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

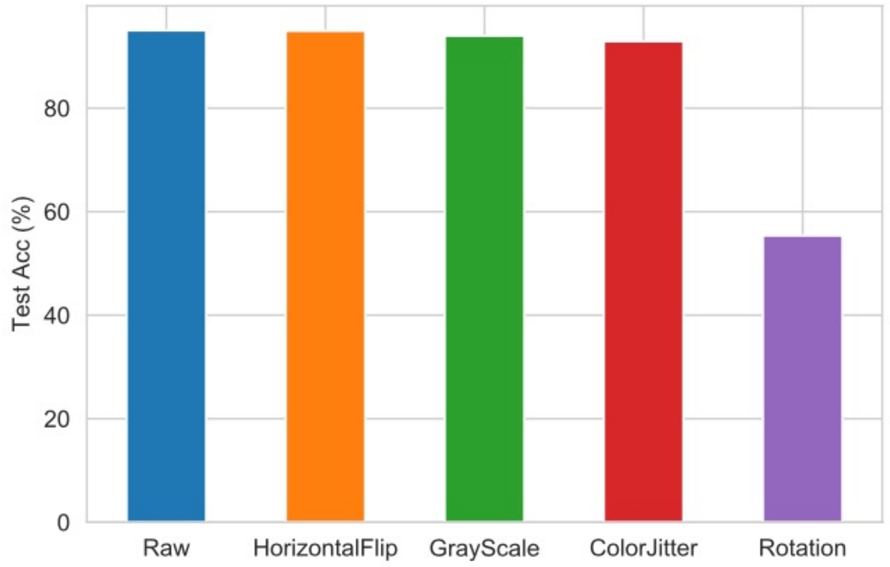
- Multi-view methods become dominant for unsupervised learning
  - SimCLR, MoCo, BYOL, SimSiam, etc
  - For each input x, we get two views, x1, x2 by random augmentation
  - Learn to align augmented views x1, x2 by minimizing representation distances

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

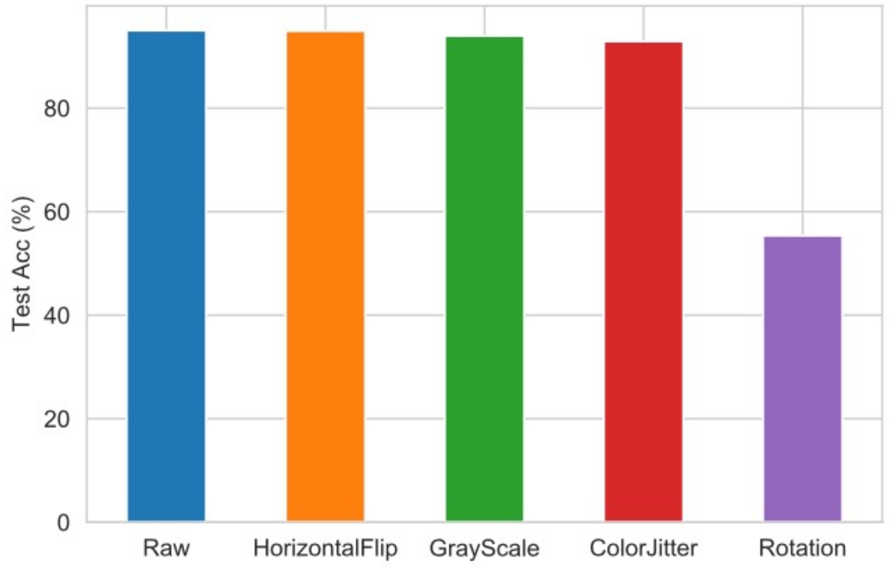
- Multi-view methods become dominant for unsupervised learning
  - SimCLR, MoCo, BYOL, SimSiam, etc
  - For each input x, we get two views, x1, x2 by random augmentation
  - Learn to align augmented views x1, x2 by minimizing representation distances
- Observation
  - Pretext (e.g. image augmentation) has a large effect on the final performance
  - Some augmentations, like rotation, are **too strong to be aligned exactly**



Method	Acc (%)
SimSiam [5]	91.8
SimSiam + margin loss	91.9
Rotation [9]	88.3
SimSiam + rotation aug.	87.9
SimSiam + Rotation loss	91.7

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
  - Some augmentations, like rotation, are **too strong to be aligned exactly**
  - However, rotation is known as **an effective signal** for Self-supervised Learning



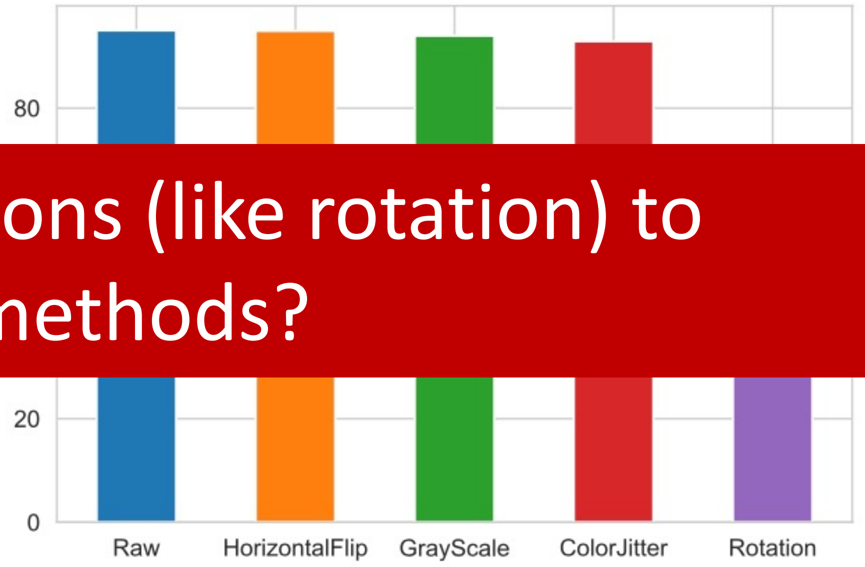
Method	Acc (%)
SimSiam [5]	91.8
SimSiam + margin loss	91.9
Rotation [9]	88.3
SimSiam + rotation aug.	87.9
SimSiam + Rotation loss	91.7

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 x1, x2 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



Method	Acc (%)
SimSiam [5]	91.8
SimSiam + margin loss	91.9
Rotation [9]	88.3
SimSiam + rotation aug.	87.9
SimSiam + Rotation loss	91.7

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**

Method	Acc (%)
SimSiam [5]	91.8
SimSiam + margin loss	91.9
Rotation [9]	88.3
SimSiam + rotation aug.	87.9
SimSiam + Rotation loss	91.7

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

$$\mathcal{L}_{\text{margin}}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) = \max\Big(\|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}')) - \mathcal{F}_{\phi}(\mathbf{x})\|_2^2 - \eta, 0\Big)$$

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

Method	Acc (%)
SimSiam [5]	91.8
SimSiam + margin loss	91.9
Rotation [9]	88.3
SimSiam + rotation aug.	87.9
SimSiam + Rotation loss	91.7

What does not work…

- Direct combination of pretext-invariant

Find an adaptive relaxation for each input!

- Use a margin loss to relax the alignment

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

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

Method	Acc (%)
SimSiam [5]	91.8
SimSiam + margin loss	91.9
Rotation [9]	88.3
SimSiam + rotation aug.	87.9
SimSiam + Rotation loss	91.7

Our Solution: Residual Relaxation

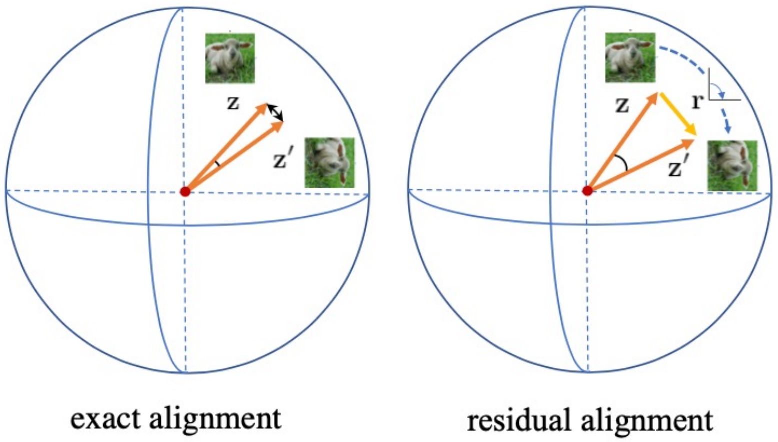
- Use residuals to account for the semantic shift brought by augmentations
- **Exact alignment** fails for strong augmentation

$$\mathbf{z}' \to\leftarrow \mathbf{z}$$

- **Identity alignment** always holds instead

$$\mathbf{z}' \to\leftarrow \mathbf{z} + \mathbf{r}$$

  - where $\mathbf{r} = \mathbf{z}' - \mathbf{z}$ encodes the semantic shift



(b) A toy example of residual relaxation.

Pretext-aware Residual Relaxation (Prelax)

- Baseline: similarity loss for x’=t(x)

$$\mathcal{L}_{\text{sim}}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) = \left\| \mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}')\right) - \mathcal{F}_{\phi}(\mathbf{x}) \right\|_2^2$$

  - F_θ online network, F_ϕ target network, G_θ online prediction network

Pretext-aware Residual Relaxation (Prelax)

- Baseline: similarity loss

$$\mathcal{L}_{\text{sim}}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) = \left\| \mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}')) - \mathcal{F}_{\phi}(\mathbf{x}) \right\|_2^2$$

  - F_θ online network, F_ϕ target network, G_θ online prediction network
- Prelax (ours)
  - Exact Alignment -> Identity Alignment

$$\mathbf{r} \triangleq \mathbf{z}'_{\boldsymbol{\theta}} - \mathbf{z}_{\boldsymbol{\theta}} = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}') - \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x})$$

$$\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}'_{\boldsymbol{\theta}}) \rightarrow\leftarrow \mathbf{z}_{\phi} \quad \Rightarrow \quad \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{z}'_{\boldsymbol{\theta}}) - \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{r}) \rightarrow\leftarrow \mathbf{z}_{\phi}$$

  - Residual Relaxed Similarity (R2S) loss (α is the interpolating coefficient)

$$\mathcal{L}_{\text{R2S}}^{\alpha}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}')\right) - \alpha\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{r}) - \mathcal{F}_{\phi}(\mathbf{x})\|_2^2.$$

Pretext-aware Residual Relaxation (Prelax)

- Prelax (ours)
  - Residual Relaxed Similarity loss (α is the interpolating coefficient)

$$\mathcal{L}^{\alpha}_{\mathrm{R2S}}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}')\right) - \alpha \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{r}) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x})\|_2^2.$$

  - 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**

$$\mathcal{L}_{\mathrm{PL}}\left(\mathbf{x}', \mathbf{x}, \mathbf{t}; \boldsymbol{\theta}\right) = \mathrm{CE}\left(\mathcal{H}^{d}_{\boldsymbol{\theta}}(\mathbf{r}), \mathbf{t}^{d}\right) + \|\mathcal{H}^{c}_{\boldsymbol{\theta}}(\mathbf{r}) - \mathbf{t}^{c}\|_2^2$$

A **non-conflicting** combination of multi-view methods and predictive methods

Pretext-aware Residual Relaxation (Prelax)

- Prelax (ours)
  - Residual Relaxed Similarity loss (α is the interpolating coefficient)

$$\mathcal{L}_{\mathrm{R2S}}^{\alpha}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}')\right) - \alpha\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{r}) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x})\|_2^2.$$

  - Predictive Learning (PL) Loss

$$\mathcal{L}_{\mathrm{PL}}\left(\mathbf{x}', \mathbf{x}, \mathbf{t}; \boldsymbol{\theta}\right) = \mathrm{CE}\left(\mathcal{H}_{\boldsymbol{\theta}}^{d}(\mathbf{r}), \mathbf{t}^{d}\right) + \|\mathcal{H}_{\boldsymbol{\theta}}^{c}(\mathbf{r}) - \mathbf{t}^{c}\|_2^2$$

  - Constraint on the Similarity
    - the residual is unbounded, and the distance between views could be very large
    - enforce small distance by adding a constraint

$$\mathcal{L}_{\mathrm{sim}} = \left\|\mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)\right) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x})\right\|_2^2 \leq \varepsilon$$

Pretext-aware Residual Relaxation (Prelax)

- Prelax (ours)
  - Residual Relaxed Similarity loss (α is the interpolating coefficient)

$$\mathcal{L}_{\mathrm{R2S}}^{\alpha}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}')\right) - \alpha\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{r}) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x})\|_2^2.$$

  - Predictive Learning (PL) Loss

$$\mathcal{L}_{\mathrm{PL}}\left(\mathbf{x}', \mathbf{x}, \mathbf{t}; \boldsymbol{\theta}\right) = \mathrm{CE}\left(\mathcal{H}_{\boldsymbol{\theta}}^{d}(\mathbf{r}), \mathbf{t}^{d}\right) + \|\mathcal{H}_{\boldsymbol{\theta}}^{c}(\mathbf{r}) - \mathbf{t}^{c}\|_2^2$$

  - Constraint on the Similarity

$$\mathcal{L}_{\mathrm{sim}} = \left\|\mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)\right) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x})\right\|_2^2 \leq \varepsilon$$

Pretext-aware Residual Relaxation (Prelax)

- Prelax (ours)
  - Residual Relaxed Similarity loss (α is the interpolating coefficient)

$$\mathcal{L}_{\mathrm{R2S}}^{\alpha}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}')\right) - \alpha \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{r}) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x})\|_2^2.$$

  - Predictive Learning (PL) Loss

$$\mathcal{L}_{\mathrm{PL}}\left(\mathbf{x}', \mathbf{x}, \mathbf{t}; \boldsymbol{\theta}\right) = \mathrm{CE}\left(\mathcal{H}_{\boldsymbol{\theta}}^{d}(\mathbf{r}), \mathbf{t}^{d}\right) + \|\mathcal{H}_{\boldsymbol{\theta}}^{c}(\mathbf{r}) - \mathbf{t}^{c}\|_2^2$$

  - Constraint on the Similarity

$$\mathcal{L}_{\mathrm{sim}} = \left\|\mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}\left(\mathbf{x}'\right)\right) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x})\right\|_2^2 \leq \varepsilon$$

  - Combined

$$\min_{\boldsymbol{\theta}} \ \mathcal{L}_{\mathrm{R2S}}^{\alpha}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) + \gamma \mathcal{L}_{\mathrm{PL}}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}), \quad s.t. \quad \|\mathcal{G}_{\boldsymbol{\theta}}\left(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}')\right) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x})\|_2^2 \leq \varepsilon.$$

Penalized →

$$\mathcal{L}_{\mathrm{R2S}}^{\alpha}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) + \gamma \mathcal{L}_{\mathrm{PL}}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}) + \beta \mathcal{L}_{\mathrm{sim}}(\mathbf{x}', \mathbf{x}; \boldsymbol{\theta}),$$

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
    - $\mathbf{S_v}$: multi-view learning, $\mathbf{S_a}$: 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
    - **X** input, **T** downstream task, **S** self-supervised signal, **Z** representation
    - $\mathbf{S_v}$: multi-view learning, $\mathbf{S_a}$: predictive learning
    - Goal: maximize mutual information I(**Z;T**) with downstream task
  - Prelax extracts more task-relevant information than multi-view ($\mathbf{Z_{mv}}$) and predictive ($\mathrm{Z}_{\mathbf{PL}}$) 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* $I(\mathbf{Z};\mathbf{T})$*:*

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

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
    - $\mathbf{S_v}$: multi-view learning, $\mathbf{S_a}$: predictive learning
    - Goal: maximize mutual information I(**Z;T**) with downstream task
  - Prelax extracts more task-relevant information than multi-view ($\mathbf{Z_{mv}}$) and predictive ($\mathrm{Z}_{\mathbf{PL}}$) methods
  - As a result, Prelax has a tighter upper bound on the downstream Bayes error

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

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

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

$$u^e_{\mathrm{sup}} \leq u^e_{\mathrm{Prelax}} \leq \min(u^e_{\mathrm{mv}}, u^e_{\mathrm{PL}}) \leq \max(u^e_{\mathrm{mv}}, u^e_{\mathrm{PL}}). \tag{12}$$

Practical Implementations of Prelax

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

$$\mathcal{L}_{\text{Simsiam}}(\mathbf{x};\boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_1)) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x}_2)\|_2^2 + \|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_2)) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x}_1)\|_2^2$$

Practical Implementations of Prelax

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

$$\mathcal{L}_{\text{Simsiam}}(\mathbf{x};\boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_1)) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x}_2)\|_2^2 + \|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_2)) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x}_1)\|_2^2$$

- Prelax-std: generalize baselines with existing augmentations
  - Residual

$$\mathbf{r}_{12} = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_1) - \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_2)$$

  - Prelax-std objective

$$\mathcal{L}_{\text{Prelax-std}}(\mathbf{x};\boldsymbol{\theta}) = \mathcal{L}^{\alpha}_{\text{R2S}}(\mathbf{x}_1,\mathbf{x}_2;\boldsymbol{\theta}) + \gamma\mathcal{L}_{\text{PL}}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{t}_1;\boldsymbol{\theta}) + \beta\mathcal{L}_{\text{sim}}(\mathbf{x}_2,\mathbf{x}_1;\boldsymbol{\theta}).$$

Practical Implementations of Prelax

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

$$\mathcal{L}_{\text{Simsiam}}(\mathbf{x};\boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_1)) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x}_2)\|_2^2 + \|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_2)) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x}_1)\|_2^2$$

- Prelax-std: generalize baselines under existing augmentations
- Prelax-rot: incorporating stronger augmentation (rotation)
  - a third view $\mathbf{x}_3$ as a randomly rotated $\mathbf{x}_1$, residual (for rotation) $\mathbf{r}_{31} = \mathbf{z}_3 - \mathbf{z}_1$
  - Rotation Residual Relaxation Similarity (R3S) loss

$$\mathcal{L}^{\alpha}_{\text{R3S}}(\mathbf{x}_{1:3};\boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_3)) - \alpha\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{r}_{31}) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x}_2)\|_2^2.$$

  - Combined

$$\mathcal{L}_{\text{Prelax-rot}}(\mathbf{x};\boldsymbol{\theta}) = \mathcal{L}^{\alpha}_{\text{R3S}}(\mathbf{x}_{1:3};\boldsymbol{\theta}) + \gamma\mathcal{L}^{\text{rot}}_{\text{PL}}(\mathbf{x}_1, \mathbf{x}_3, a;\boldsymbol{\theta}) + \beta\mathcal{L}_{\text{sim}}(\mathbf{x}_2, \mathbf{x}_1;\boldsymbol{\theta}).$$

Practical Implementations of Prelax

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

$$\mathcal{L}_{\text{Simsiam}}(\mathbf{x};\boldsymbol{\theta}) = \|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_1)) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x}_2)\|_2^2 + \|\mathcal{G}_{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}}(\mathbf{x}_2)) - \mathcal{F}_{\boldsymbol{\phi}}(\mathbf{x}_1)\|_2^2$$

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

$$\mathcal{L}_{\text{Prelax-all}}(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{2}\left(\mathcal{L}_{\text{R2S}}^{\alpha_1}(\mathbf{x}_1,\mathbf{x}_2;\boldsymbol{\theta}) + \mathcal{L}_{\text{R3S}}^{\alpha_2}(\mathbf{x}_{1:3};\boldsymbol{\theta})\right) + \frac{\gamma_1}{2}\mathcal{L}_{\text{PL}}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{t}_1;\boldsymbol{\theta}) + \frac{\gamma_2}{2}\mathcal{L}_{\text{PL}}^{\text{rot}}(\mathbf{x}_1,\mathbf{x}_3,a;\boldsymbol{\theta}) + \beta\mathcal{L}_{\text{sim}}(\mathbf{x}_2,\mathbf{x}_1;\boldsymbol{\theta}),$$

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.

(a) CIFAR-10.

Method	Acc. (%)
Supervised [12] (re-produced)	95.0
Rotation [9] (re-produced)	88.3
BYOL [10] (re-produced)	91.1
SimCLR [2]	91.1
SimSiam [5]	91.8
SimSiam + Prelax	**93.4**

(b) ImageNette.

Method	Acc. (%)
Supervised	91.0
Supervised + TTA	92.2
BYOL [10] (ResNet-18)	91.9
BYOL [10] (ResNet-50)	92.3
BYOL + Prelax (ResNet-18)	**92.6**

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

(a) In-domain linear evaluation.

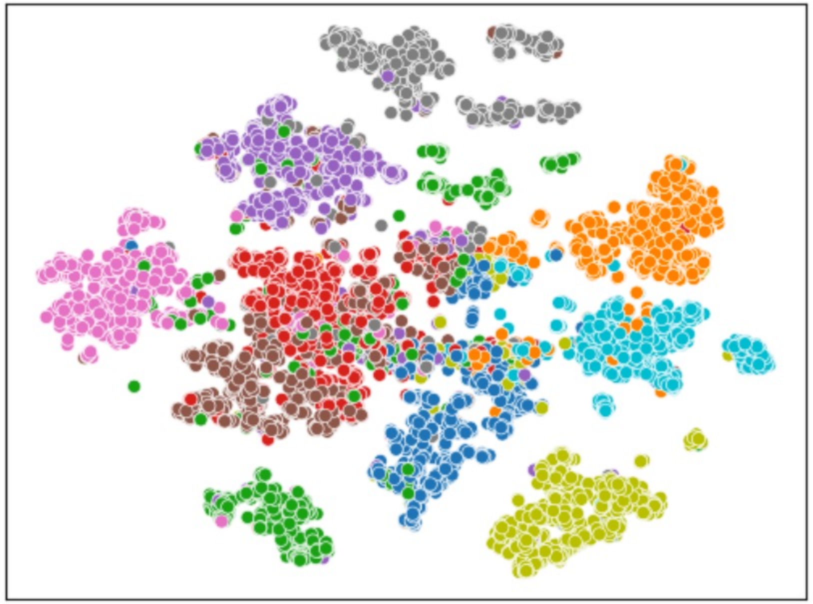
Method	CIFAR-10	CIFAR-100	Tiny-ImageNet-200
SimSiam [5]	91.8	66.9	47.7
SimSiam + Prelax-std	92.5	67.5	47.9
SimSiam + Prelax-rot	92.4	67.3	47.1
SimSiam + Prelax-all	**93.4**	**70.0**	**49.2**

(b) Out-of-domain linear evaluation.

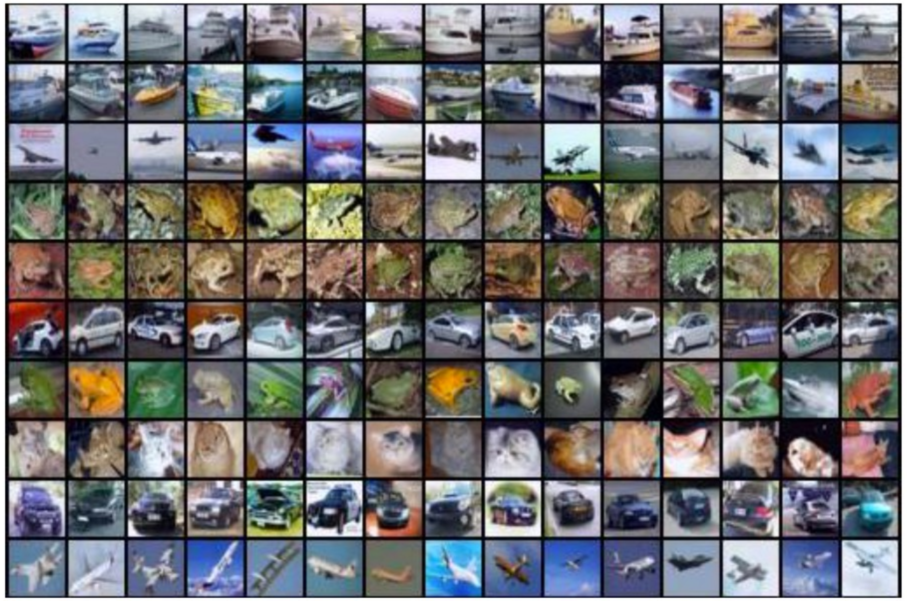
Method	C100 $\rightarrow$ C10	Tiny200 $\rightarrow$ C10	Tiny200 $\rightarrow$ C100
SimSiam [5]	44.1	43.9	21.8
SimSiam + Prelax-std	**45.0**	**45.1**	21.8
SimSiam + Prelax-rot	**45.0**	**45.1**	22.0
SimSiam + Prelax-all	44.9	44.6	**22.1**

Experiments

- Empirical understandings



(a) Representation visualization.



(b) Nearest image retrieval.

Experiments

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

(a) Comparison against alternative options.

Method	Acc. (%)
SimSiam [5]	91.8
SimSiam + margin loss	91.9
Rotation [9]	88.3
SimSiam + rotation aug.	87.9
SimSiam + Rotation loss	91.7
SimSiam + Prelax (ours)	**93.4**

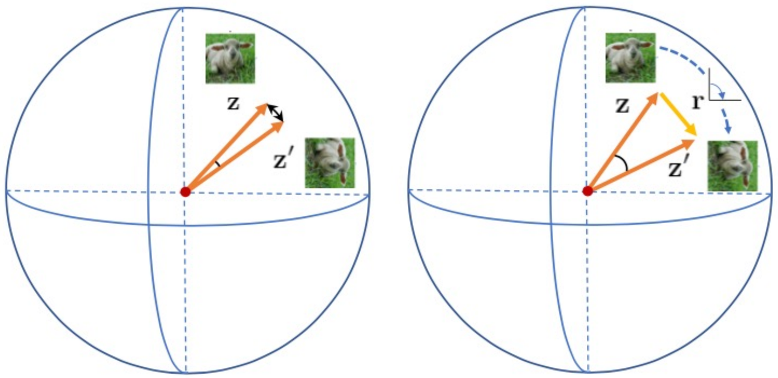
(b) Ablation study.

Method	Acc. (%)
Prelax-std (R2S + Sim + PL)	**92.5**
Prelax-std w/o R2S	92.2
Prelax-std w/o Sim	91.7
Prelax-std w/o PL	91.5
Prelax-rot (R3S + Sim + RotPL)	**92.4**
Prelax-rot w/o R3S	91.1
Prelax-rot w/o Sim	79.8
Prelax-rot w/o RotPL	91.9

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!



exact alignment residual alignment

(b) A toy example of residual relaxation.

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