The Canonical Correlation Analysis Family: VICReg/BarlowTwins/SWAV/W-MSE

Xunyi Jiang

23 Summer Study - Week7

xunyijiang001@gmail.com

Aug 3, 2023

Overview

Overview

SSL(Self-Supervise Learning) Methods Summary

- Contrastive learning: MoCo/SimCLR
  - Many authors use the InfoNCE loss in which the repulsive force is larger for contrastive samples that are closer to the reference.
- Distillation methods: BYOL/SimSiam/OBoW:
  - These methods have shown that collapse can be avoided by using architectural tricks inspired by knowledge distillation.
- Clustering methods: SeLa/SwAV
  - Instead of viewing each sample as its own class, clustering-based methods group them into clusters based on some similarity measure
- Information maximization methods: W-MSE/Barlow Twins/VICReg
  - A principle to prevent collapse is to maximize the information content of the embeddings.

Notations

Symbol	Definition
Image	i
Transformation	$t, t' \sim \mathcal{T}$
Views/Augments	$x = t(i), x' = t'(i)$
Representation NN	f_θ
Representations	$y = f_\theta(x)$
Prediction heads	h_ϕ (MLP & SoftMax)
Embeddings	$z = h_\phi(y), z \in \mathbb{R}^K$
Hole NN	$h_\phi \circ f_\theta$
Batch size	B

Table: Notations

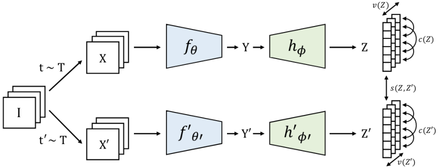


Figure: VICReg

Information Maximization Methods: W-MSE/Barlow Twins/VICReg

W-MSE

W-MSE: Whitening for Self-supervised Learning

- Object: Learn f_θ that can extract good features.
- Input: Batch of Images X
- Output: $v = Whitening(z)$ to calculate loss
- IDEA: Use whihtening to "scatter" the embeddings, avoiding degenerate solutions.

Figure: Framework of W-MSE

W-MSE: Whitening

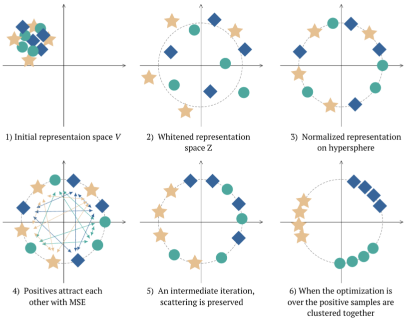


Figure: Whiten Visualization

Whitening

$$Whitening() : \mathbb{R}^K \to \mathbb{R}^K$$
$$Whitening(z) = W_z(z - \mu_Z)$$
$$\text{where } W_Z^T W_Z = \Sigma_Z^{-1} =$$
$$\left(\frac{1}{B-1} \sum_b (z_b - \mu_Z)(z_b - \mu_Z)^T\right)^{-1},$$
$$\mu_Z = \frac{1}{B} \sum_b z_b$$

Denote: $v = Whitening(z)$

Suppose we have d views, N samples in the dataset. t_i, t_j means different augmentaions.

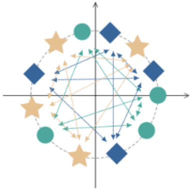


Figure: Whitening MSE Calculation

$$L_{W-MSE}(\mathcal{D}) = \frac{2}{Nd(d-1)} \sum_b \sum_{ij} dist(v_{b,t_i}, v_{b,t_j})$$

$$\text{where } dist(v_{b,t_i}, v_{b,t_j}) = \left\| \frac{v_{b,t_i}}{\|v_{b,t_i}\|_2} - \frac{v_{b,t_j}}{\|v_{b,t_j}\|_2} \right\|^2$$

$$= 2 - 2\frac{\langle v_{b,t_i}, v_{b,t_j} \rangle}{\|v_{b,t_i}\|_2 \|v_{b,t_j}\|_2}$$

W-MSE: Experiment & Conclusion

- Experiments:
  1. Freezing the network encoder (Res-Net50)
  2. Add fully connected layer followed by softmax./ KNN
  3. Hyperparameters: learning rate from $10^{-2} to 10^{-6}$, epochs $= 500$
- Advantages:
  - No Negatives
  - No asymmetric networks
  - Conceptually Simple

Barlow Twins

Barlow Twin: Overview

- Object: Learn efficient representations.
- Input: Image
- Output: Embeddings z
- IDEA: Minimize the redundancy between the components of embeddings. CCA!

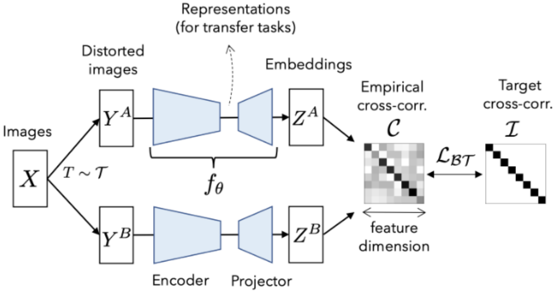


Figure: Barlow Twins Frame

Barlow Twins: Loss

Barlow Twins Loss

$$\mathcal{L}_{\mathcal{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2$$

where λ is a positive constant trading off the importance of the first and the second terms of the loss.

$$\mathcal{C}_{ij} = \frac{\sum_b z_{b,i}^{t_1} z_{b,j}^{t_2}}{\sqrt{\sum_b (z_{b,i}^{t_1})^2} \sqrt{\sum_b (z_{b,j}^{t_2})^2}}$$

$\mathcal{C}$ is the cross-correlation matrix of z^{t_1}, z^{t_2}.

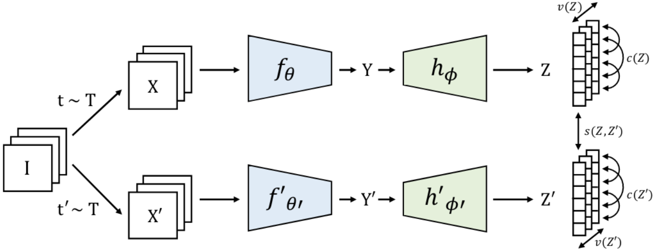
Barlow Twins: Experiment & Conclusion

- Experiments (Semi-supervised):
  - Fine-tune a Res-Net50 pretrained with BT on a subset of Image-Net
- Advantages:
  - Not require large batches
  - No asymmetric network
  - BARLOW TWINS outperforms previous methods on ImageNet for semi-supervised classification in the low-data regime.

VICReg: Variance-Invariance-Covariance Regularization for SSL

VICReg: Variance-Invariance-Covariance Regularization for SSL

- Object: Learn efficient representaions
- Input: Images I
- Output: Embeddings Z
- IDEA:
  Different pictures have **V**ariance,
  same picture with different views maintain **I**nvariance,
  different features need to de-**C**ovariance



VICReg: Loss

In order to avoid collapse, different pictures will produce different embeddings.

- Variance Regularization (Hinge Loss):

Variance Regularization (Hinge Loss)

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} max(0, \gamma - S(Z_{j\cdot}, \epsilon))$$

where S is the regularized standard deviation defined by:

$$S(x, \epsilon) = \sqrt{Var(x) + \epsilon}$$

where ϵ is a small scalar preventing numerical instabilities.

VICReg: Loss

Inspired by Barlow Twins, the difference is the covariance matrix rather than cross-correlation matrix.

- Covariance Regularization: The covariance matrix is:

$$C(Z) = \frac{1}{B-1}\sum_{b}(z_i - \bar{z})(z_i - \bar{z})^T$$

where $\bar{z} = \frac{1}{B}\sum_b z_i$
The Covariance Regularization is:

Covariance Regularization

$$c(Z) = \frac{1}{d}\sum_{i \neq j}[C(Z)]^2_{i,j}$$

VICReg: Loss

The same image with different transformation will result in the similar vectors.

Invariant Regularization

$$s(Z, Z') = \frac{1}{B} \sum_b \|z_b - z'_b\|_2^2$$

Then the overall loss is:

Overall Loss

$$\ell(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')]$$

$$\mathcal{L} = \sum_{I \in D} \sum_{t, t' \sim \mathcal{T}} \ell(Z, Z')$$

where λ, μ, ν are the hyper-parameters.

Conclusion

Advantages:

- No negatives
- No batch normalization, feature-wise normalization
- No output quantization
- No stop gradiant
- No memory bank

Clustering Method: SeLa/SwAV

SeLa: Self-labelling

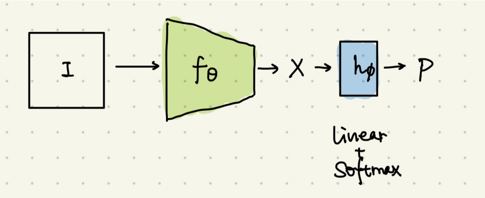


Figure: Sela framework

Suppose $x \in \mathbb{R}^D$, $h : \mathbb{R}^D \to \mathbb{R}^K$, where K is the cluster number.

$$p(y = \cdot|x_i) = \text{softmax}(h_\phi(x_i))$$

If we have the label, then we can minimize the cross-entropy and solve the problem.
IDEA: Learn pseudo labels, calling $q(y|x_i)$, which can minimize:

$$E(p, q) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{y=1}^{K} q(y|x_i) \log p(y|x_i)$$

subject to

$$\forall y : q(y|x_i) \in \{0, 1\},$$

$$\sum_{i=1}^{N} q(y|x_i) = \frac{N}{K} \text{ and } \sum_{y=1}^{K} q(y|x_i) = 1$$

Formulate this Problem

We want to minimize:

$$E(p, q) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{K} q(y|x_i) \log p(y|x_i)$$

subject to

$$\forall y : q(y|x_i) \in \{0, 1\},$$

$$\sum_{i=1}^{N} q(y|x_i) = \frac{N}{K} \text{ and } \sum_{y=1}^{K} q(y|x_i) = 1$$

Alternative Optimization

- Step1: **representation learning:** Fix q, update f_θ we can considering this problem as a supervised problem, using cross-entropy:

$$\min_p E(p, q)$$

- Step2: **self-labelling:** Fix f_θ, we can get fixed P, then using the aforementioned method (Sinkhorn-Knopp algorithm) to update q.

$$\min_{q \in U} E(p, q)$$

where U is the constraints aforementioned.

SeLa Summarize

- No negative, but the constrains implies negatives
- Cluster the data into K classes
- Use Sinkhorn-Knopp algorithm to solve Q
- Not use transformation information

SwAV: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

SwAV Framework

Similar to SeLa, but it uses the information of augmentation.

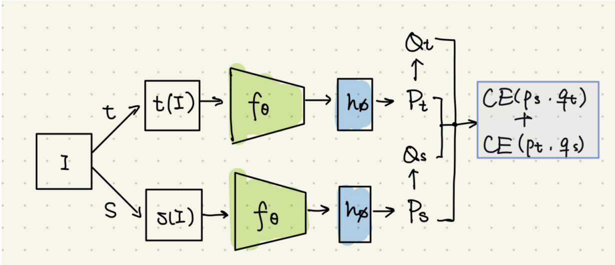


Figure: SwAV Framework

SwAV Loss

$$-\frac{1}{N}\sum_{n=1}^{N}\sum_{s,t\sim\mathcal{T}}[\sum_{k} q_{ns}^{(k)} \log p_{nt}^{(k)} + \sum_{k} q_{nt}^{(k)} \log p_{ns}^{(k)}]$$

SwAV Summary

- **IDEA:** Same images under different transformations should have similar features.
- Using augmentation
- Introduce clustering prior thought
- Good Performance
- Introduce multi-crop, which uses a mix of views with different resolutions in place of two full-resolution views, without increasing the memory or compute requirements.

Summary of all these methods

- No Negatives, but implies negatives
    - W-MSE: whitening process can scatter the points.
    - VICReg: Add variance threshold to different sample in the same batch.
    - SeLa/SwAV: add constrains to cluster size.
- Use covariance to maximize information. (Different feature should be decorrelation-CCA).
- Cluster the features and use other optimization methods.

Thanks!