Preventing Collapses in Non-Contrastive Self-Supervised Learning

Ayoub Echchahed
Department of Computer Science
University of Montreal
ayoub.echchahed@umontreal.ca

Nassim El Massaudi
Department of Computer Science
University of Montreal
nassim.el.massaudi@umontreal.ca

Abstract

Recent self-supervised learning methods for image representation are either classified as contrastive or non-contrastive. Contrastive methods operate by simultaneously minimizing the distance between two augmented views of the same data point (positive pairs) and maximizing views from different data points (negative pairs). While those techniques achieve state of the art results for image classification with few labeled samples, their use of negative pairs is making them limited in many aspects as their scaling potential is limited by the ability of finding suitable contrastive samples for training. Recently, many Non-Contrastive methods have shown competitive results without using any negative pairs, which raises the question of how those methods avoid any type of collapsing solutions to their loss function? Driven by those recent advances, the goal of this project report was to study in detail how and why some recent non-contrastive methods avoid any type of collapses via some specific architectural changes and regularization to their loss functions. In other words, how those methods can avoid trivial solutions that output constant solutions while keeping a high information content in their representations.

1 Introduction

Self-Supervised learning (SSL) has recently emerged as a scalable solution for learning useful representations without expensive labeling. For image representation, most recent methods are classified into two paradigms: contrastive methods and non-contrastive ones. Contrastive approaches learned representations by minimizing the distance between similar data points while also maximizing the distance between dissimilar data points, allowing them to create rich representations during pretraining that can be used later with much fewer labeled samples in order to reach SOA performances in image classification tasks [10, 26]. Nonetheless, those methods are often said to suffer from severe scaling limitations in higher dimensions due to the importance of finding appropriate negative pairs.

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 (trivial or informational), those methods are introducing changes in their architectures/loss function that are not always backed by strong theoretical foundations when it comes to their precise effect [2, 4, 5, 19, 20]. Hence, our work was driven by the following problematic: ***How and why Non-Contrastive Self-Supervised learning methods avoid any type of strong collapsing solutions?***

To make our problem more specific, we decided to focus our analysis on one type of Non-Contrastive SSL methods which are called the information maximization type as those build representations without collapsing solutions by maximizing the Mutual Information between representations of different views extracted from a shared context. The key points of this report are:

Preprint. Under review.

- **Section 2** presents a review of some key concepts in SSL, including the differences between contrastive and non-contrastive methods.
- **Section 3** examines the different types of collapses that SSL methods can suffer, which can lead to poor performances on downstream tasks.
- **Section 4** provides an in-depth review of some non-contrastive methods, including explanations supplemented by experimentations & discussions.
- **Section 5** presents an information-theoretic view of some non-contrastive methods and the related collapse phenomena.

2 Self-Supervised Learning

Self-supervised learning is a technique that obtains supervisory signals from data itself, using the underlying structure in the data to make predictions about any unobserved or hidden part of the input from any observed or unhidden part of the input. It can be used across co-occurring modalities and across large data sets, without relying on any labels. Indeed, by learning to predict what's next or what's missing, this induces a strong learning of the underlying representation of dependencies between high dimensional signals, which could then be used for improving performances in downstream tasks without requiring so many labels during the training phase. For example, self-supervised learning is critical for Natural Language Processing (NLP), allowing the training of large LLMs on massive unlabeled datasets via the masked language model (MLM) training paradigm. Now when it comes to higher dimensional signals such as images in computer vision, self-supervised learning techniques cannot be easily extended to new domains without changes significant changes to the architectures.

One way to learn strong representations of high-dimensional signals like images is to learn to understand the similarity between images that are sharing a similar context. This can be thought of within the unified framework of an energy-based model (EBM), which can be seen as a trainable system that evaluates the incompatibility between two inputs, x and y, by producing a single scalar called the energy. Training an EBM involves two parts: (1) training the model to produce a low energy for compatible x and y pairs, and (2) ensuring that incompatible pairs produce a higher energy than compatible pairs for a particular x. For example, the model could be trained to produce low energies of slightly distorted versions of images coming from the same context, such as pictures of the same dog taken from different angles.

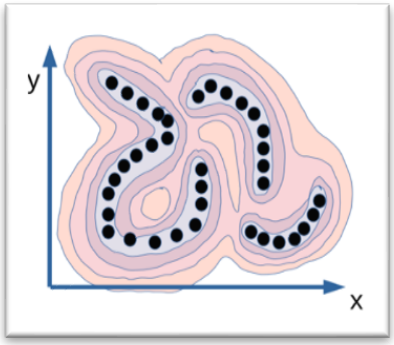


Figure 1: Example of an EBM landscape that act as our similarity assessment system between input x and y. Black dots represent positive pairs.

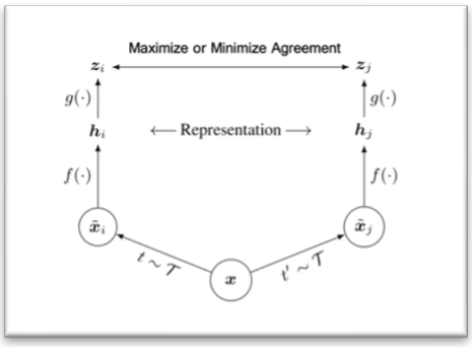


Figure 2: Example of a Joint-Embedding Architecture where a single image x is augmented via sampled transformations t and t' to produce two separate views x_i and x_j.

2.1 Joint-Embedding Architecture

To train this incompatibility system, we need the required SSL architecture and an optimal way of training it. The most common architecture for this type of applications is called the Joint-Embedding Architecture. Given a dataset D with individual image samples x, each data point is augmented via

sampled transformations $t \sim T$ and $t' \sim T$ to obtain two views x and x', which are then fed through a pair of neural networks f_θ and $f'_{\theta'}$ called the encoders. As such, we obtain the representations z and z', which are later used for downstream tasks when the training is over. After that, a pair of expanders (or projectors) h_θ and $h'_{\theta'}$ map the obtained representations into an embedding space of higher dimensions (or lower in case of projectors) where the loss function will be computed on the two resulting embeddings y and y'. In other words, y is defined as $h_\theta\ (f_\theta(x_i))$ and same for y'. We also denote the matrices of embeddings K and K', where $K, K' \in R^{M \times N}$, with M being the embedding size while N the batch size.

2.2 Training Paradigms

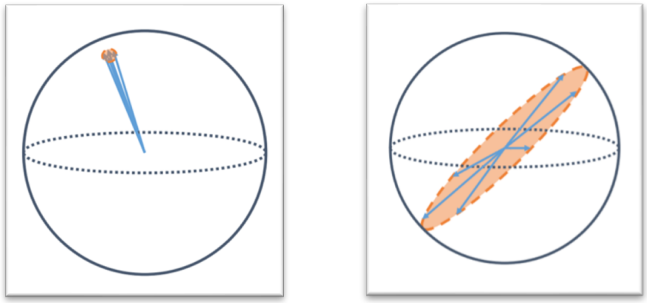
We define two main training paradigms for our similarity system: Contrastive and Non-Contrastive Methods. When training our system using a contrastive method, the objective is to decrease the energy of compatible (Positive) pairs (x, y) while simultaneously increasing the energy of incompatible pairs (Negative Pairs) (x, $\hat{y}$). In other words, when trying to learn what features makes a dog "a dog" to us, we use positive pairs like two augmented views of the same dog while also using negative pairs like a dog and any other objects that "is not" a dog. The contrastive samples should be picked in such a way as to ensure that the EBM assigns higher energies to points outside the regions of high data density. Hence, a suitable loss function should be an increasing function of $F\ (x, y)$ and a decreasing function of $F\ (x, \hat{y})$ in order to create a strong contrast between energies of positive and negative pairs. Now we can the question of whether this paradigm of defining things by what they are not is sustainable at long term. Indeed, selecting suitable incompatible pairs to shape the energy landscape is difficult and computationally expensive as we scale in higher dimensions due to the exponential growth of required negative samples necessary to make an energy surface adopt a good shape. Popular examples of contrastive methods are SimCLR [10] and MoCo [27], both of which are using the InfoNCE loss introduced in [26].

On the other hand, non-contrastive methods are using specific architectural/loss function tricks to rely only on positive pairs while training, which makes them much more promising than their counterpart when it comes to scaling potential. In fact, some of those specific tricks will be reviewed as part of this investigation due to their importance in avoiding different type of collapsing solutions. For example, a network without those tricks could quickly reach a global minimum where producing constant vectors can minimize effectively the training objective, which we would later call a complete collapse. Overall, three main categories of non-contrastive methods exist: Clustering methods, Distillation methods, Information Maximization methods. Clustering methods group samples into clusters based on some similarity measure. On the other hand, distillation-based methods such as BYOL/SimSiam [23] use architectural tricks inspired by distillation to avoid collapsing solutions. Information maximization methods such as VICReg, Barlow-Twins, or W-MSE [2, 4, 5] maximize the informational content of the representations while avoiding most of the architectural tricks introduced by others. For this reason, we decided to focus most of our attention to those recent Information Maximization methods due to their simplicity in avoiding architectural asymmetries via some clever regularization of their loss functions.

3 Collapses Phenomena

3.1 Total Collapse

The first type of collapse that can affect non-contrastive methods is what is called a total collapse, where the model converges during training to a trivial constant solution to the loss function (Global Minima) where all representation vectors are clustered around a single point in the representation space. In other words, the exact same embeddings are produced for all possible input, whether it is a cat, a dog, or a house. This is a consequence of only maximizing the similarity between representations without constraining the model to some specific implicit or explicit constraints. In the context of Energy-Based Models, this means that the energy landscape suffered from a total collapse when given an x, the energy landscape is flat, hence producing the same energy to all values of y. Contrastive methods avoid this type of collapse via their addition of a second term that maximizes the distance between the negative pairs in the training procedure. As an example, the contrastive training will map all representations of dogs close together while separating them from representations of other animals or objects.



(a) Total Collapse (b) Dimensional Collapse

Figure 3: Different types of Collapse

3.2 Dimensional / Informational Collapse

The second type of collapse known as informational or dimensional collapse is much more difficult to avoid than the first one. Also, it can affect both contrastive methods and non-contrastive ones [20]. In theory, an optimal model would produce embedding vectors that span the entire embedding space, therefore maximizing the information content it can encode. However, it was observed that in many contrastive methods, certain dimensions in the embeddings end up collapsing, which produce a lower-dimensional subspace that limit the amount of information that can be encoded in the representation space [2, 20]. Although there is still some learning happening in the training procedure, it is not optimized or simply inefficient as the different vector components of the embeddings are not decorrelated from each other's, hence carrying redundant information between the output units. This means that the goal should be to create full-rank representation vectors, where the represented space corresponds to the singular value spectrum of the covariance matrix $C \in R^{d \times d}$ on the embeddings.

According to the study carried by [20], two main mechanisms can be responsible for this dimensional collapse in contrastive methods. The first one happens when too strong augmentations are produced from the original sample, from which the resulting images are no longer similar enough to be considered as positive pairs. Indeed, the authors found that if strong augmentation produces more variance within a specific feature than traditionally in the data distribution, the weights will collapse in that specific dimension. This was reported to happen when the contrastive covariance matrix (the weighted data distribution covariance matrix minus the weighted augmentation covariance matrix) is not positive semidefinite. The second mechanism that can cause dimensional collapses in contrastive methods is the use of overparameterized linear networks that tend to find low-rank solutions during training. According to their study, gradient descent can cause adjacent layers to align and small initialized singular values to evolve exponentially more slowly than others, resulting in collapsed dimensions. This happens at the condition that the contrastive covariance matrix must be positive semidefinite, hence the opposite of the condition that causes dimensional collapse due to strong augmentations. We limited our investigation of those mechanism due to our focus on tricks used by non-contrastive methods for avoiding collapsing solutions, but their study of the subject is really recommended.

4 Non-Contrastive SSL Methods

SimSiam / BYOL

SimSiam [28] and BYOL [23] are two non-contrastive methods inspired by the concept of Distillation between Siamese networks. The core idea behind those two models relies on the use of architectural modifications instead of regularizations terms. Although those are not the main focus of our analysis as they are not part of the information maximization category, we still decided to provide a brief overview of those methods due to their importance in understanding the overall topic of non-contrastive self-supervised learning. In their most common configurations, both methods use the concept of online and target networks in their architecture. It consists of two neural networks that interact and learn from each other. The online network, defined by a set of weights θ, typically consists of an encoder,

a projector, and a predictor. The target network has the same architecture as the online network but uses a different set of weights usually denoted as ξ. It ideally provides training targets that can improve the online network's representation and does not contribute a gradient. While both BYOL and SimSiam are similar in their architectural aspect which is based, respectively, on an encoder, a projector and a second projector on the online network, they diverge in some specific details as both methods are not strictly using gradient descent. SimSiam relies on a momentum encoder which slowly follows the online network in a delayed fashion through an Exponential Moving Average (EMA). This implementation aims to guarantee a more stable learning. On the other hand, BYOL relies on the stop-gradient approach which consists of freezing the gradient of the target network to stabilize the learning and preventing trivial solution.

As we will not analyze further this method, we include here the 2 key components that are responsible for creating the asymmetric architecture that avoid collapses: the predictor head and the stop-gradient. According to one detailed studies of the dynamics of those methods during training [19], the 2 operations are absolutely essential and removing either of them leads to representational collapse in BYOL and SimSiam. Also, it is interesting to note that experiments performed in [2] showed that even without explicit covariance regularization terms, which are necessary in the following methods, SimSiam and BYOL naturally minimize the average correlation coefficient of the representations, which provide a glimpse on the effect of asymmetric network architectures when it comes to the capacity to avoid informational collapses. Finally, finding from another study suggest that the Batch Normalization layers included in the projection/prediction heads of BYOL are one of the key ingredients which help to avoid degenerate solutions [5].

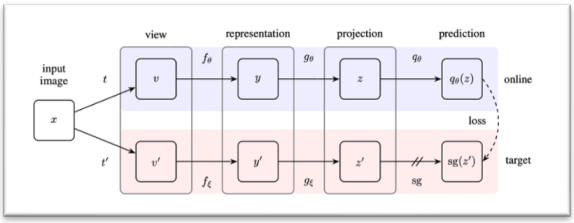


Figure 4: Architecture for BYOL [23]

Barlow-Twins

Barlow Twins [4] is a non-contrastive self-supervised learning technique that is said to be inspired by the neuro-inspired principle of redundancy-reduction, which was introduced by H. Barlow in 1961. In brief, this principle state that the goal of sensory processing is to recode highly redundant sensory inputs into a code with statistically independent components. Inspired by this idea, the authors created a method with an objective function that measures the cross- correlation matrix between the embeddings of two identical networks fed with distorted versions of a batch of samples. Then, it tries to make this matrix as similar as possible to the identity matrix, which causes the embedding vectors of distorted versions of a sample to be similar, while minimizing the redundancy between the components of these vectors. Therefore, the trivial solutions are avoided by design. Although the method does not rely on architectural tricks like stop-gradients or predictor heads, it still requires the use of batch normalization, which is often considered by many to play an important role for avoiding collapses. One of its impressive attributes is its strong robustness in low batch size regimes and its increased performances as the dimensionality of the projector increases. One interesting aspect that was experimented by the authors was whether the addition of symmetry-breaking architectural tricks would improve the performance further, which finally did not have a positive impact on the performance.

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

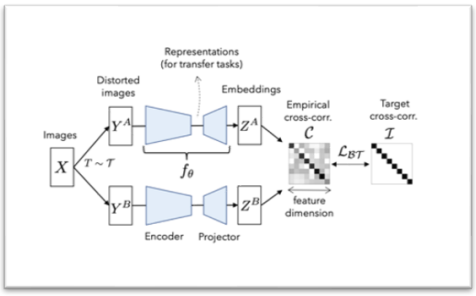


Figure 5: Architecture for Barlow Twins

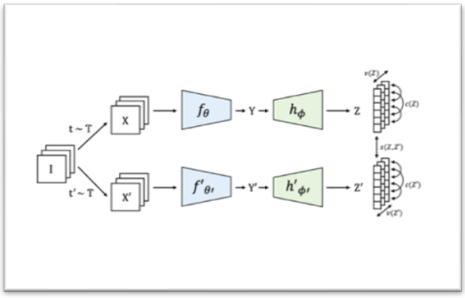


Figure 6: Architecture for VICReg

VICReg

Inspired by Barlow Twins [4], VICReg [2] is an elegant non-contrastive method that is fully "regularized", which means that collapses are only avoided by the different regularizations which are applied to the loss function. Looking at the loss function below, the first term corresponds to the "Invariance component", which reduces the distance between representations. The second term corresponds to the "Variance component", which maintains a threshold on the variance of each embedding dimension. Finally, the third term corresponds to the "Covariance component", which decorrelates each pair of variables in each embedding vectors. In fact, this last term correspond to the redundancy reduction terms of Barlow-Twins. But the methods are nonetheless different in some respects. Indeed, VICReg branches are both regularized independently, which provides the flexibility of working with different type of backbone branches which uses different architectures and data modalities. This is done specifically via the flexibility of imposing various amount of regularization to each branch depending on the statistics of each branch. On the other hand, Barlow Twins' regularization is applied on the cross-correlation matrix, which is more adapted when the two branches are produce outputs with similar statistics.

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}\left(\alpha \operatorname{Var}\left(Z_k\right) + \beta \operatorname{Cov}\left(Z_k, Z_{k'}\right)\right) + \gamma \operatorname{Inv}\left(Z_k, Z_{k'}\right) \tag{2}$$

$$\operatorname{Var}\left(Z_k\right) = \max\left(0, \gamma - \sqrt{C_{k,k} + \epsilon}\right) \tag{3}$$

$$\operatorname{Cov}\left(Z_k, Z_{k'}\right) = \sum_{k' \neq k}\left(C_{k,k'}\right)^2 \tag{4}$$

$$\operatorname{Inv}\left(Z_k, Z_{k'}\right) = \left\|Z_k - Z_{k'}\right\|_F^2 / N \tag{5}$$

Although this method is not contrastive on samples due to the usage of positive pairs only, it is possible to consider them as dimension contrastive. Indeed, given a matrix $A \in R^{n \times n}$, and its extracted diagonal diag(A) $\in R^{n \times n}$, it is possible to define sample-contrastive methods as those that minimize the following contrastive objective: $L_c = \left\|K^{\mathrm{T}}\mathrm{K} - \operatorname{diag}\left(\mathrm{K}^{\mathrm{T}}\mathrm{K}\right)\right\|_{\mathrm{F}}^2$ (Frobenius Norm). In the same optic, we can define dimension-contrastive methods as those that minimize the following non-contrastive objective: $\mathrm{L_{nc}} = \left\|\mathrm{K}\mathrm{K}^{\mathrm{T}} - \operatorname{diag}\left(\mathrm{K}\mathrm{K}^{\mathrm{T}}\right)\right\|_{\mathrm{F}}^2$. In other words, the sample-contrastive objective penalizes the similarity between different pairs of images. On the other hand, the dimension-contrastive objective penalizes the off-diagonal terms of the covariance matrix of the embeddings. Hence, the goal of those objectives can be resumed as respectively trying to make pairs of samples or dimensions orthogonal to each other. [18]

Whitening MSE

Finally, the last method we will review is called W-MSE, which introduces a loss function which is based on the following steps that can be followed using the picture to the right: (1) First, we start with a representation of the batch features in the space V. (2) A whitening operation is applied on the representation space using the Cholesky decomposition-based whitening transform proposed by

(Siarohin et al., 2019) to project the latent-space vectors into a spherical distribution. (3) L2 Normalization is applied on the whitened representation space Z, which constrains the representations tightly on the hypersphere. (4) The MSE is computed over the normalized z features, which encourages the network to move the representations of positive pairs closer together via the $min_\theta E[dist(z_i, z_j)]$ term. As the MSE is computed between normalized vectors, this is equivalent to the cosine similarity objective. Also, another term in the loss function "s.t. $cov(z_j, z_j) = cov(z_j, z_j) = I$" plays a key role in applying a constrain on the distribution of the z values so it needs to be non-degenerate. This is key for avoiding any type of collapse and thus also making all components of z to be linearly independent from each other, which encourages the different dimensions of z to represent different semantic content, hence maximizing the entropy of Z. (5) Subsequent optimization iterations are applied, which move closer the positive pairs, while enforcing the spherical distribution constraint on the other samples. Also, a method called batch slicing is used to avoid high variance in the estimation of the MSE (depends on the whitening matrix W_V, which may have a high variance over consecutive batch iterations). (6) When the optimization is over, positive samples should be clustered together.

An interesting fact is that when the author tried to use Batch Normalization (BN) alone without the whitening operation, this was not sufficient to prevent the network in finding a solution where some high degree of dimensional/informational collapse is introduced. This proves the importance of the whitening operation here. More details on this method are available in the appendix.

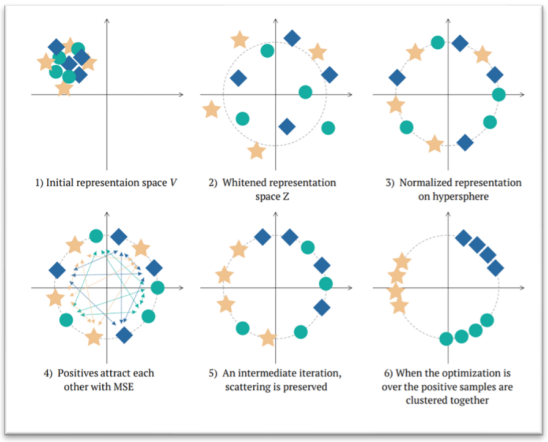


Figure 7: W-MSE Training Steps

4.1 Experimentations

We decided to perform some small-scale experiments that aimed to provide a clear way to quantify the amount of collapse present in the training procedure for different kind of methods and highlights via experimental results the different effects of the regularization terms on the degree of collapse obtained. All the experiments can be accessed and reproduced via a Jupyter Notebook : `https://colab.research.google.com/github/nasselm4i/Deep-Theoretical/blob/main/SSL_Methods_V1.ipynb`

Metrics for tracking the collapse

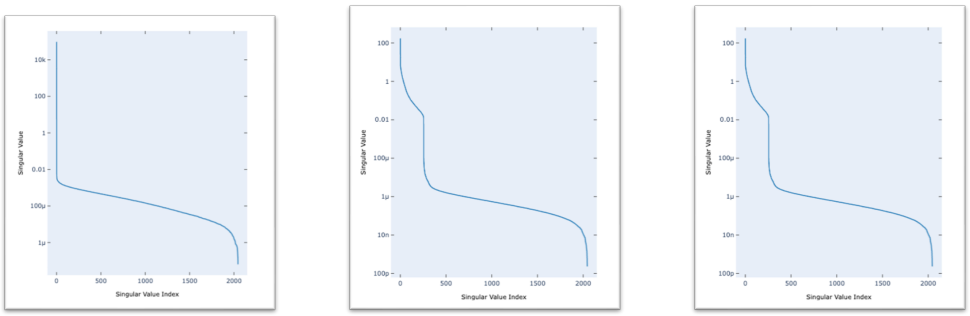
We investigated and considered three different metrics to track the amount of collapse present during the training of various methods. First, we used the singular value decomposition (SVD). If we identify the embedding space by the singular value spectrum of the covariance matrix C on the embedding, dimensional collapses can be quantified by examining the singular values decomposition of this specific matrix. Indeed, if some singular values collapse to zero, it indicates that the corresponding dimensions of the embedding space have collapsed [20]. In other words, the goal is to avoid vanishing singular values as this means that C is low rank, hence indicating collapsed dimensions. In a more mathematical way, given a matrix $A \in \mathbb{R}^{m \times n}$, its SVD is given by $A = U\Sigma V^T$, where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with non-negative diagonal elements, known as singular values, sorted in descending order. The columns of U and V

are called left and right singular vectors, respectively. In the context of dimensional collapse, the covariance matrix $C \in \mathbb{R}^{d \times d}$ of the embedding vectors is computed, and then SVD is applied to this matrix . The singular values of the covariance matrix represent the variances of the principal components of the data. If some singular values collapse to zero, it means that the corresponding dimensions of the embedding space have simply collapsed.

The second measure we decided to use for tracking the amount of collapse/information is the average correlation coefficient, which are measured by averaging the off-diagonal terms of the correlation matrix of the representations. Those provide a reliable source of information when it comes to informational collapse of the embeddings. Finally, we also contemplated the idea of using directly an estimator for the entropy of the distribution of embeddings. Indeed, if the entropy H (Z) goes to 0, a total collapse is observed. However, interpreting this measure in case it does not converge to 0 could potentially provide us with interesting facts when it comes to the amount of information present in the representations during training. Altough we compared two entropy estimators (Pairwise Distance Estimator and the Log-Determinant Estimator) presented in [1] as a quantifiable indicator of collapse, our degree of certainty when it comes to the added benefits in comparaison to the two previous ones was very limited, hence we only used the first ones due to our limitation in time.

Implementations details

The experiments were conducted using the CIFAR-10 dataset due to its practical size when it comes to making fast experimentations loops in a limited time scale with only a few GPUs at hand. The data augmentations were generated using the SimCLR transform library and the VICReg transform library. When it comes to the architecture, the same ResNet18 backbone was used for every model tested. Also, each model was trained with a batch size of 256, a learning rate of 0.06 and approximatively 20 epochs. Those rules were derived from the guidance provided in [2]. Ideally, our number of epochs should have been much higher (around 500), but our limitations for the scope of this project restricted us to this number. Although we could have performed linear evaluations using our representations to assess the quality of the training, we decided that it was not absolutely required for analyzing specifically the collapse phenomena.



(a) VICReg Vanilla (b) VICReg w/o Covariance (c) VICReg w/o Variance

Figure 8: Singular value distribution of the embeddings (not representations) computed on the training set of CIFAR-10 for VICReg Vanilla, VICReg without Variance, and VICReg without Covariance. All methods use 2048 dimensional embeddings

Analysis & Discussion

Empirical results on the VICReg regularization terms have shown some different effects of collapse that we theoretically studied. In Figure 8, the Log Singular Values have been computed from different VICReg methods with a regularization term removed. Thus, we could highlight a clear total collapse when removing the Variance term as this regularization acted as a threshold on the sparsity in each feature of the embedding. Removing it made the method converge to a trivial solution, which caused a total collapse. Meanwhile, Figure 8(b) highlights an Information Collapse; the different features are not decorrelated, resulting in a lower dimensional space, hence a loss of information. This loss is captured as a truncated latent space, which restricts the capacity to encode a wider set of distinct representations.

5 Information-Theoretical View

In this section, we decided to analyze how some of those non-contrastive methods which maximize mutual information across similar views can be linked and analyzed through an Information-Theoretic lens. We first review the Information Bottleneck principle applied to Self-Supervised Learning, then proceed by providing a view of how authors of some previous non-contrastive methods made interesting links with those information-theoretic objectives.

5.1 Information Bottleneck applied to Self-Supervised Learning

The information bottleneck method was introduced in [17] as a sort of rate distortion problem, where the goal is to find the best trade-offs between accuracy and complexity when compressing the information contained in a set of random variables. In other words, we can use a distortion function which measures how well the target Y is predicted from a compressed representation Z compared to a direct prediction from the source X in order to find the optimal amount of compression necessary for Z. This is done in order to maintain our ability to predict Y accurately (up to a certain degree) without needing all the information from the source X. If we reformulate this objective in a self-supervised learning context similar to how [4] presented it, the goal is to obtain a desirable representation Z with the following properties: (A) Being maximally informative about the represented sample. (B) Being maximally invariant (or simply non-informative) to possible distortions applied to the sample.

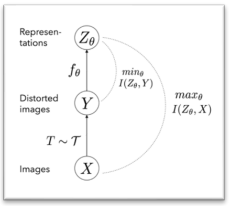


Figure 9: IB applied to SSL

This is equivalent to the formula (6), where β represents a positive scalar that balance the desire of preserving information and the desire of being invariant to distortions. Using the definition of mutual information, the equation can be rewritten into (7). From there, the conditional entropy $H(Z_\theta|Y)$ which represents the entropy of the representation conditioned on a specific distorted sample amount to 0 due to the deterministic nature of the function f_θ. In other words, the representation Z_θ conditioned on the input sample Y is perfectly known and carry zero uncertainty. And as the overall scaling factor of the loss function is not important, the equation can be rearranged into equation (8).

$$\mathcal{IB}_\theta \triangleq I\left(Z_\theta, Y\right) - \beta I\left(Z_\theta, X\right) \tag{6}$$

$$\mathcal{IB}_\theta = \Big[H\left(Z_\theta\right) - \underline{H\left(Z_\theta Y\right)}\Big] - \beta\left[H\left(Z_\theta\right) - H\left(Z_\theta \mid X\right)\right] \tag{7}$$

$$\mathcal{IB}_\theta = H\left(Z_\theta \mid X\right) + \frac{1-\beta}{\beta} H\left(Z_\theta\right) \tag{8}$$

5.2 Recovering Barlow Twins Method

Looking at equation (8), we can see that measuring the entropy is necessary to obtain a measure of our objective function. However, estimating the entropy of a distribution of high-dimensional vectors similar to the obtained representations normally requires large amount of data. A solution to that entropy estimation problem is to make some assumptions about the distribution of the representations. In the case of Barlow-Twins [4], a Gaussian assumption was made on the distribution of representations Z, which allows the entropy of the distribution to be estimated by the logarithm of the determinant of its covariance function, up to a discretization constant that was ignored in their analysis. Therefore, the loss function becomes (9), which they will add some further assumptions and simplifications in order to arrive at their original objective (10).

$$\mathcal{IB}_\theta = \mathbb{E}_X \log \left|\mathcal{C}_{Z_\theta|X}\right| + \frac{1-\beta}{\beta} \log \left|\mathcal{C}_{Z_\theta}\right| \tag{9}$$

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i \left(1 - \mathcal{C}_{ii}\right)^2}_{\text{invariance}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy}} \tag{10}$$

The first term of (9) corresponds to the minimization of the information the representation contains about the distortions. It has the same global optimum than the first term of (10), which maximizes the alignment between representations of pairs of distorted samples. The second term of (9) represents the maximization of the information about samples. It corresponds after some assumptions and simplifications to the second term of (10), also called the redundancy reduction term, which decorrelate all output units to maximize the information content of embeddings, hence eschewing the possibility of informational collapses. To recapitulate, the Barlow Twins loss can be seen as a proxy entropy estimator of the distribution of representations under a Gaussian assumption, which allows to estimate the variability of the embedding from much fewer samples, and on very large dimensional embeddings.

5.3 Link to VICReg

Although the assumption of Gaussian distribution of the embeddings may be questionable at first, a recent study [1] demonstrated in detail that this assumption is nonetheless realistic in many scenarios. Also, the same study shows how the VICReg method presented earlier can be analyzed in terms of information-theoretic quantities, where they derived an objective composed of a regularization part and an invariance part.

Starting from the MultiView InfoMax principle [8], which aims to maximize the mutual information between the representations of two different views, X and X', and their corresponding representations, Z and Z'. To maximize their information, we need to maximize $I(Z; X')$ and $I(Z'; X)$ using the lower bound in (11). There, the authors from the study precise an important point: in equation (11), the entropies are not constant as they can be optimized throughout the learning process. Hence, only minimizing the log loss will cause the representations to collapse to a trivial solution (like we saw earlier), where the entropy term $H(Z)$ will go to 0. To avoid this scenario, regularizing the entropy is necessary via different approaches that implicitly maximize the information content of those embeddings.

$$I\left(Z, X'\right) = H(Z) - H\left(Z \mid X'\right) \geq H(Z) + \mathbb{E}_{x'}\left[\log q\left(z \mid x'\right)\right] \tag{11}$$

$$L \approx \frac{1}{N} \sum_{i=1}^{N} \underbrace{H(Z) - \log\left(\left|\Sigma\left(x_i\right)\right| \cdot \left|\Sigma\left(x_i'\right)\right|\right)}_{\text{Regularizer}} - \underbrace{\frac{1}{2}\left(\mu\left(x_i\right) - \mu\left(x_i'\right)\right)^2}_{\text{Invariance}} \tag{12}$$

$$L \approx \sum_{n=1}^{N} \log \frac{\left|\Sigma_Z\right|}{\left|\Sigma\left(x_i\right)\right| \cdot \left|\Sigma\left(x_i'\right)\right|} - \frac{1}{2}\left(\mu(x) - \mu\left(x'\right)\right)^2 \tag{13}$$

However, the authors of the study precise that computing the regularization term which contains the $H(Z)$ term is not a trivial task as we saw earlier that estimating the entropy of random variables is a classic problem in information theory. After defining Σ_Z as the covariance matrix of Z, they proceed by using the first two moments to approximate the entropy that need to be maximized in order to avoid the collapsing solutions. Then, they end up with the approximation (13), which they use to illustrate how the problem of maximizing the log determinant of Z can be solved by diagonalizing the covariance matrix and increase its diagonal elements. A solution is to push the off-diagonal terms of Σ_Z to be zero and maximize the sum of its log diagonal, which they link to the covariance term of VICReg. And to avoid scenarios where the values on the diagonal become close to zero, which can cause instability when computing the logarithm, they propose to calculate the sum of the diagonal elements directly, which they finally link to the variance term of VICReg.

5.4 Future directions

Although methods like VICReg estimates the entropy of Z only based on the second moment, the authors of the previous study [1] showed that this estimator can be problematic in some instances. Therefore, finding alternative ways to estimate the entropy is an interesting research direction and their study review some key alternatives to this issue.

Also, one method that we found very promising but did not have time to cover is the Principle of Maximal Coding Rate Reduction [14], which uses an alternative measure to the entropy due to the fact that this measure is not well-defined for continuous random variables with degenerate distributions. Therefore, they use another concept in information theory that measures the compactness of a random distribution: the rate distortion measure. This can be seen in simple terms as the minimal number of binary bits needed to encode a random variable z such that the expected decoding error is less than ϵ. Their framework [14] rest on the idea that 2 principles should be respected for having good representations: First, for learned features to be discriminative, features of different classes should be maximally incoherent to each other. This means that they together should span the largest possible space (max. dimensions) and the coding rate of the whole set Z should be as large as possible. Second, learned features of the same class should be highly correlated and coherent, which means that each class should only span a small space that correspond to a very small volume and the coding rate should be as small as possible. Merging those two ideas together gives us a good definition of an optimal representation Z of X: Given a partition Π of Z, an optimal representation Z achieves a large difference between the coding rate for the whole and that for all the subsets. This can be translated into the equation (14) seen below, where the final goal will eventually be to learn a set of features $Z(\theta) = f(X, \theta)$ and their partition Π (if not given in advance) such that they maximize the reduction between the coding rate of all features and that of the sum of features with respect to their classes. Unfortunately, we did have time to go deeper into the details of this non-contrastive SSL method that looks promising as a future research direction. Also, a recent article [6] proposes to enhance the original Principle of Maximal Coding Rate Reduction by improving the significant computational cost intrisic to this method, which we also found worth future investigations.

$$\Delta R(\boldsymbol{Z}, \boldsymbol{\Pi}, \epsilon) \doteq R(\boldsymbol{Z}, \epsilon) - R^c(\boldsymbol{Z}, \epsilon \mid \boldsymbol{\Pi}) \tag{14}$$

6 Conclusion

In summary, we started this project to investigate how and why some non-contrastive self-supervised learning methods avoid any type of strong collapsing solutions due to their use of positive pairs only. For that, we reviewed some key concepts related to the different type of collapses possible, then analyzed a limited number of recent non-contrastive methods and their mechanisms for avoiding collapsing solutions. In fact, most of the methods we focused on use regularization tricks to maximize the information content of the embeddings, which successfully prevent a model collapse. In addition, we performed some limited experiments to quantify precisely the degree of collapsing solutions generated by each method when some specific changes were performed. Finally, we finished our analysis by providing an overview of the collapse phenomena and some non-contrastive methods through an Information-Theoretic lens, where many parallels can be drawn between information-theoretic measures and the previously analyzed methods.

References

[1] Shwartz-Ziv, R., Balestriero, R., Kawaguchi, K., Rudner, T. G., et LeCun, Y. (2023). An Information-Theoretic Perspective on Variance-Invariance-Covariance Regularization. arXiv preprint arXiv:2303.00633.

[2] Bardes, Adrien, Jean Ponce, and Yann LeCun. "Vicreg: Variance-invariance-covariance regularization for self-supervised learning." arXiv preprint arXiv:2105.04906 (2021).

[3] Shwartz-Ziv, Ravid, Randall Balestriero, and Yann LeCun. "What Do We Maximize in Self-Supervised Learning?." arXiv preprint arXiv:2207.10081 (2022).

[4] Zbontar, Jure, et al. "Barlow twins: Self-supervised learning via redundancy reduction." International Conference on Machine Learning. PMLR, 2021.

[5] Ermolov, Aleksandr, et al. "Whitening for self-supervised representation learning." International Conference on Machine Learning. PMLR, 2021.

[6] Baek, Christina, et al. "Efficient maximal coding rate reduction by variational forms." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[7] Cabannes, Vivien, et al. "The SSL Interplay: Augmentations, Inductive Bias, and Generalization." arXiv preprint arXiv:2302.02774 (2023).

[8] Bachman, Philip, R. Devon Hjelm, and William Buchwalter. "Learning representations by maximizing mutual information across views." Advances in neural information processing systems 32 (2019).

[9] Belghazi, Mohamed Ishmael, et al. "Mine: mutual information neural estimation." arXiv preprint arXiv:1801.04062 (2018).

[10] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

[11] Hjelm, R. Devon, et al. "Learning deep representations by mutual information estimation and maximization." arXiv preprint arXiv:1808.06670 (2018).

[12] LeCun, Yann (2022). A Path Towards Autonomous Machine Intelligence, Version 0.9.2, 2022-06-27.

[13] Tschannen, Michael, et al. "On mutual information maximization for representation learning." arXiv preprint arXiv:1907.13625 (2019).

[14] Yu, Yaodong, et al. "Learning diverse and discriminative representations via the principle of maximal coding rate reduction." Advances in Neural Information Processing Systems 33 (2020): 9422-9434.

[15] Tschannen, Michael, et al. "On mutual information maximization for representation learning." arXiv preprint arXiv:1907.13625 (2019).

[16] Assran, Mahmoud, et al. "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture." arXiv preprint arXiv:2301.08243 (2023).

[17] Tishby, Naftali; Pereira, Fernando C.; Bialek, William (September 1999). The Information Bottleneck Method (PDF). The 37th annual Allerton Conference on Communication, Control, and Computing. pp. 368–377.

[18] Garrido, Quentin, et al. "On the duality between contrastive and non-contrastive self-supervised learning." arXiv preprint arXiv:2206.02574 (2022).

[19] Tian, Yuandong, Xinlei Chen, and Surya Ganguli. "Understanding self-supervised learning dynamics without contrastive pairs." International Conference on Machine Learning. PMLR, 2021.

[20] Jing, Li, et al. "Understanding dimensional collapse in contrastive self-supervised learning." arXiv preprint arXiv:2110.09348 (2021).

[21] Becker, Suzanna, and Geoffrey E. Hinton. "Self-organizing neural network that discovers surfaces in random-dot stereograms." Nature 355.6356 (1992): 161-163.

[22] Zhang, Chaoning, et al. "How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning." arXiv preprint arXiv:2203.16262 (2022).

[23] Grill, Jean-Bastien, et al. "Bootstrap your own latent-a new approach to self-supervised learning." Advances in neural information processing systems 33 (2020): 21271-21284.

[24] Richemond, Pierre H., et al. "Byol works even without batch statistics." arXiv preprint arXiv:2010.10241 (2020).

[25] Wen, Zixin, and Yuanzhi Li. "The mechanism of prediction head in non-contrastive self-supervised learning." arXiv preprint arXiv:2205.06226 (2022).

[26] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).

[27] He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[28] Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

7 Appendix

7.1 Contrastive against Regularized Methods

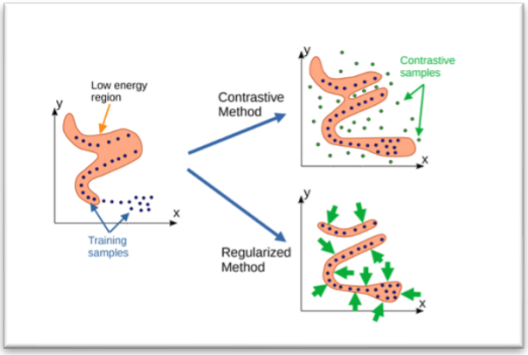


Figure 10: Two types of training paradigms [12]

7.2 Complete objective function based on information-theoretical first principles (See [1])

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

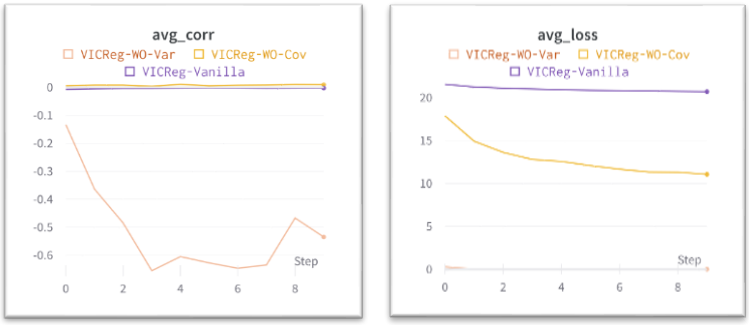
$$I\left(Z; X'\right) = H(Z) + \mathbb{E}_{x,z|x,x',z'|x'}\left[\log q\left(z \mid z'\right)\right]$$

$$= H(Z) + \frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{x,x'}\left[\left(\mu(x) - \mu\left(x'\right)\right)^2 + \log\left(|\Sigma(x)| \cdot |\Sigma\left(x'\right)|\right)\right]$$

$$L \approx \frac{1}{N}\sum_{i=1}^{N} \underbrace{H(Z) - \log\left(|\Sigma\left(x_i\right)| \cdot |\Sigma\left(x_i'\right)|\right)}_{\text{Regularizer}} \underbrace{- \frac{1}{2}\left(\mu\left(x_i\right) - \mu\left(x_i'\right)\right)^2}_{\text{Invariance}}$$

$$L \approx \sum_{n=1}^{N} \log \frac{|\Sigma_Z|}{|\Sigma\left(x_i\right)| \cdot |\Sigma\left(x_i'\right)|} - \frac{1}{2}\left(\mu(x) - \mu\left(x'\right)\right)^2$$

7.3 Additional Figures



(a) Average Correlation Coefficient (b) Average loss

Figure 11: Effect of removing the regularization terms on the loss and the average correlation coefficient, which is computed as the off-diagonal mean of the correlation matrix for each embedding

7.4 W-MSE Details

Going into slightly more details, the W-MSE loss involves computing the mean squared error (MSE) over all positive pairs of features in a batch of size $K = Nd$, where N is the number of original images and d is the number of positive pair. We use the reparameterization of the feature variables $\mathbf{v}$ with the whitened variables $\mathbf{z}$, obtained through the whitening function $Whitening(\mathbf{v}) = W_V (\mathbf{v} - \boldsymbol{\mu}_V)$, where W_V is computed using the Cholesky decomposition and the inverse of L, a lower triangular matrix obtained through the decomposition of the covariance matrix Σ_V of V. The W-MSE loss is defined as:

$$L_{W-MSE}(V) = \frac{2}{Nd(d-1)} \sum \text{dist}\left(\mathbf{z}_i, \mathbf{z}_j\right) \tag{15}$$

where $\mathbf{z} = Whitening(\mathbf{v})$, and $\boldsymbol{\mu}_V$ is the mean of the elements in V. The full whitening of each $\mathbf{v}_i \in V$ is performed using the resulting set of vectors $Z = \mathbf{z}_1, \ldots, \mathbf{z}_K$. The Cholesky decomposition used to compute W_V is fully differentiable. The inverse of L is obtained as $W_V = L^{-1}$, where L satisfies the factorization $\Sigma_V = LL^\top$, with L being a lower triangular matrix. The circular shape comes from the product $\Sigma_V = LL^\top$ which results in a hypersphere shape.

7.5 Limitations to our experiments

We initially started comparing 5 methods: BYOL/SimSiam, VICReg, Barlow-Twins, and W-MSE. But due to our specific focus on information maximization methods and the slow training of W-MSE caused by the heavy matrix inversions procedures, we decided to focus our efforts only on VICReg and Barlow-Twins, although both methods share many similarities. If we were making this project on a longer time scale, we would have decided to compare the collapsing degree of all those methods. In addition to previous limitations mentioned, we also lacked time to analyze the degree of collapse in relation with the sensibility of the methods to specific hyperparameters, batch sizes, and embedding dimensions. Therefore, better experiments should be performed with bigger datasets (ex: ImageNet), more epochs, a higher number of different non-contrastive methods, more hyperparameters tuning, and more changes to the vanilla architectures in order to analyze the impact of designs on the collapsing degree.