Preventing Collapses in Non-Contrastive Self-Supervised Learning

Ayoub.E *Nassim E.M*

Université de Montréal

Introduction

- Self-Supervised learning (SSL) has recently emerged as a **scalable solution for learning useful representations without expensive labeling**.
- By understanding dependencies between streams of multimodal data, those methods are promising for building a grounded understanding & accurate world models for future AI methods.
- Traditionally, contrastive approaches learned representations by minimizing the distance between similar data points while maximizing the distance between dissimilar data points.
- On the other hand, recent non-contrastive SSL methods are showing **remarkable performance without any usage of negative pairs.**
- For avoiding any type of collapse in the learning process, those methods are **introducing changes in their architectures/loss function that are not always well understood.**
- Hence, our work was driven by the following question:

How and why successful Non-Contrastive SSL methods avoid any type of collapsing solution?

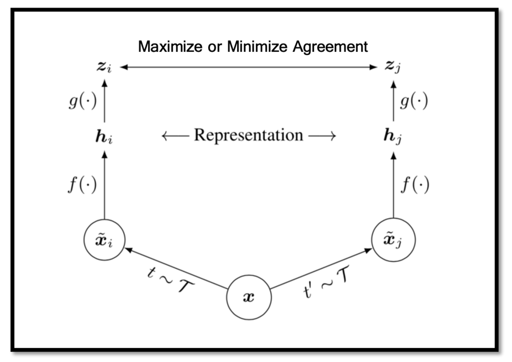
SSL

SSL:

- Capturing dependencies between high dimensional signals
- Learning to predict what's next or what's missing induces a strong representation
- Generating a good representation for downstream tasks without labels during training

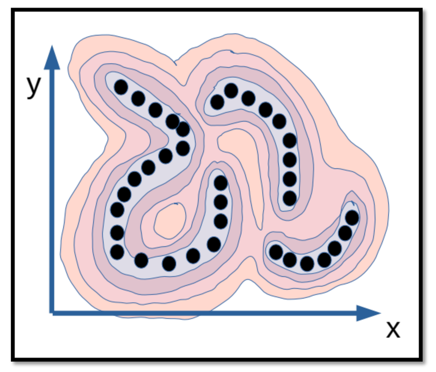
Architectures

- Predictive, Joint-Embedding, Joint-Embedding-Predictive, …
- **Our focus**: Joint-Embedding Architecture (Siamese networks)



- Randomly sample a minibatch of samples
- Apply randomly sampled augmentations
- Representations h produced by base encoder $f(.)$
- Loss operates on an extra projector/expander space from h
- Only the representation is used for downstream tasks

EBM Framework:

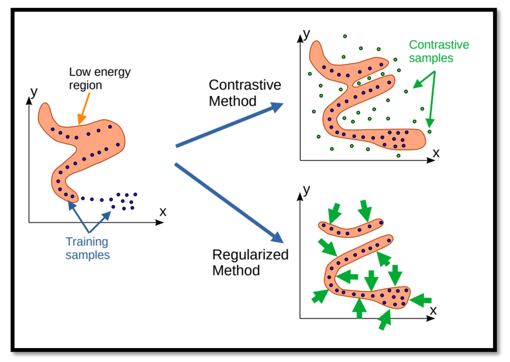


- EBM as a trainable function for assessing incompatibility
- Assign **high energy** to incompatible pairs of points
- Assign **low energy** to compatible pairs of points
- Problem: Fitting the energy landscape

Training Paradigms:

A) Contrastive

- Training samples (low-E) vs contrastive samples (high-E)
- Loss function should push:
- Positive pairs closer / Negative pairs away
- Examples: InfoNCE
- Problems: Poor scaling in high dimensions, hard negative mining, …

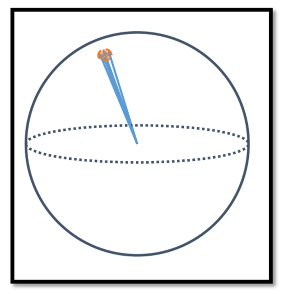


B) Non-Contrastive

- No contrastive (negative) samples used
- Regularizer that minimize the space of possible low-energy

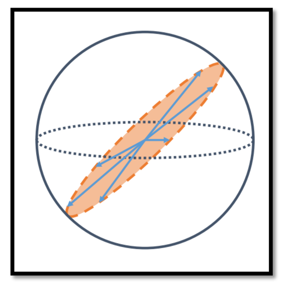
Collapses

A) Total Collapse



- Trivial solution to a loss function that brings closer similar representations
- Ignore the inputs and produce identical and constant output vectors
- Total collapse of the energy landscape where all points are low-energy
- Prevented in contrastive methods via pushing away embeddings of negative pairs

B) Dimensional / Information Collapse



- Across a batch of different inputs:
- Embedding vectors only span a lower-dimensional subspace
- Variables in the latent representations carry redundant information
- Tools to avoid:
  - Loss function
  - Architectural

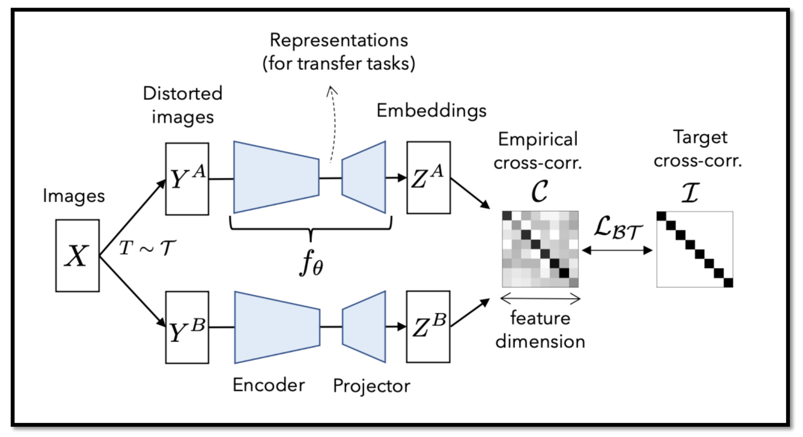
Non-Contrastive Methods

- **Different Categories**: Info Maximization, Self-Distillation, Clustering.
- **Our focus**: Information Maximization Methods
- Maximize the Mutual Information between representations of different views from a shared context

- **Barlow-Twin:**

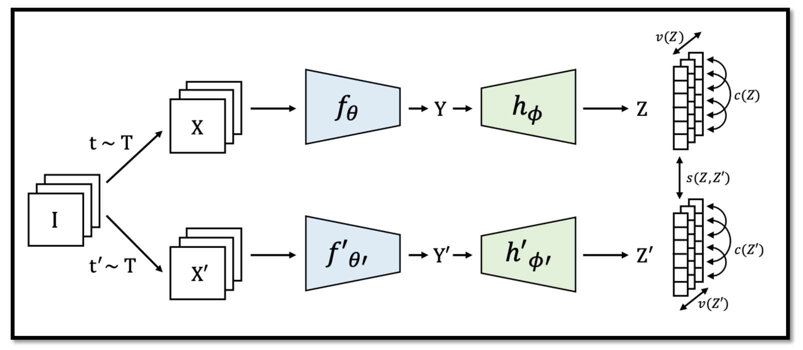
$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1-\mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j\neq i} {\mathcal{C}_{ij}}^2}_{\text{redundancy reduction term}}$$

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b (z^A_{b,i})^2}\sqrt{\sum_i (z^B_{b,j})^2}}$$



- **VICREG:**

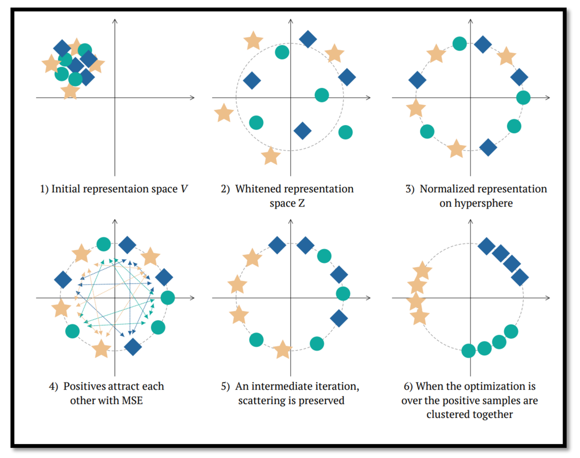
$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}(\alpha \mathrm{Var}(Z_k) + \beta \mathrm{Cov}(Z_k, Z_{k'})) + \gamma \mathrm{Inv}(Z_k, Z_{k'}),$$
$$\mathrm{Var}(Z_k) = \max(0, \gamma - \sqrt{\boldsymbol{C}_{k,k} + \epsilon})$$
$$\mathrm{Cov}(Z_k, Z_{k'}) = \sum_{k'\neq k} (\boldsymbol{C}_{k,k'})^2$$
$$\mathrm{Inv}(Z_k, Z_{k'}) = \|\boldsymbol{Z}_k - \boldsymbol{Z}_{k'}\|_F^2 / N.$$



Invariance: Reduce distance between representations
Variance: Maintains variance of each embedding dimension above a threshold
Covariance: Decorrelates each pair of variables

- **W-MSE**

$$\min_\theta \mathbb{E}\left[\mathrm{dist}(\mathbf{z}_i, \mathbf{z}_j)\right],$$
$$\text{s.t. } \mathrm{cov}(\mathbf{z}_i, \mathbf{z}_i) = \mathrm{cov}(\mathbf{z}_j, \mathbf{z}_j) = I,$$
$$L_{W-MSE}(V) = \frac{2}{Nd(d-1)} \sum dist(\mathbf{z}_i, \mathbf{z}_j)$$
$$Whitening(\mathbf{v}) = W_V(\mathbf{v} - \boldsymbol{\mu}_V).$$



Adding a **whitening operation** on the embeddings (Cholesky decomposition)
This projects vectors onto a spherical distribution (zero-mean and identity-matrix covariance)
1) Computing the inverse covariance matrix of the embeddings
2) Use its square root as a whitening operator on the embeddings

Tools for Avoiding Collapses

Tracking the Dimensional / Information Collapse:

- Singular Value Decomposition
- Embedding space is identified by the singular value spectrum of the covariance matrix on the embedding.
- If the weight matrix W has vanishing singular values, C is also low-rank, indicating collapsed dimensions.

$$C = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$$
$$C = USV^T, S = diag(\sigma^k)$$

- Entropy of embeddings vectors
- Based on the *MultiView InfoMax principle:*
- Maximize the mutual information between the representations of two different views, X and X′, and their corresponding representations, Z and Z′:

$$I(Z, X') = H(Z) - H(Z|X') \geq H(Z) + \mathbb{E}_{x'}[\log q(z|x')]$$

- Only minimizing the cross-entropy loss will result to collapse to a trivial solution, thus a collapse.

- Average correlation coefficient
- Measured by averaging the off-diagonal terms of the correlation matrix of the representations.

Barlow Twins

- Drives the normalized cross-correlation matrix of the two embeddings towards the identity

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1-\mathcal{C}_{ii})^2}_{\text{Diagonal values to Identity}} + \lambda \underbrace{\sum_i \sum_{j\neq i} {\mathcal{C}_{ii}}^2}_{\text{Off Diagonal to zero}}$$

VICReg

- Avoids the collapses with two regularization terms applied to **both embeddings separately**.
  - Multi-Modality advantage against B.T
- Use the covariance matrix of each branch individually for imposing variance / decorrelation
- Fewer constraints on the architecture compared to other methods

W-MSE

- Using a full whitening of the latent space features is sufficient to avoid collapsed representations

- First scatters all the sample representations in a spherical distribution
- Then penalizes the positive pairs which are far from each other

- Downside to the whitening operator on the embeddings:
  - Matrix inversion is a very costly and potentially unstable operation.

Information-Theoretic View

- **Information Bottleneck Principle for SSL:**
- Desirable representation should be as informative as possible about the sample represented
- While being as invariant (non-informative) as possible to distortions (data augmentations)

$$\mathcal{IB}_\theta \triangleq I(Z_\theta, Y) - \beta I(Z_\theta, X)$$

- β is a positive scalar trading off the desire of preserving information and being invariant to distortions.

$$\mathcal{IB}_\theta = [H(Z_\theta) - \overbrace{H(Z_\theta|Y)}^{0}] - \beta[H(Z_\theta) - H(Z_\theta|X)]$$

- Entropy of the representation conditioned on a specific distorted sample cancels to 0 as the function f_θ is deterministic
- Hence the representation Z_θ conditioned on the input sample Y is perfectly known and has zero entropy.

$$\mathcal{IB}_\theta = H(Z_\theta|X) + \frac{1-\beta}{\beta}H(Z_\theta) \rightarrow \mathcal{IB}_\theta = \mathbb{E}_X log\,|\mathcal{C}_{Z_\theta|X}| + \frac{1-\beta}{\beta} log\,|\mathcal{C}_{Z_\theta}|$$

- Simplifying assumption: Representation Z is distributed as a Gaussian (For Friendly Entropy Estimation)
- Entropy of a Gaussian distribution: logarithm of the determinant of its covariance function

- *Additional simplifications and approximations: Replacing the 1−β / β by a new positive constant λ, preceded by a negative sign. Replace the second term of the loss (maximizing the information about samples) by simply minimizing the Frobenius norm of the cross-correlation matrix (off-diagonal terms to 0) (diagonal terms fixed due to rescaling), which creates the surrogate objective that decorrelate all output units*