}_2$

$$\mathcal{L}_{\text{Simsiam}}(x; \theta) = \|G_\theta(F_\theta(x_1)) - F_\phi(x_2)\|_2^2 + \|G_\theta(F_\theta(x_2)) - F_\phi(x_1)\|_2^2$$
Practical Implementations of Prelax

• Backbone (e.g. SimSiam) between two augmented views $x_1, x_2$

$$\mathcal{L}_{\text{Simsiam}}(x; \theta) = \|G_{\theta}(F_{\theta}(x_1)) - F_{\phi}(x_2)\|^2_2 + \|G_{\theta}(F_{\theta}(x_2)) - F_{\phi}(x_1)\|^2_2$$

• Prelax-std: generalize baselines with existing augmentations
  • Residual
    $$r_{12} = F_{\theta}(x_1) - F_{\theta}(x_2)$$
  • Prelax-std objective

$$\mathcal{L}_{\text{Prelax-std}}(x; \theta) = \mathcal{L}^\alpha_{R2S}(x_1, x_2; \theta) + \gamma \mathcal{L}_{PL}(x_1, x_2, t_1; \theta) + \beta \mathcal{L}^\text{sim}_{\text{sim}}(x_2, x_1; \theta).$$
Practical Implementations of Prelax

• Backbone (e.g. SimSiam) between two augmented views $x_1, x_2$

$$L_{\text{Simsiam}}(x; \theta) = \|G_{\theta}(F_{\theta}(x_1)) - F_{\phi}(x_2)\|^2_2 + \|G_{\theta}(F_{\theta}(x_2)) - F_{\phi}(x_1)\|^2_2$$

• Prelax-std: generalize baselines under existing augmentations

• Prelax-rot: incorporating stronger augmentation (rotation)
  • a third view $x_3$ as a randomly rotated $x_1$, residual (for rotation) $r_{31} = z_3 - z_1$
  • Rotation Residual Relaxation Similarity (R3S) loss

$$L_{\text{R3S}}^{\alpha}(x_{1:3}; \theta) = \|G_{\theta}(F_{\theta}(x_3)) - \alpha G_{\theta}(r_{31}) - F_{\phi}(x_2)\|^2_2.$$  

• Combined

$$L_{\text{Prelax-rot}}(x; \theta) = L_{\text{R3S}}^{\alpha}(x_{1:3}; \theta) + \gamma L_{\text{PL}}^{\text{rot}}(x_1, x_3, a; \theta) + \beta L_{\text{sim}}(x_2, x_1; \theta).$$
Practical Implementations of Prelax

• Backbone (e.g. SimSiam) between two augmented views $x_1, x_2$

$$\mathcal{L}_{\text{Simsiam}}(x; \theta) = \|G_\theta(F_\theta(x_1)) - F_\phi(x_2)\|_2^2 + \|G_\theta(F_\theta(x_2)) - F_\phi(x_1)\|_2^2$$

• Prelax-std: generalize baselines under existing augmentations
• Prelax-rot: incorporating stronger augmentation (rotation)
• Prelax-all: best of both worlds

$$\mathcal{L}_{\text{Prelax-all}}(x; \theta) = \frac{1}{2} \left( \mathcal{L}_{R2S}^{a_1}(x_1, x_2; \theta) + \mathcal{L}_{R3S}^{a_2}(x_1:3; \theta) \right) + \frac{\gamma_1}{2} \mathcal{L}_{\text{PL}}(x_1, x_2, t_1; \theta)$$
$$+ \frac{\gamma_2}{2} \mathcal{L}_{\text{PL}}^{\text{rot}}(x_1, x_3, a; \theta) + \beta \mathcal{L}_{\text{sim}}(x_2, x_1; \theta),$$
Experiments

- Two backbone methods: SimSiam and BYOL
- Two benchmark datasets: CIFAR-10 and ImageNette (10 classes from ImageNet)
- Default hyperparameters + ResNet-18

Table 1: Linear evaluation on CIFAR-10 (a) and ImageNette (b) with ResNet-18 backbone. TTA: Test-Time Augmentation.

<table>
<thead>
<tr>
<th>(a) CIFAR-10.</th>
<th>Method</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised [12] (re-produced)</td>
<td>95.0</td>
<td></td>
</tr>
<tr>
<td>Rotation [9] (re-produced)</td>
<td>88.3</td>
<td></td>
</tr>
<tr>
<td>BYOL [10] (re-produced)</td>
<td>91.1</td>
<td></td>
</tr>
<tr>
<td>SimCLR [2]</td>
<td>91.1</td>
<td></td>
</tr>
<tr>
<td>SimSiam [5]</td>
<td>91.8</td>
<td></td>
</tr>
<tr>
<td>SimSiam + Prelax</td>
<td>93.4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) ImageNette.</th>
<th>Method</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>91.0</td>
<td></td>
</tr>
<tr>
<td>Supervised + TTA</td>
<td>92.2</td>
<td></td>
</tr>
<tr>
<td>BYOL [10] (ResNet-18)</td>
<td>91.9</td>
<td></td>
</tr>
<tr>
<td>BYOL [10] (ResNet-50)</td>
<td>92.3</td>
<td></td>
</tr>
<tr>
<td>BYOL + Prelax (ResNet-18)</td>
<td>92.6</td>
<td></td>
</tr>
</tbody>
</table>
Experiments

- Effectiveness of Prelax-variants
  - Three benchmark datasets

- In-domain linear evaluation

- Out-of-domain linear evaluation

- Residual Relaxation can benefit from both existing (Prelax-std) and stronger (Prelax-rot) augs
Experiments

• Empirical understandings

(a) Representation visualization.  
(b) Nearest image retrieval.
Experiments

- Ablation Study
  - best among alternative algorithmic options
  - each component is necessary in Prelax

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSiam [5]</td>
<td>91.8</td>
</tr>
<tr>
<td>SimSiam + margin loss</td>
<td>91.9</td>
</tr>
<tr>
<td>Rotation [9]</td>
<td>88.3</td>
</tr>
<tr>
<td>SimSiam + rotation aug.</td>
<td>87.9</td>
</tr>
<tr>
<td>SimSiam + Rotation loss</td>
<td>91.7</td>
</tr>
<tr>
<td>SimSiam + Prelax (ours)</td>
<td>93.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prelax-std (R2S + Sim + PL)</td>
<td>92.5</td>
</tr>
<tr>
<td>Prelax-std w/o R2S</td>
<td>92.2</td>
</tr>
<tr>
<td>Prelax-std w/o Sim</td>
<td>91.7</td>
</tr>
<tr>
<td>Prelax-std w/o PL</td>
<td>91.5</td>
</tr>
<tr>
<td>Prelax-rot (R3S + Sim + RotPL)</td>
<td>92.4</td>
</tr>
<tr>
<td>Prelax-rot w/o R3S</td>
<td>91.1</td>
</tr>
<tr>
<td>Prelax-rot w/o Sim</td>
<td>79.8</td>
</tr>
<tr>
<td>Prelax-rot w/o RotPL</td>
<td>91.9</td>
</tr>
</tbody>
</table>
Takeaways

• Stronger augmentations like rotation are harmful for multi-view learning, but they contain useful semantics
• Residuals can be used to account for large semantic shift
• Residual relaxation generalizes multi-view learning to benefit from stronger augmentations
• Multi-view learning and self-supervised learning can be combined to encode richer semantics and yield better performance
Thanks!

Q & A

Find more stuff about this work at https://yifeiwang77.github.io/
Contact:
yifei_wang AT pku.edu.cn; yisen.wang AT pku.edu.cn