What Do We Maximize in Self-Supervised Learning?

Ravid Shwartz-Ziv [1] **Randall Balestriero** [2] **Yann LeCun** [1 2]

Abstract

In this paper, we examine self-supervised learning methods, particularly VICReg, to provide an information-theoretical understanding of their construction. As a first step, we demonstrate how information-theoretic quantities can be obtained for a deterministic network, offering a possible alternative to prior work that relies on stochastic models. This enables us to demonstrate how VICReg can be (re)discovered from first principles and its assumptions about data distribution. Furthermore, we empirically demonstrate the validity of our assumptions, confirming our novel understanding of VICReg. Finally, we believe that the derivation and insights we obtain can be generalized to many other SSL methods, opening new avenues for theoretical and practical understanding of SSL and transfer learning.

1. Introduction

Self-Supervised Learning (SSL) algorithms (Bromley et al., 1993) learn representations using a proxy objective (i.e., SSL objective) between inputs and self-defined signals. The results indicate that the learned representations can generalize well to a wide range of downstream tasks (Chen et al., 2020; Misra & Maaten, 2020), even when the SSL objective does not use downstream supervision during training. In SimCLR (Chen et al., 2020), for example, a contrastive loss is defined between images with different augmentations (i.e., one as input and the other as a self-supervised signal). Then, we take our pre-learned model as a feature extractor and adopt the features to various applications, including image classification, object detection, instance segmentation, and pose estimation (Caron et al., 2021). However, despite the success in practice, only a few works (Arora et al., 2019; Lee et al., 2021a) provide theoretical insights into the learning efficacy of SSL.

In recent years, information theory methods have played a key role in several notable deep learning achievements, from practical applications in representation learning as the variational information bottleneck (Alemi et al., 2016), to theoretical investigations (e.g., the generalization bound induced by mutual information (Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020; Shwartz-Ziv, 2022). Moreover, different deep learning problems have been successfully approached by developing and applying novel estimators and learning principles derived from information-theoretic quantities, such as mutual information estimation. Many works have attempted to analyze SSL from an information theory perspective. An example is the use of the mutual information neural estimator (MINE) (Belghazi et al., 2018) in representation learning (Hjelm et al., 2018) in conjunction with the renowned information maximization (InfoMax) principle (Linsker, 1988). However, looking at these works may be confusing. Numerous objective functions are presented, some contradicting each other, as well as many implicit assumptions. Moreover, these works rely on a crucial assumption: a stochastic (often Gaussian) DN mapping, which is rarely the case nowadays.

This paper presents a unified framework for SSL methods from an information theory perspective which can be applied to deterministic DN training. We summarize our contributions into two points: (i) Firdt, in order to study deterministic DNs from an information theory perspective, we shift stochasticity to the DN input, which is a much more faithful assumption for current training techniques. (ii) Second, based on this formulation, we analyze how current SSL methods that use deterministic networks optimize information-theoretic quantities.

2. Background

Continuous Piecewise Affine (CPA) Mappings. A rich class of functions emerges from piecewise polynomials: spline operators. In short, given a partition Ω of a domain $\mathbb{R}^D$, a spline of order k is a mapping defined by a polynomial of order k on each region $\omega \in \Omega$ with continuity constraints on the entire domain for the derivatives of order $0,\ldots,k-1$. As we will focus on affine splines ($k=1$), we define this case only for concreteness. An K-dimensional affine spline

[*]Equal contribution [1]New York Universityy [2]Meta AI Research. Correspondence to: Ravid Shwartz-Ziv <ravidziv@gmail.com>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022.

f produces its output via

$$f(\boldsymbol{z}) = \sum_{\omega \in \Omega} (\boldsymbol{A}_\omega \boldsymbol{z} + \boldsymbol{b}_\omega) \mathbb{1}_{\{\boldsymbol{z} \in \omega\}}, \tag{1}$$

with input $\boldsymbol{z} \in \mathbb{R}^D$ and $\boldsymbol{A}_\omega \in \mathbb{R}^{K \times D}, \boldsymbol{b}_\omega \in \mathbb{R}^K, \forall \omega \in \Omega$ the per-region *slope* and *offset* parameters respectively, with the key constraint that the entire mapping is continuous over the domain $f \in \mathcal{C}^0(\mathbb{R}^D)$. Spline operators and especially affine spline operators have been widely used in function approximation theory (Cheney & Light, 2009), optimal control (Egerstedt & Martin, 2009), statistics (Fantuzzi et al., 2002), and related fields.

Deep Networks. A deep network (DN) is a (non-linear) operator f_Θ with parameters Θ that map a *input* $\boldsymbol{x} \in \mathbb{R}^D$ to a *prediction* $\boldsymbol{y} \in \mathbb{R}^K$. The precise definitions of DNs operators can be found in Goodfellow et al. (2016). We will omit the Θ notation for clarity unless needed. The only assumption we require for our study is that the non-linearities present in the DN are CPA, as is the case with (leaky-) ReLU, absolute value, and max-pooling. In that case, the entire input-output mapping becomes a CPA spline with an implicit partition Ω, the function of the weights and architecture of the network (Montufar et al., 2014; Balestriero & Baraniuk, 2018). For smooth nonlinearities, our results hold from a first-order Taylor approximation argument.

Self-Supervised Learning. Joint embedding methods learn the DN parameters Θ without supervision and input reconstruction. Due to this formulation, the difficulty of SSL is to produce a good representation for downstream tasks whose labels are not available during training —while avoiding a trivially simple solution where the model maps all inputs to constant output. Many methods have been proposed to solve this problem. *Contrastive methods* learn representations by contrasting positive and negative examples, e.g. SimCLR (Chen et al., 2020) and its InfoNCE criterion (Oord et al., 2018). Other recent work introduced *non-contrastive methods* that employ different regularization methods to prevent collapsing of the representation. Several papers used stop-gradients and extra predictors to avoid collapse (Chen & He, 2021; Grill et al., 2020) while Caron et al. (2020) uses an additional clustering step. As opposed to contrastive methods, noncontrastive methods do not explicitly rely on negative samples. Of particular interest to us is the *VICReg* method (Bardes et al., 2021) that considers two embedding batches $\boldsymbol{Z} = [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_N)]$ and $\boldsymbol{Z}' = [f(\boldsymbol{x}'_1), \dots, f(\boldsymbol{x}'_N)]$ each of size $(N \times K)$. Denoting by $\boldsymbol{C}$ the $(K \times K)$ covariance matrix obtained from $[\boldsymbol{Z}, \boldsymbol{Z}']$ we obtain the VICReg triplet loss

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

Our goal will now be to formulate SSL as an information-theoretic problem from which we can precisely relate VICReg to known methods even with a deterministic network.

Deep Networks and Information-Theory. Recently, information-theoretic methods have played a key role in several remarkable deep learning achievements (Alemi et al., 2016; Xu & Raginsky, 2017; Steinke & Zakynthinou, 2020; Shwartz-Ziv & Tishby, 2017). Moreover, different deep learning problems have been successfully approached by developing and applying information-theoretic estimators and learning principles (Hjelm et al., 2018; Belghazi et al., 2018; Piran et al., 2020; Shwartz-Ziv et al., 2018). There is, however, a major problem when it comes to analyzing information-theoretic objectives in deterministic deep neural networks: the source of randomness. The mutual information between the input and the representation in such networks is infinite, resulting in ill-posed optimization problems or piecewise constant, making gradient-based optimization methods ineffective (Amjad & Geiger, 2019). To solve these problems, researchers have proposed several solutions. For SSL, stochastic deep networks with variational bounds could be used, where the output of the deterministic network is used as parameters of the conditional distribution (Lee et al., 2021b; Shwartz-Ziv & Alemi, 2020). Dubois et al. (2021) suggested another option, which assumed that the randomness of data augmentation among the two views is the source of stochasticity in the network. For supervised learning, Goldfeld et al. (2018) introduced an auxiliary (noisy) DN framework by injecting additive noise into the model and demonstrated that it is a good proxy for the original (deterministic) DN in terms of both performance and representation. Finally, Achille and Soatto (Achille & Soatto, 2018) found that minimizing a stochastic network with a regularizer is equivalent to minimizing cross-entropy over deterministic DNs with multiplicative noise. However, all of these methods assume that the noise comes from the model itself, which contradicts current training methods. In this work, we explicitly assume that the stochasticity comes from the data, which is a less restrictive assumption and does not require changing current algorithms.

3. Information Maximization of Deep Networks Outputs

This section first sets up notation and assumption on the information-theoretic challenges in self-supervised learning (Section 3.1) and on our assumptions regarding the data distribution (Section 3.2) so that any training sample $\boldsymbol{x}$ can be seen as coming from a single Gaussian distribution as in $\boldsymbol{x} \sim \mathcal{N}(\mu_{\boldsymbol{x}}, \Sigma_{\boldsymbol{x}})$. From this we obtain that the output of any deep network $f(\boldsymbol{x})$ corresponds to a mixture of truncated Gaussian (Section 3.3). In particular, it can fall back to a single Gaussian under some small noise ($\det(\Sigma) \to \epsilon$) assumptions. These results will enable information measures to be applied to deterministic DNs. We then recover the known SSL methods (Bardes et al., 2021) by making different assumptions about the data distribution and estimating their information.

3.1. SSL as an Information-Theoretic Problem

To better grasp the difference between key SSL methods, we first formulate the general SSL goal from an information-theoretical perspective.

We start with the *MultiView InfoMax* principle, i.e., maximizing the mutual information between the representations of the two views. To do so, as shown in Federici et al. (2020), we need to maximize $I(Z; X')$ and $I(Z'; X)$. We can do so by a lower bound using

$$I(Z,X'){=}H(Z){-}H(Z|X'){\geq}H(Z){+}E[\log q(z|x')] \quad (2)$$

where $H(Z)$ is the entropy of Z. In supervised learning, where we need to maximize $I(Z; Y)$, the labels (Y) are fixed, the entropy term $H(Y)$ is constant, and you only need to optimize the log-loss $E[\log q(z|x)]$ (cross-entropy or square loss). However, in SSL, the entropy $H(Z)$ and $H(Z')$ are not constant and can be optimized throughout the learning process. Therefore, only maximizing $E[\log q(z|x')]$ will cause it to collapse to the trivial solution of making the representations constant (where the entropy goes to zero). To regularize these entropies, i.e., prevent collapse, different methods utilize different approaches to implicit regularizing information. To recover them in Section 4, we must first introduce the notation and results around the data distribution (Section 3.2) and how a DN transforms that distribution (Section 3.3).

3.2. Data Distribution Hypothesis

Our first step is to assess how the output random variables of the network are represented, assuming a distribution on the data itself. Under the manifold hypothesis, any point can be seen as a Gaussian random variable with a low-rank covariance matrix in the direction of the manifold tangent space of the data. Therefore, we will consider throughout this study the conditioning of a latent representation with respect to the mean of the observation, i.e., $X|\boldsymbol{x}^* \sim \mathcal{N}(\boldsymbol{x}^*, \Sigma_{\boldsymbol{x}^*})$ where the eigenvectors of $\Sigma_{\boldsymbol{x}^*}$ are in the same linear subspace than the tangent space of the data manifold at $\boldsymbol{x}^*$, which varies with the position of $\boldsymbol{x}^*$ in space.

Hence a dataset is considered to be a collection of $\{\boldsymbol{x}^*_n, n = 1, \ldots, N\}$ and the full data distribution to be a sum of low-rank covariance Gaussian densities as in

$$X \sim \sum_{n=1}^{N} \mathcal{N}(\boldsymbol{x}^*_n, \Sigma_{\boldsymbol{x}^*_n})^{T=n}, T \sim \mathrm{Cat}(N), \qquad (3)$$

with T the uniform Categorical random variable. To keep things simple and without loss of generality, we consider that the effective support of $\mathcal{N}(\boldsymbol{x}^*_i, \Sigma_{\boldsymbol{x}^*_i})$ and $\mathcal{N}(\boldsymbol{x}^*_j, \Sigma_{\boldsymbol{x}^*_j})$ do not overlap. This keeps things general, as it is enough to cover the domain of the data manifold overall, without overlap between different Gaussians. Hence, in general, we have that.

$$p(\boldsymbol{x}) \approx \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{x}^*_{n(\boldsymbol{x})}, \Sigma_{\boldsymbol{x}^*_{n(\boldsymbol{x})}}\right)/N, \qquad (4)$$

where $\mathcal{N}(\boldsymbol{x}; ., .)$ is the Gaussian density at $\boldsymbol{x}$ and with $n(\boldsymbol{x}) = \arg\min_n (\boldsymbol{x}-\boldsymbol{x}^*_n)^T \Sigma_{\boldsymbol{x}^*_n} (\boldsymbol{x}-\boldsymbol{x}^*_n)$. This assumption that a dataset is a mixture of Gaussians with nonoverlapping support will simplify our derivations below, which could be extended to the general case if needed. Note that this is not restrictive since, given a sufficiently large N, the above can represent any manifold with an arbitrarily good approximation.

3.3. Data Distribution After Deep Network Transformation

Consider an affine spline operator f (Eq. 1) that goes from a space of dimension D to a space of dimension K with $K \geq D$. The span, that we denote as image, of this mapping is given by

$$Im(f) \triangleq \{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^D\} = \bigcup_{\omega \in \Omega} \mathrm{Aff}(\omega; \boldsymbol{A}_\omega, \boldsymbol{b}_\omega) \quad (5)$$

with $\mathrm{Aff}(\omega; \boldsymbol{A}_\omega, \boldsymbol{b}_\omega) = \{\boldsymbol{A}_\omega \boldsymbol{x} + \boldsymbol{b}_\omega : \boldsymbol{x} \in \omega\}$ the affine transformation of region ω by the per-region parameters $\boldsymbol{A}_\omega, \boldsymbol{b}_\omega$, and with Ω the partition of the input space in which $\boldsymbol{x}$ lives in. We also provide an analytical form of the per-region affine mappings in Section 2. Hence, the DN mapping consists of affine transformations on each input space partition region $\omega \in \Omega$ based on the coordinate change induced by $\boldsymbol{A}_\omega$ and the shift induced by $\boldsymbol{b}_\omega$.

When the input space is equipped with a density distribution, this density is transformed by the mapping f. In general, finding the density of $f(X)$ is an intractable task. However, given our disjoint support assumption provided in Section 3.2, we can arbitrarily increase the representation power of the density by increasing the number of prototypes N. In doing so, the support of each Gaussian is included with the region ω in which its means lie in, leading to the following result.

Theorem 1. *Given the setting of Equation (4) the unconditional DN output density denoted as Z is a mixture of the affinely transformed distributions $\boldsymbol{x}|\boldsymbol{x}^*_{n(\boldsymbol{x})}$ e.g. for the Gaussian case*

$$Z \sim \sum_{n=1}^{N} \mathcal{N}\Big(\boldsymbol{A}_{\omega(\boldsymbol{x}^*_n)} \boldsymbol{x}^*_n + \boldsymbol{b}_{\omega(\boldsymbol{x}^*_n)}, \boldsymbol{A}^T_{\omega(\boldsymbol{x}^*_n)} \Sigma_{\boldsymbol{x}^*_n} \boldsymbol{A}_{\omega(\boldsymbol{x}^*_n)}\Big)^{T=n},$$

where $\omega(\boldsymbol{x}^*_n) = \omega \in \Omega \iff \boldsymbol{x}^*_n \in \omega$ is the partition region in which the prototype $\boldsymbol{x}^*_n$ lives in.

The proof of the above involves the fact that if $\int_\omega p(\boldsymbol{x}|\boldsymbol{x}^*_{n(\boldsymbol{x})})d\boldsymbol{x} \approx 1$ then f is linear within the effective support of p. Therefore, any sample from p will almost surely lie within a single region $\omega \in \Omega$ and therefore the entire mapping can be considered linear with respect to p. Thus, the output distribution is a linear transformation of the input distribution based on the per-region affine mapping.

4. Information Optimization and Optimality

Based on our analysis, we now show how specific SSL algorithms can be derived. According to Section 3.1, we want to maximize $I(Z;X')$ and $I(Z';X)$. When our input noise is small, we can reduce the conditional output density $p_{\boldsymbol{z}|\boldsymbol{x}*}$ to a single Gaussian.

$$(Z'|X' = x_n) \sim \mathcal{N}(\mu_n, \Sigma_n),$$

where we abbreviated the parameters. Using that, and the result from Section 3.1, we see that one should optimize both $H(Z|X')$ and $H(Z)$. As in a standard regression task, we assume a Gaussian observation model, which means $p(z|z') \sim \mathcal{N}(z', \Sigma_r)$. Using the mean square error as a loss function in regression tasks is a particular application of this assumption, where $\Sigma_r = I$. To compute the expected loss, we need to marginalize out the stochasticity in Z', which means that the conditional decoding map is a Gaussian:

$$q(z|X' = x^*_n) = \int q(z|z')p(z'|x^*_n)dz', \tag{6}$$

which gives the distribution $\mathcal{N}(\mu_{n'}, \Sigma_r + \Sigma_{n'})$ meaning that we can lower bound the mutual information with

$$I(Z;X') \geq H(Z) + E[\log(q(z|x')] = H(Z) - \frac{d}{2}\log 2\pi\Sigma_r \\ - \sum_{n=1}^{N}\frac{1}{2}(z_n - z'_n)^T\Sigma_r^{-1}(z_n - z'_n) - \log|\Sigma_n|. \tag{7}$$

What happens if we attempt to optimize this objective? The only intractable component is the entropy of Z. We will begin by examining Z itself. It is natural to ask why the entropy of Z will not increase to infinity. Intuitively, the answer is that $H(Z)$ and $H(Z|X')$ are tied together, and one cannot increase without the other. Now, recalling that under our distribution assumption, Z is a mixture of Gaussian (recall Thm. 1), we can see how the existing upper and lower bounds could be used for this case; for example, the ones in Moshksar & Khandani (2016).

4.1. Deriving VICReg From First Principles

We now propose to recover VICReg from the first principles per the above information-theoretic principle.

Recall that our goal is to estimate the entropy $H(Z)$ in Equation (7), where Z is a Gaussian mixture. This quantity is not known for a mixture of Gaussians due to the logarithm of a sum of exponential functions, except for the special case of a single Gaussian density. There are, however, several approximations in the literature that include both upper and lower bounds. Among the methods, some use the logarithmic sum of the probability (Kolchinsky & Tracey, 2017), and some use entropy-adjusted logarithmic probabilities (Huber et al., 2008).

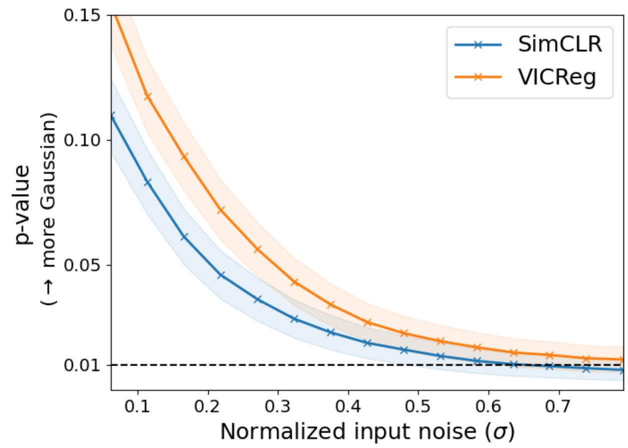

Figure 1. **The network output with VICReg training is more gaussian for small input noise**. The P-value of the normality test for different SSL models trained on CIFAR-10 for different input noise levels. The x-axis is the coefficient that multiplies the data distribution standard deviation to obtain the Gaussian standard deviation that samples around each image. The dashed line represents the point at which the null hypothesis (Gaussian distribution of the network output) can be rejected with 99% confidence.

An even simpler solution is to approximate the entire mixture as a single Gaussian by capturing only the first two moments of the mixture distribution. Since the Gaussian distribution maximizes the entropy for a given covariance matrix, this method provides an upper bound approximation of our entropy of interest $H(Z)$. In this case, denoting by Σ_Z is the covariance matrix of Z, we find that we should maximize the following objective:

$$\max_{Z,Z'}\sum_{n=1}^{N}\log\frac{|\Sigma_Z|}{|\Sigma_n|} - \frac{1}{2}(z_n - z'_n)^T\Sigma_r^{-1}(z_n - z'_n) - \frac{d}{2}\log 2\pi\Sigma_r$$

where Σ_r is constant with respect to our optimization process, and the second term is the prediction performance of one representation from the other. A key result from Shi et al. (2009) connects the eigenvectors and eigenvalues of Σ_Z and those of each component $\Sigma_i, \forall i$, and showed that under the assumption that the separation $(\mu_i - \mu_j)\Sigma_i^{-1}(\mu_i - \mu_j)^T$ between the different components is large enough —which holds true in our case as per our data distribution model— the eigenfunctions of $\Sigma_i, \forall i$ are approximately the eigenfunctions of Σ_Z. Therefore, in our case, this means that since all those eigenvalues are tied, we only need to find the most efficient way to maximize $|\Sigma_Z|$.

We know that the determinant of the matrix is the product of its eigenvalues. For every positive matrix, the maximum eigenvalue is greater than or equal to each diagonal element. Therefore, under the constraint of the eigenvalues of the matrix, the most efficient way is to decrease the off-diagonal terms and increase the diagonal terms. By setting $\Sigma_r = I$, *we therefore fully recover the VICReg objective*.

4.2. SimCLR vs VICReg

Lee et al. (2021b) connected the SimCLR objective (Chen et al., 2020) to the variational bound on the information between representations (Poole et al., 2019), by using the von Mises-Fisher distribution as the conditional variational family. Based on our analysis in Section 4.1, we can identify two main differences between SimCLR and VICReg: (i) **The conditional distribution** $p\boldsymbol{z}|\boldsymbol{x}'$. SimCLR assumes a von Mises-Fisher distribution for the encoder, while VICReg assumes a Gaussian distribution. (ii) (Entropy estimation). The entropy term in SimCLR, $H(Z) = \int p(z|x')p(x')dx'$, is approximate based on the finite sum of the samples. VICReg, however, uses a different approach and estimates the entropy of Z only from the first second moments. Creating self-supervised methods that combine these two differences would be an interesting future research direction. In theory, none of these assumptions is more valid than the other, and it depends on the specific task and our computational constraints.

4.3. Empirical Evaluation

The next step is to verify the validity of our assumptions. Based on the theory presented in Section 3.3, the conditional output density $p_{\boldsymbol{z}|T=i}$ reduces to a single Gaussian with decreasing input noise. We validated it using a ResNet-18 model trained with SimCLR or VICReg on the CIFAR-10 dataset (Krizhevsky, 2009). From the test dataset, we sample 512 Gaussian samples for each image and analyzed whether these samples remain Gaussian (for each image) at the penultimate layer of the DN, that is, before the linear classification layer, independently for each output dimension. Then, we employ D'Agostino and Pearson's test (D'Agostino, 1971) to compute the p-value of the normality test under the null hypothesis that the sample represents a normal distribution. In this test, the deviation from normality is measured, and the test aims to determine whether the sample represents a normally distributed population. A kurtosis and skewness transformation is used to perform the test. The process is repeated for different noise standard deviations. Figure 1 shows the p-value as a function of the normalized standard deviation. We can observe that the network's output is indeed Gaussian with a high probability for small input noise. As we increase the input noise, the network's output becomes less Gaussian until the noise distribution can be rejected with 99% confidence. Moreover, we can see that VICReg is, interestingly, more "Gaussian" than SimCLR, which may have to do with the fact that it optimizes only the second moments of the density distribution to regularize $H(Z)$.

5. Conclusions

In this study, we examine SSL's objective function from an information-theoretic perspective. Our analysis, based on transferring the required stochasticity to the input distribution, shows how to derive SSL objectives. Therefore, it is possible to obtain an information-theoretic analysis even when using deterministic DNs. In the second part, we rediscovered VICReg's loss function from first principles and showed its implicit assumptions. In short, VICReg performs a crude lower bound estimate of the output density entropy by approximating this distribution with a Gaussian matching the first two moments. Finally, we empirically validated that our assumptions are valid in practice, thus confirming the validity of our novel understanding of VICReg. Our work opens many new paths for future research; A better estimation of information-theoretic quantities fits our assumptions. Another exciting research direction is to identify which SSL method is preferred based on data properties.

References

Achille, A. and Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Amjad, R. A. and Geiger, B. C. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Balestriero, R. and Baraniuk, R. A spline theory of deep networks. In *Proc. ICML*, volume 80, pp. 374–383, Jul. 2018.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

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

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Cheney, E. W. and Light, W. A. *A course in approximation theory*, volume 101. American Mathematical Soc., 2009.

D'Agostino, R. B. An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–348, 1971. ISSN 00063444. URL http://www.jstor.org/stable/2334522.

Dubois, Y., Bloem-Reddy, B., Ullrich, K., and Maddison, C. J. Lossy compression for lossless prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

Egerstedt, M. and Martin, C. *Control theoretic splines: optimal control, statistics, and path planning*. Princeton University Press, 2009.

Fantuzzi, C., Simani, S., Beghelli, S., and Rovatti, R. Identification of piecewise affine models in noisy environment. *International Journal of Control*, 75(18):1472–1485, 2002.

Federici, M., Dutta, A., Forré, P., Kushman, N., and Akata, Z. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

Goldfeld, Z., van den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., and Polyanskiy, Y. Estimating Information Flow in Neural Networks. *ArXiv e-prints*, 2018.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*, volume 1. MIT Press, 2016.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Huber, M., Bailey, T., Durrant-Whyte, H., and Hanebeck, U. On entropy approximation for gaussian mixture random vectors. pp. 181 – 188, 09 2008. doi: 10.1109/MFI.2008.4648062.

Kolchinsky, A. and Tracey, B. D. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

Lee, K.-H., Arnab, A., Guadarrama, S., Canny, J., and Fischer, I. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34, 2021b.

Linsker, R. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

Misra, I. and Maaten, L. v. d. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. On the number of linear regions of deep neural networks. In *Proc. NeurIPS*, pp. 2924–2932, 2014.

Moshksar, K. and Khandani, A. K. Arbitrarily tight bounds on differential entropy of gaussian mixtures. *IEEE Transactions on Information Theory*, 62(6):3340–3354, 2016. doi: 10.1109/TIT.2016.2553147.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Piran, Z., Shwartz-Ziv, R., and Tishby, N. The dual information bottleneck. *arXiv preprint arXiv:2006.04641*, 2020.

Poole, B., Ozair, S., Oord, A. v. d., Alemi, A. A., and Tucker, G. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Shi, T., Belkin, M., and Yu, B. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, pp. 3960–3984, 2009.

Shwartz-Ziv, R. Information flow in deep neural networks. *arXiv preprint arXiv:2202.06749*, 2022.

Shwartz-Ziv, R. and Alemi, A. A. Information in infinite ensembles of infinitely-wide neural networks. In *Symposium on Advances in Approximate Bayesian Inference*, pp. 1–17. PMLR, 2020.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Shwartz-Ziv, R., Painsky, A., and Tishby, N. Representation compression and generalization in deep neural networks, 2018.

Steinke, T. and Zakynthinou, L. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pp. 3437–3452. PMLR, 2020.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.