Barlow Twins: Self-supervised Learning via Redundancy Reduction (J. Zbontar et al., ICML'21)

Zhiming Xu
zx2rw@virginia.edu

UNIVERSITY of VIRGINIA | ENGINEERING

School of Engineering and Applied Science
University of Virginia

March 7, 2022

Table of Contents

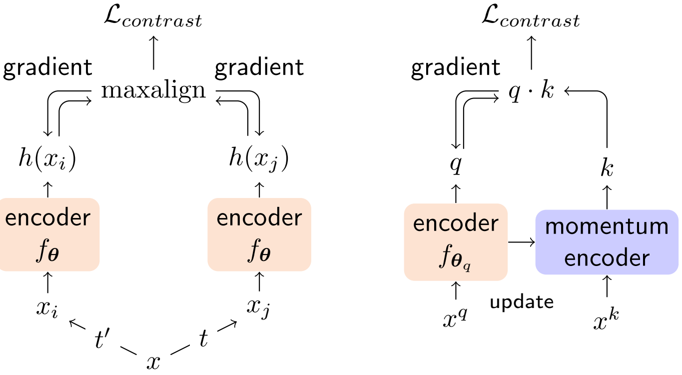
Objective of Self-supervised Learning

Self-supervised learning (SSL) has achieved large success recently. A mainstream approach is to learn representations based on several perturbed versions of the input data.

This approach could lead to trivial solutions such as a constant representation for all data, which is avoided by implementation details of previous models.

Methods to Prevent Trivial Solutions

- *SimCLR* [2]. Leverage positive and negative samples to ensure the learned representations will not collapse to the same.
- *MoCo* [3]. Leverage asymmetric update to update momentum encoder separately from the main encoder.



(a) Positives and negatives (b) Asymmetric update

Table of Contents

Motivation

H. Barlow's *redundancy-reduction* principle: the goal of sensory processing is recode highly redundant sensory inputs into a factorial code (a code with statistically independent components).

Based on this hypothesis, the *Barlow Twins* objective function is proposed to ensure the cross-relation matrix computed from the twin embeddings as close to identity matrix as possible.

Formulating the Objective Function

Barlow Twins works on a pair of representations of perturbed data. It computes the correlation between the representations of two distorted data Y^A and Y^B.

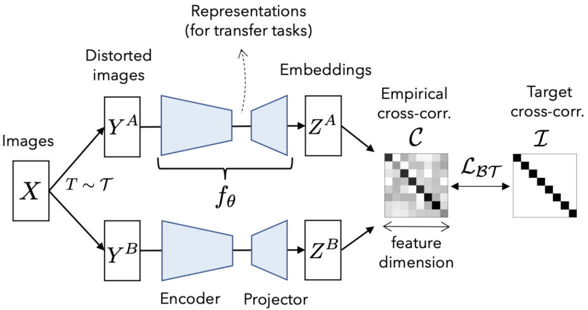


Figure 1: Barlow Twins objective function.

Formulation the Objective Function (cont.)

Assuming that the encoded representations Z^A and Z^B are centered along the batch dimension, the Barlow Twins objective function is defined as

$$\mathcal{L}_{BT} = \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance}} + \underbrace{\lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction}} \quad . \tag{1}$$

λ is the weighting factor, and $\mathcal{C}$ is the correlation matrix.

$$\mathcal{C}_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b \left(z_{b,i}^A\right)^2} \sqrt{\sum_b \left(z_{b,j}^B\right)^2}}. \tag{2}$$

b indexes the batch dimension while i, j index the feature dimension. $\mathcal{C}_{ij}$ ranges from -1, perfect anti-correlation to 1, perfect correlation.

Objectives of Barlow Twins

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

- The *invariance term* tries to ensure the diagonal elements of the correlation matrix equals 1. This makes the representations invariant to the distortion applied.
- The *redundancy reducion term* tries to ensure the off-diagonal elements of the correlation matrix equals 0. This de-correlates the different vector components of the representation.

The objective of information bottleneck (IB) aims to find a representation that converses as much information about the sample and as little information about the distortion as possible.

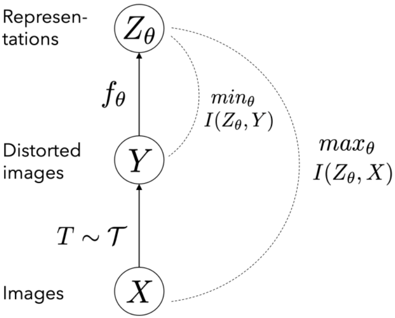


Figure 2: The objective of IB in representation learning.

Compare with InfoNCE

The InfoNCE loss, which maximizes the alignment with positives and minimizing that with negatives, can be reformulated as

$$\mathcal{L}_{InfoNCE} = \underbrace{-\sum_b \frac{\left(z_a^A \cdot z_b^B\right)_i}{\tau \left\|z_b^A\right\|_2 \left\|z_b^B\right\|_2}}_{\text{similarity}} + \underbrace{\sum_b \log\left(\sum_{b' \neq b} \exp\left(\frac{(z_b^A \cdot z_{b'}^B)_i}{\tau \left\|z_b^A\right\|_2 \left\|z_{b'}^B\right\|_2}\right)\right)}_{\text{contrastive}} \tag{3}$$

The two terms here serve similar purpose to those in Barlow Twins. However, InfoNCE increases the variability of the representations by maximizing the pairwise distance between all pairs of samples.

Compare with Asymmetric Twins

Asymmetric updates can use a simple cosine similarity between twin representations as an objective function w/o contrastive term

$$\mathcal{L}_{cosine} = -\sum_b \frac{\left(z_b^A \cdot z_b^B\right)_i}{\left\|z_b^A\right\|_2 \left\|z_b^B\right\|_2}. \tag{4}$$

These models avoid trivial solutions by introducing certain asymmetry in the twin neural networks, such as addtional predictor network and stop-gradient in BYOL [4].

Compare with IMAX

Earlier works on SSL proposed a twin loss function defined as

$$\mathcal{L}_{IMAX} = \log |\mathcal{C}_{Z^A - Z^B}| - \log |\mathcal{C}_{Z^A + Z^B}| . \tag{5}$$

$|\cdot|$ denotes the determinant of a matrix, and $\mathcal{C}_{Z^A \pm Z^B}$ is the covariance matrix of $Z^A \pm Z^B$. It is similar to Barlow Twins in that there is one similarity term and one de-correlation term.

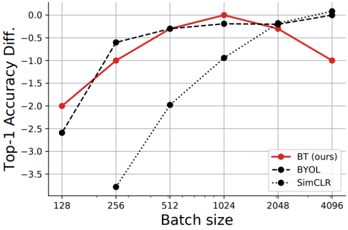
Table of Contents

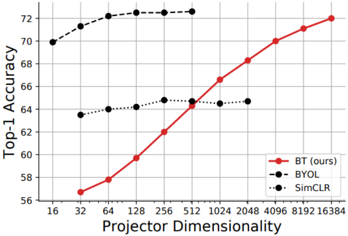
Effects of Components in Barlow Twins

- Baseline: $\mathcal{L}_{BT} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2$.
- Cross-entropy with temperature: $\mathcal{L}_{CE} = -\log \sum_i \exp\left(\mathcal{C}_{ii}/\tau\right) + \lambda \log \sum_i \sum_{j \neq i} \exp\left(\max\left(\mathcal{C}_{ij}, 0\right)/\tau\right)$.

Loss function	Top-1	Top-5
Baseline	71.4	90.2
Only invariance term (on-diag term)	57.3	80.5
Only red. red. term (off-diag term)	0.1	0.5
Normalization along feature dim.	69.8	88.8
No BN in MLP	71.2	89.7
No BN in MLP + no Normalization	53.4	76.7
Cross-entropy with temp.	63.3	85.7

- Barlow Twins does not require a very large batch size to achieve the best performace.
- A wider projector representation will benefit Barlow Twins, but has little influence on other models.





Hyperparameter Sensitivity

Tuning the hyperparameter λ weighting the invariance and de-correlation term does not have a strong effect on the model performance.

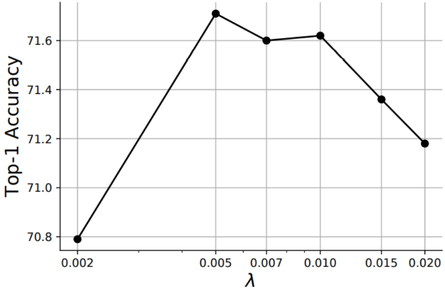


Table of Contents

Summary

- Propose Barlow Twins objective function for self-supervised learning that maximizes similarity between twin representations and minimizes redundancy among their components.
- Draw connections and comparisons with IB and previous SSL models via theoretical analyses.
- Achieve good results on benchmark datasets and alleviate certain disadvantages in previous works, e.g., large batch size.

Thank You for Your Attention

Q & A

References

[1] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, PMLR, 2021, pp. 12 310–12 320.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[3] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[4] J.-B. Grill, F. Strub, F. Altché, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.