Paper Review: Barlow Twins

Beomsu Kim*

July 25, 2021

Paper Information.

- Jure Zbontar et. al. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. arXiv preprint arXiv:2103.03230, 2021.

1 Introduction

- In this paper, we propose a new method, *Barlow Twins*, which applies redundancy reduction — a principle first proposed in neuroscience — to self-supervised learning.
- In his influential article *Possible Principles Underlying the Transformation of Sensory Messages*, neuroscientist H. Barlow hypothesized that the goal of sensory processing is to recode highly redundant sensory inputs into a factorial code (a code with statistically independent components).
- Based on this principle, we propose an objective function which tries to make the cross-correlation matrix computed from twin embeddings as closed to the identity matrix as possible.

2 Method

- Like other methods for SSL, Barlow Twins operates on a joint embedding of distorted images.
- It produces two distorted views for all images of a batch X sampled from a dataset.
- The distorted views are obtained via a distribution of data augmentations $\mathcal{T}$.
- The two batches of distorted views Y^A and Y^B are then fed to a function f_θ, typically a deep network with trainable parameters θ, producing batches of embeddings Z^A and Z^B, respectively.
- To simply notations, Z^A and Z^B are assumed to be mean-centered along the batch dimension, such that each unit has mean output 0 over the batch.
- Barlow Twins distinguishes itself from other methods by innovative loss function

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

where $\mathcal{C}$ is the cross-correlation matrix

$$\mathcal{C}_{ij} = \frac{\langle Z^A_{\cdot,i}, Z^B_{\cdot,j} \rangle}{\|Z^A_{\cdot,i}\| \|Z^B_{\cdot,j}\|}$$

where $Z^A_{\cdot,i}$ denotes the ith column of Z^A (and likewise for $Z^B_{\cdot,j}$).

*Department of Mathematical Sciences, KAIST. Email beomsu.kim@kaist.ac.kr

- $\mathcal{C}$ is a square matrix with size the dimension of the network's output, and with values between -1 (perfect anti-correlation) and 1 (perfect correlation).
- Intuitively, the invariance term of the objective, by trying to equate the diagonal elements of the cross-correlation matrix to 1, makes the embeddings invariant to the distortions applied.
- The redundancy reduction term, by trying to equate the off-diagonal elements of the cross-correlation matrix to 0, decorrelates the different vector components of the embedding.
- More formally, Barlow Twins' objective can be understood through the lens of information theory, and specifically as an instantiation of the Information Bottleneck (IB) objective.
- Applied to self-supervised learning, the IB principle posits that a desirable representation should be as informative as possible about the sample represented while being as invariant as possible to distortions of that sample.
- This trade-off is captured by the following loss function

  $$\mathcal{L}_{\texttt{IB}}(\theta) = I(Z_\theta, Y) - \beta I(Z_\theta, X).$$

  Here, X denotes images, Y denotes transformed images, and Z_θ denotes representations, i.e.,

  $$X \xrightarrow{T\sim\mathcal{T}} Y \xrightarrow{f_\theta} Z_\theta.$$

- Using a classical identity for mutual information, we can rewrite $\mathcal{L}_{\texttt{IB}}(\theta)$ as

  $$\mathcal{L}_{\texttt{IB}}(\theta) = [H(Z_\theta) - H(Z_\theta \mid Y)] - \beta[H(Z_\theta) - H(Z_\theta \mid X)]$$

  and $H(Z_\theta \mid Y) = 0$ since Z_θ is a deterministic function of Y. It follows that

  $$\mathcal{L}_{\texttt{IB}}(\theta) = H(Z_\theta \mid X) + \frac{1-\beta}{\beta} H(Z_\theta).$$

- In the case $\beta \leq 1$, the minimum of $\mathcal{L}_{\texttt{IB}}(\theta)$ occurs when the representation is set to a constant that does not depend on the input. For then, $H(Z_\theta \mid X) = H(Z_\theta) = 0$.
- In the case $\beta > 1$, $(1-\beta)/\beta$ becomes negative, so we can replace it by $-\lambda$ for some $\lambda > 0$.

  $$\mathcal{L}_{\texttt{IB}}(\theta) = H(Z_\theta \mid X) - \lambda H(Z_\theta).$$

- Small $H(Z_\theta \mid X)$ means Z_θ is nearly a deterministic function of X, i.e., Z_θ is invariant to $T \sim \mathcal{T}$. This corresponds to the goal of the invariance term in $\mathcal{L}_{\texttt{BT}}(\theta)$.
- Large $H(Z_\theta)$ means Z_θ takes on a diverse set of values. This corresponds to the goal of the redundancy reduction term in $\mathcal{L}_{\texttt{BT}}(\theta)$. Rigorously speaking, for direct correspondence between $H(Z_\theta)$ and the redundancy reduction term in $\mathcal{L}_{\texttt{BT}}(\theta)$, the redundancy reduction term should be computed from the autocorrelation matrix of one of the twin networks, instead of the cross-correlation matrix between two networks. In practice, we do not see a strong difference in performance between these two alternatives.