Unbiased VICReg

S. Bezyazichniy,
E. Serov,
D. Ivanov,
P. Lisov,
N. Groza

MIPT

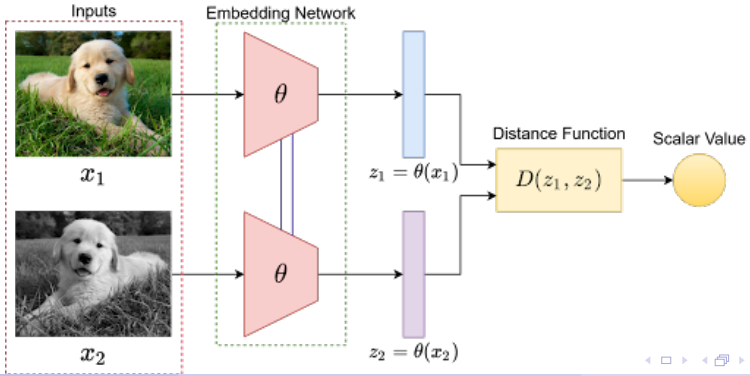
October 2, 2024

Contents

1. Basic SSL Idea
2. SimCLR
3. VICReg
4. SimCLR vs VICReg
5. Unbiased VICReg

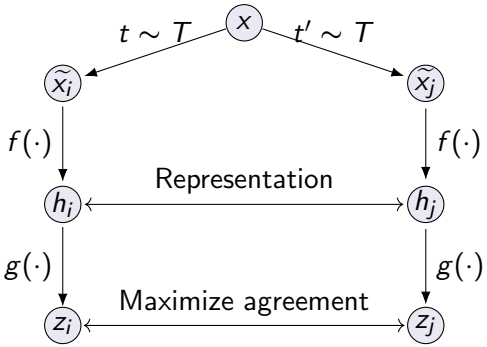
Preventing Collapse in Self-Supervised Learning

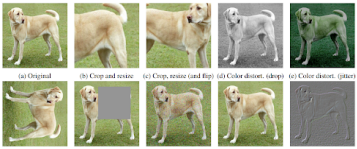
- **Self-Supervised Learning (SSL)** has made significant progress, allowing models to learn without labeled data by maximizing agreement between embeddings of different views of the same image or other object.
- **Main Challenge:** In joint embedding architectures models often *collapse* producing constant, non-informative embeddings.



Simple Framework for Contrastive Learning of Visual Representations

1. **Augmentation:** generate two correlated views from given data using random cropping, resizing, horizontal flipping, colour distortion, grey scaling and/or Gaussian blur.
2. **Encoder:** extract representation vectors from augmented data.
3. **Projection head:** map representations to the embedding space where contrastive loss is applied. This small neural network with one hidden layer improves the quality of learned representations.





Contrastive Loss

SimCLR aims to learn representations by *maximizing* agreement between differently augmented views of the same image (positive pairs) and *minimizing* it for different images (negative pairs).

$$\text{sim}(u, v) = \frac{u^\top v}{\|u\| \cdot \|v\|}$$

τ — hyperparameter

$$\mathfrak{l}_{i,j} = -\log \frac{\exp \frac{\text{sim}(z_i, z_j)}{\tau}}{\sum_{k=1}^{2N} \mathbb{I}_{(k \neq i)} \exp \frac{\text{sim}(z_i, z_k)}{\tau}}$$

Batch Size and Learning Rate Scaling

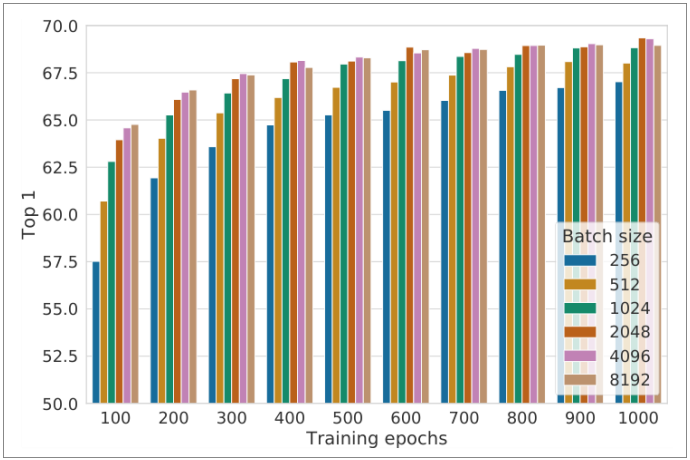
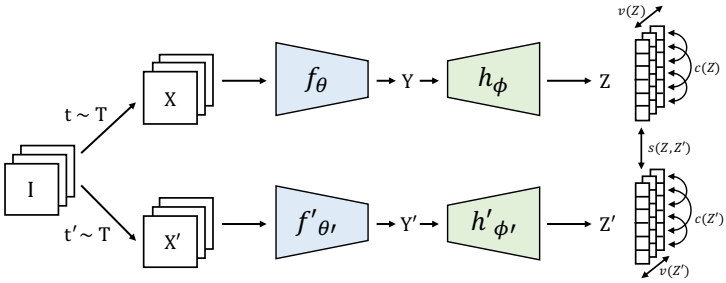


Figure: Linear evaluation models (ResNet-50) trained with different batch sizes and epochs. Each bar is a single run from scratch

VICReg



v	: maintain variance
c	: bring covariance to zero
s	: minimize distance
T	: distribution of transformations
t,t'	: random transformations
$f_\theta, f'_{\theta'}$	: encoders
$h_\phi, h'_{\phi'}$	: expanders
I	: batch of images
X, X'	: batches of views
Y, Y'	: batches of representations
Z, Z'	: batches of embeddings

VIC

Consider $Z = (z_1, \ldots, z_n)$ and $Z' = (z'_1, \ldots, z'_n)$ be the two batches composed of n vectors of dimension d.

- **Invariance:**

$$s(Z, Z') = \frac{1}{n} \sum_{i=1}^{n} \|z_i - z'_i\|_2^2.$$

- **Variance:**

$$v(Z) = \frac{1}{d} \sum_{i=1}^{d} \max(0, \gamma - \sqrt{\mathbb{V} z^j + \varepsilon}),$$

where γ — minimal standard deviation, ε — small scalar preventing numerical instabilities.

- **Covariance:**

$$c(Z) = \frac{1}{d} \sum_{i=1}^{d} |C(Z)|_{(i,j)}^2, \text{ where } C(Z) = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - z)(z_i - z)^\intercal.$$

Reg

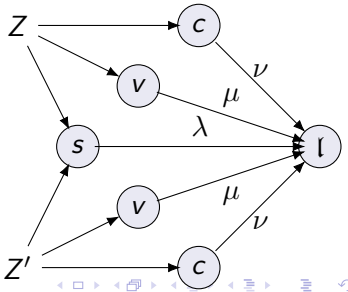
The loss function over batch is

$\mathfrak{l}(Z, Z') = \lambda s(Z, Z') + \mu(v(Z) + v(Z')) + \nu(c(Z) + c(Z'))$, where $\lambda = \mu > \nu$ in most cases.

The overall objective function on all images over dataset $\mathfrak{D}$ is

$$\mathfrak{L} = \sum_{I \in \mathfrak{D}} \sum_{t, t' \sim T} \mathfrak{l}(g(f(t(I))), g(f(t'(I))).$$

Model's performance is less dependent on large batch sizes because it does not rely on negative samples. Instead, it focuses on regularizing the variance, invariance and covariance of the learned embeddings, making it more flexible and effective even with smaller batch sizes.



SimCLR vs VICReg

Method	Batch size				
	256	512	1024	2048	4092
SimCLR	57.5	60.7	62.8	64.0	64.6
VICReg	67.9	68.2	68.3	68.6	67.8

Figure: Impact of batch size to Top-1 accuracy after 100 epochs on ImageNet.

Unbiased VICReg

Assumption. By making the VICReg gradient unbiased, we can eliminate the dependency on batch size, leading to more stable training with smaller batches. To accomplish this, we need to identify an unbiased estimator for the following loss function:

$$\mathfrak{L} = \lambda \mathop{\mathbb{E}}_{X\sim\mathfrak{D},t,t'\sim T} \|f_\theta(T(X)) - f_\theta(T'(X))'\|_2^2 + \frac{\mu}{d}\sum_{i=1}^{d} \max(0, \gamma - \sqrt{\mathbb{V}Z_i}) + \nu\|\mathrm{cov}(Z) - I\|_F^2 \to \min,$$

where $Z = f_\theta(t(X))$.
But how to find unbiased estimator for the second summand?

Any questions???

Thanks for your attention! :D