FORMULATING A NON-CONTRASTIVE LOSS AS A DIFFERENCE OF ENTROPIES USING A VON NEUMANN ENTROPY BOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive loss has been successfully exploited in the latest visual unsupervised representation learning methods. Contrastive loss is based on a lower-bound estimation of mutual information where its known limitations include batch size dependency expressed as $O(log(n))$. It is also commonly known as a negative sampling size problem. To cope with the limitation, non-contrastive methods have been proposed and they have been shown to achieve outstanding performance. The non-contrastive methods, however, are limited in that their designs are typically based on heuristics and their learning dynamics can be unstable. In this work, we propose a ~~derive a principled~~ non-contrastive method where the loss design begins from a formulation of mutual information as a difference of entropies such that there is no need for a negative sampling. ~~With our best knowledge, this is the first successful implementation of difference of entropies for visual unsupervised representation learning.~~ Our method performs on par with ~~or better than~~ the state-of-the-art contrastive and non-contrastive methods. The main idea of our approach is to extend Shannon entropy $H(\mathbf{Z})$ to von Neumann entropy ~~$S(\mathbf{Z})$~~. The von Neumann entropy ~~can be~~ has been shown to be a lower bound of Shannon entropy and the corresponding loss enables a stable learning with a small sample size. Additionally, we ~~prove~~ show that the conditional entropy term $H(\mathbf{Z}_1|\mathbf{Z}_2)$ is upper bounded by the negative cosine similarity for the case of weak Gaussian noise augmentation. Even though the derivation is limited to a special case of augmentation, it provides a justification of the commonly used cosine similarity as the measure between positive samples.

Updates in response to multiple reviewers are highlighted in yellow; Reviewer 1 (7tZk); Reviewer 2 (heVY); Reviewer 3 (g1dt); Reviewer 4 (6s5e)

1 INTRODUCTION

Visual unsupervised representation learning focuses on learning useful representations from unannotated visual examples, and recent works have achieved remarkable performance on par with supervised learning (Chen et al., 2020; Grill et al., 2020; Chen & He, 2020; He et al., 2020; Caron et al., 2020; Zbontar et al., 2021). Many of the works are based on contrastive loss (Oord et al., 2018; Henaff et al., 2020), where minimizing contrastive loss is equivalent to maximizing a lower bound estimation of Mutual Information (MI). The bound is known as InfoNCE, and the popular contrastive loss is simply a variational implementation of InfoNCE where positive samples and negative samples are exploited for the MI estimation. It is worthwhile to note that the need for negative samples in contrastive learning is directly related to the need for approximating marginal distributions of mutual information. In this sense, the negative samples may be explained without any relation to noise-contrastive estimation (Gutmann & Hyvärinen, 2010) because they are naturally required for marginal calculation. The use of negative samples, instead of random noise samples, is also more consistent with the view point of marginal approximation.

InfoNCE estimation is known to be vulnerable to two problems. The first problem is the formal limitation articulated in McAllester & Stratos (2020). Specifically, it was shown that any distribution-

free high-confidence lower bound on mutual information estimated from n samples cannot be larger than $O(log(n))$. Often, this limitation is considered to be the reason why many negative samples are needed or why mini-batch size needs to be very large for contrastive learning. The second problem is the high variance of mutual information estimation when a variational estimator, such as InfoNCE estimator, is used. While techniques for trading or reducing the high variance have been proposed (Poole et al., 2019; Song & Ermon, 2019), it is largely an unsolved or perhaps unsolvable problem (McAllester & Stratos, 2020). The exact effect of the high variance to contrastive learning is unclear, but it might be reasonable to assume a negative effect on the learning dynamics. Overall, InfoNCE estimator has turned out to be extremely effective for learning, but finding a better way of estimating mutual information is an important and still open problem.

While many of the visual unsupervised representation learning works are based on the contrastive loss, more recent developments are frequently based on a non-contrastive loss, where the benefits in terms of a superior performance or a simplified training are argued. The non-contrastive methods, however, are typically more challenging to analyze because they tend to be based on heuristics (Grill et al., 2020; Chen & He, 2020), clustering (Caron et al., 2020), neuroscience (Zbontar et al., 2021), or statistical motivation (Bardes et al., 2021). In contrast, contrastive learning is based on a theoretically profound and historically influential concept, mutual information.

In our work, we adhere to the mutual information ~~estimator as the loss~~ as the starting point of loss design, ~~because it is theoretically principled,~~ and develop a non-contrastive ~~mutual information estimator~~ learning method to alleviate the $O(log(n))$ and high variance problems. As shown in Figure 1, we follow the recent self-supervised learning works where the learning is based on an invariant mapping. Our method starts from the mutual information formulation as a Difference of Entropies (DoE) as suggested by McAllester & Stratos (2020).

$$I(\mathbf{Z}_1; \mathbf{Z}_2) = H(\mathbf{Z}_1) - H(\mathbf{Z}_1|\mathbf{Z}_2)$$

Unlike McAllester & Stratos (2020), however, we adopt von Neumann entropy (Nielsen & Chuang, 2002; Wilde, 2013) ~~to estimate~~ as a lower bound of $H(\mathbf{Z}_1)$ and empirically show that it allows a stable learning without a need for a large batch size. For the conditional entropy term $H(\mathbf{Z}_1|\mathbf{Z}_2)$, ~~we adopt cosine similarity~~ its quantum generalization requires an evaluation of the joint system that is computationally disadvantageous. Instead, we recognize that $H(\mathbf{Z}_1|\mathbf{Z}_2)$ can be bounded by the well-known cosine similarity over positive pairs for a special case of augmentation, and apply the bound even for other general augmentations. While not being a rigorous approach, the choice will be empirically shown to provide promising results.

A possible interpretation of von Neumann entropy $S(\mathbf{Z})$ is that it is an extension of Shannon entropy $H(\mathbf{Z})$. This interpretation is based on the fact that $H(\mathbf{Z})$ is a special case of $S(\mathbf{Z})$, and further explanations will be provided in later sections. In fact, $S(\mathbf{Z})$ can be shown to be a lower bound of $H(\mathbf{Z})$ and ~~our derivation is based on the inequality~~ our loss design is based on the bound. While the theory of quantum information theory, the background that is needed for fully understanding von Neumann entropy, is not straightforward, the calculation of $S(\mathbf{Z})$ turns out to be quite simple. For a given mini-batch of n samples, the normalized representations can be stacked into a matrix form as

$$\boldsymbol{U} = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n]^T,$$

where $\boldsymbol{z}_i \in \mathbb{R}^d,\ ||\boldsymbol{z}_i||^2 = 1$, and d is the size of the representation vector. Then, von Neumann entropy is calculated as Shannon entropy over the eigenvalues of $\boldsymbol{U}^T\boldsymbol{U}/n$, where $\boldsymbol{U}^T\boldsymbol{U}/n = 1/n \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T$ and $tr(\boldsymbol{U}^T\boldsymbol{U}/n) = 1$. Note that the n samples are used to estimate each element of the representation covariance matrix $\boldsymbol{U}^T\boldsymbol{U}/n$. Therefore, the estimation of von Neumann entropy does not suffer from $O(log(n))$ problem. We will also empirically show that von Neumann entropy is a numerically stable loss.

Cosine similarity has been a popular loss for positive pairs. In SimCLR (Chen et al., 2020), it was empirically shown that maximizing cosine similarity (i.e. normalized dot product) performs better than maximizing the un-normalized dot product. In SimSiam (Chen & He, 2020), it was empirically shown that maximizing cosine similarity performs better than minimizing cross-entropy. In our work, the choice of cosine similarity as the estimation of $H(\mathbf{Z}_1|\mathbf{Z}_2)$ can be formally justified for the case of weak Gaussian noise as the augmentation where we prove that $H(\mathbf{Z}_1|\mathbf{Z}_2)$ is upper bounded by negative cosine similarity. Even though the proof is for a special case, we adopt the cosine similarity

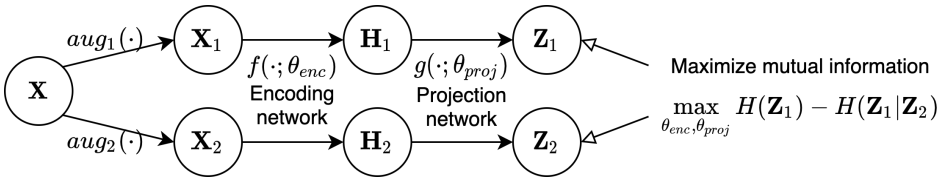


Figure 1: Structure of unsupervised representation learning by maximizing mutual information as a difference of entropies. $\mathbf{H}_1$ and $\mathbf{H}_2$ are the representations used for downstream tasks. $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are normalized representations on a unit hypersphere.

as the ~~estimation~~ surrogate loss of $H(\mathbf{Z}_1|\mathbf{Z}_2)$ because its calculation is quite stable and it has been empirically shown to be effective in the previous works.

Our main contribution is in ~~deriving a principled mutual information estimation method that is not contrastive and utilizes difference of entropies formulation~~ proposing a new non-contrastive loss that can be related to a basic mutual information formulation as a difference of entropies. We also provide empirical results of visual unsupervised representation learning where our proposed method performs on par with the state-of-the-art methods despite the use of a small batch size. In particular, our method is able to achieve the best results so far for COCO transfer learning tasks. ~~The entire formulation of our method is based on pure mutual information estimation, and no other trick is used.~~

2 Related works

Quantum theory in deep representation learning: The quantum theory provides a mathematical framework for representing and manipulating probabilistic distributions of quantum states in a Hilbert space $\mathcal{H}$ (Nielsen & Chuang, 2002; Wilde, 2013). Compared to the representation of deep learning, the quantum state is theoretically well defined and understood. Interestingly, some of the fundamental concepts in quantum theory, such as superposition and entanglement, can be related to popular methods in deep learning. Superposition in quantum theory states that any linear combination of two states forms a superposition state. Some of the deep learning methods (e.g. sum fusion (Feichtenhofer et al., 2016), model superposition (Cheung et al., 2019), and mixup (Zhang et al., 2017)) also utilize linear combination as a basic operator of representations. Entanglement in quantum theory refers to the strong correlation that can be formed between two or more quantum particles. In deep learning, each element of a representation cannot be described separately in general (Szegedy et al., 2013) and only partial disentanglements have been achieved with a variety of training methods (e.g. (Chen et al., 2016; Higgins et al., 2016)). Among the existing studies, some have considered topics relevant to quantum theory. In language modeling, several works have incorporated quantum theory as a general probabilistic framework to model semantic information (Sordoni et al., 2013; Zhang et al., 2018) where the main goal is to address words with multiple meanings using superposition.

Mutual information motivated methods: While measuring mutual information in a high dimensional vector space is a notoriously difficult problem, variational bounds implemented with deep neural network based critics have been studied for estimating and optimizing mutual information (Hjelm et al., 2018; Poole et al., 2019). After a popular Shannon MI estimator InfoNCE was used as a principled loss of unsupervised representation learning by Oord et al. (2018) and Henaff et al. (2020), the later works largely focused on developing variants of the InfoNCE loss so as to improve the downstream task performance (Chen et al., 2020; He et al., 2020).

Non-contrastive methods: While variational mutual information bounds including InfoNCE suffer from the inherent statistical limitations that are addressed in (McAllester & Stratos, 2020), some of the recent non-contrastive methods have successfully resolved such limitations by designing learning methods that do not need negative pairs. A variety of strategies had to be adopted to prevent the representation collapse that results in a constant output to all inputs, and the prevention of such a representation collapse is often considered as the main challenge. In clustering based methods (Caron et al., 2020; Asano et al., 2019), collapse prevention is achieved by balancing the number of elements in each cluster. Other works utilize stop-gradient. The stop-gradient in Siamese network freezes

one side of the network and was empirically shown to be capable of preventing representation collapse (Grill et al., 2020; Chen & He, 2020). Stop-gradient was theoretically studied by Tian et al. (2021). Barlow twins (Zbontar et al., 2021) avoids representation collapse by regularizing the cross-correlation. VICReg (Bardes et al., 2021) adopts variance-invariance-covariance regularization as in an earlier work on covariance and variance regularization (Choi & Rhee, 2019). Besides the existing non-contrastive works, it might to possible to develop new unsupervised learning methods by adopting recently developed mutual information estimation methods such as MIND (Samo, 2021).

3 PRELIMINARIES OF QUANTUM INFORMATION THEORY

While quantum theory encompasses a broad scope of subjects, quantum information theory or quantum Shannon theory is a sub-field that focuses on the quantum equivalent of Shannon information theory (Wilde, 2013). A brief introduction is provided in Appendix A. Among the extensive results, we utilize only the basic concepts of von Neumann entropy, conditional entropy, and mutual information in this work.

While Shannon entropy is calculated for a classical probability distribution, *von Neumann entropy* (also called quantum entropy) is calculated for a density operator ρ (Nielsen & Chuang, 2002), a positive semidefinite hermitian matrix in a Hilbert space $\mathcal{H}$ with the trace value of one. More details can be found in Appendix A. Similar to Shannon information theory, it measures the uncertainty associated with a quantum system.

Definition 3.1 (von Neumann entropy (Nielsen & Chuang, 2002))**.** *The von Neumann entropy (quantum entropy) of a quantum state with density operator ρ is defined as below.*

$$S(\rho) \equiv -tr(\rho \log \rho) = -\sum_x \lambda_x \log \lambda_x, \text{ where } \lambda_x \text{ are the eigenvalues of } \rho.$$

In case some of the singular values are zero, we exclude the corresponding dimensions with $0 \cdot log(0) = 0$. $S(\rho)$ ranges from zero (when ρ is a pure state) to $\log d$ (when ρ is uniformly mixed), where ρ is in a d-dimensional space. The von Neumann entropy agrees with the Shannon entropy if and only if the states $|\psi_i\rangle$ are orthogonal (Nielsen & Chuang, 2002).

When we have two quantum systems A and B, their composite system is called AB. The corresponding density operators are denoted as ρ_A, ρ_B, and ρ_{AB}, respectively. Then, conditional entropy and mutual information are defined as below, where the quantum system name instead of the density operator is used as the argument's notation.

Definition 3.2 (Conditional entropy and mutual information (Nielsen & Chuang, 2002))**.** *The quantum conditional entropy and mutual information of a composite system AB with components A and B are defined as below.*

$$S(A|B) \equiv S(AB) - S(B).$$
$$S(A;B) \equiv S(A) - S(A|B).$$

The value of conditional entropy $S(A|B)$ is reduced as the two systems become entangled. In fact, it can be even negative (Cerf & Adami, 1997). In general, measuring quantum entanglement is computationally challenging (Huang, 2014). The quantum mutual information $S(A;B)$ can be understood in a similar way as in Shannon information theory, and it is a measure of shared information between the two systems.

4 DERIVATION OF ENTROPY BOUND AND CONDITIONAL ENTROPY BOUND

In this section, we provide the derivation of von Neumann entropy lower bound for $H(\mathbf{Z})$ and derive the negative cosine similarity upper bound for $H(\mathbf{Z}_1|\mathbf{Z}_2)$. In Section 5, the two bounds are combined to form a new ~~lower bound of mutual information.~~ non-contrastive loss for unsupervised learning.

4.1 VON NEUMANN ENTROPY $S(\mathbf{Z})$ AS A LOWER BOUND OF SHANNON ENTROPY $H(\mathbf{Z})$

As discussed in Section 3, von Neumann entropy is defined for a quantum state that lies on a unit hypersphere in a Hilbert space $\mathcal{H}$. In our work, the state vector $|\psi\rangle$ corresponds to the representation

vector $\boldsymbol{z}$ (i.e. $|\psi\rangle \equiv \boldsymbol{z}$). Von Neumann entropy of a random variable $\mathbf{Z}$, that is on a unit hypersphere in $\mathbb{R}^d$, has been shown to be a lower bound of Shannon entropy (Nielsen & Chuang, 2002; Wilde, 2013). We merely apply an estimation with samples to obtain the desired result as shown in Theorem 4.1.

Theorem 4.1. *Let $\boldsymbol{z}_i \in \mathbf{Z}$ be a unit column vector with probability $p(i)$. For $\rho \equiv \sum_i p(i)\boldsymbol{z}_i(\boldsymbol{z}_i)^T \approx \sum_{i=1}^n \frac{1}{n}\boldsymbol{z}_i(\boldsymbol{z}_i)^T = \frac{1}{n}\boldsymbol{U}^T\boldsymbol{U} = \tilde{\rho}$, where $\boldsymbol{U} = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n]^T$,*

$$H(\mathbf{Z}) \geq S(\rho) \approx S(\tilde{\rho}) = -\sum_j \lambda_j \log \lambda_j, \text{ where } \lambda_j \text{ denote the eigenvalues of } \boldsymbol{U}^T\boldsymbol{U}/n.$$

Proof. By the quantum data-processing inequality, $H(\mathbf{Z}) \geq S(\rho)$ holds for the Shannon entropy $H(\mathbf{Z})$ and the von Neumann entropy $S(\rho)$ (Nielsen & Chuang, 2002; Wilde, 2013). For sufficiently large n, the density operator ρ can be approximated as the empirical density operator $\tilde{\rho}$ that satisfies $\tilde{\rho} \geq 0$ and $tr(\tilde{\rho}) = 1$. Then, $S(\tilde{\rho})$ can be evaluated with the eigenvalues of $\tilde{\rho}$ following the Definition 3.1. □

With a slight abuse of notation, we use both $S(\mathbf{Z})$ and $S(\rho)$ to refer von Neumann entropy.

4.2 Negative cosine similarity as an upper bound of conditional entropy $H(\mathbf{Z}_1|\mathbf{Z}_2)$

Consider two random variables $\mathbf{Z}_1$ and $\mathbf{Z}_2$, where $\mathbf{Z}_2$ is assumed to be on a unit hypersphere in $\mathbb{R}^d$. If $\mathbf{Z}_1$ is an augmented version of $\mathbf{Z}_2$ with a weak Gaussian perturbation, the two random variables are related as $\mathbf{Z}_1 \sim \mathcal{N}(\mathbf{Z}_2, \boldsymbol{\Sigma})$. For this special augmentation, the main result of Theorem 4.4 can be proved. For the two lemmas, it is assumed that $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix whose eigendecomposition is $\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$, $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_d)$, and $0 < \lambda_d \leq ... \leq \lambda_1 \ll 1$. Also, note that $||\boldsymbol{z}_1||^2 \approx ||\boldsymbol{z}_2||^2 = 1$ and $\boldsymbol{z}_1 \cdot \boldsymbol{z}_2 \approx \frac{\boldsymbol{z}_1^T\boldsymbol{z}_2}{||\boldsymbol{z}_1|| \cdot ||\boldsymbol{z}_2||} = \cos(\boldsymbol{z}_1, \boldsymbol{z}_2) > 0$ due to the weak Gaussian noise perturbation condition.

Lemma 4.2. $-\frac{1}{2}(\boldsymbol{z}_1 - \boldsymbol{z}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_1 - \boldsymbol{z}_2) \geq -1/\lambda_d + \boldsymbol{z}_1 \cdot \boldsymbol{z}_2/\lambda_1$.

Proof. It suffices to prove that $(\boldsymbol{z}_1 - \boldsymbol{z}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_1 - \boldsymbol{z}_2) \leq 2(1/\lambda_d - \boldsymbol{z}_1 \cdot \boldsymbol{z}_2/\lambda_1)$.

$$\begin{aligned}
&(\boldsymbol{z}_1 - \boldsymbol{z}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_1 - \boldsymbol{z}_2) = (\boldsymbol{z}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_1 + \boldsymbol{z}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_2) - (\boldsymbol{z}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_2 + \boldsymbol{z}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_1) \\
&= (\boldsymbol{z}_1^T\boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^T\boldsymbol{z}_1 + \boldsymbol{z}_2^T\boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^T\boldsymbol{z}_2) - (\boldsymbol{z}_1^T\boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^T\boldsymbol{z}_2 + \boldsymbol{z}_2^T\boldsymbol{Q}\boldsymbol{\Lambda}^{-1}\boldsymbol{Q}^T\boldsymbol{z}_1) \\
&\leq 1/\lambda_d(\boldsymbol{z}_1 \cdot \boldsymbol{z}_1 + \boldsymbol{z}_2 \cdot \boldsymbol{z}_2) - 1/\lambda_1(\boldsymbol{z}_1 \cdot \boldsymbol{z}_2 + \boldsymbol{z}_2 \cdot \boldsymbol{z}_1) = 2(1/\lambda_d - \boldsymbol{z}_1 \cdot \boldsymbol{z}_2/\lambda_1)
\end{aligned}$$

□

Lemma 4.3. $-\log p(\boldsymbol{z}_1|\boldsymbol{z}_2) \leq \frac{d}{2}\log 2\pi\lambda_1 + \frac{1}{\lambda_d} - \frac{1}{\lambda_1}\boldsymbol{z}_1 \cdot \boldsymbol{z}_2$.

Proof. It suffices to prove that $p(\boldsymbol{z}_1|\boldsymbol{z}_2) \geq (2\pi\lambda_1)^{-\frac{d}{2}} \exp\{-1/\lambda_d + \boldsymbol{z}_1 \cdot \boldsymbol{z}_2/\lambda_1\}$.

$$\begin{aligned}
p(\boldsymbol{z}_1|\boldsymbol{z}_2) &= \frac{1}{(2\pi)^{d/2}} \frac{1}{\det(\boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{z}_1 - \boldsymbol{z}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_1 - \boldsymbol{z}_2)\right\} \\
&\geq \frac{1}{(2\pi)^{d/2}} \frac{1}{\lambda_1^{d/2}} \exp\{-1/\lambda_d + \boldsymbol{z}_1 \cdot \boldsymbol{z}_2/\lambda_1\}.
\end{aligned}$$

Where the inequality follows from $\det(\boldsymbol{\Sigma}) = \Pi_{i=1}^d \lambda_i \leq \lambda_1^d$ and by Lemma 4.2. □

Theorem 4.4. $H(\mathbf{Z}_1|\mathbf{Z}_2) \leq \frac{d}{2}\log 2\pi\lambda_1 + \frac{1}{\lambda_d} - \frac{1}{\lambda_1}\mathbb{E}[\cos(\mathbf{Z}_1, \mathbf{Z}_2)]$.

Proof.

$$\begin{aligned}
H(\mathbf{Z}_1|\mathbf{Z}_2) &= -\int_{\boldsymbol{z}_1, \boldsymbol{z}_2} p(\boldsymbol{z}_1, \boldsymbol{z}_2) \log p(\boldsymbol{z}_1|\boldsymbol{z}_2) d\boldsymbol{z}_1 d\boldsymbol{z}_2 \\
&\leq \frac{d}{2}\log 2\pi\lambda_1 + \frac{1}{\lambda_d} - \frac{1}{\lambda_1}\int_{\boldsymbol{z}_1, \boldsymbol{z}_2} p(\boldsymbol{z}_1, \boldsymbol{z}_2)(\boldsymbol{z}_1 \cdot \boldsymbol{z}_2) d\boldsymbol{z}_1 d\boldsymbol{z}_2 \text{ (by Lemma 4.3)} \\
&= \frac{d}{2}\log 2\pi\lambda_1 + \frac{1}{\lambda_d} - \frac{1}{\lambda_1}\mathbb{E}[\mathbf{Z}_1 \cdot \mathbf{Z}_2] = \frac{d}{2}\log 2\pi\lambda_1 + \frac{1}{\lambda_d} - \frac{1}{\lambda_1}\mathbb{E}[\cos(\mathbf{Z}_1, \mathbf{Z}_2)].
\end{aligned}$$

□

5 NON-CONTRASTIVE MUTUAL INFORMATION LOSS FOR UNSUPERVISED REPRESENTATION LEARNING

We have derived two bounds in Section 4. In this section, we utilize the two bounds to develop an unsupervised learning method that can be used with the standard diagram shown in Figure 1. ~~As usual, representation learning is accomplished as a byproduct of mutual information maximization. Our method differs from the previous works only in that the mutual information loss is non-contrastive and in that the mutual information loss is implemented as a difference of entropies.~~

5.1 MAXIMIZING ENTROPY $H(\mathbf{Z})$

A direct application of Theorem 4.1 is used, and the following von Neumann entropy is maximized.

$$S(\tilde{\rho_1}) = -\sum_j \lambda_j \log \lambda_j, \ \lambda_j \text{ denote the eigenvalues of } \boldsymbol{U}_1^T \boldsymbol{U}_1/n. \tag{1}$$

From equation 1, we obtain the desired empirical loss $-S(\tilde{\rho_1}) = \sum_j \lambda_j \log \lambda_j$, and it is used to learn $(\boldsymbol{\theta}_{enc}, \boldsymbol{\theta}_{proj})$ that maximizes the entropy $H(\mathbf{Z}_1)$ for the given input data distribution $\mathbf{X}_1$. Computing a more accurate density operator requires a larger batch size of n. However, a large batch size is not really needed as long as the empirical density operator is sufficiently accurate for training the network. Our experiments in Figure 3 show that the default batch size of 128 is sufficient. In fact, even smaller values of 32 and 64 are effective as well.

5.2 MINIMIZING CONDITIONAL ENTROPY $H(\mathbf{Z}_1|\mathbf{Z}_2)$

The Theorem 5.1 was derived for random variables $\mathbf{Z}_1$ and $\mathbf{Z}_2$. For unsupervised representation learning, however, augmentation is applied to input random variables $\mathbf{X}_1$ and $\mathbf{X}_2$. Therefore, we assume the commonly used ReLU as the activation vector and utilize its piecewise linearity to prove the following theorem.

Theorem 5.1. *Let $\mathbf{X}_2$ be an input distribution and $\mathbf{X}_1 (= \{\boldsymbol{x}_1 | \boldsymbol{x}_1 = \boldsymbol{x}_2 + \mathcal{N}(0, c \cdot I), \boldsymbol{x}_2 \in \mathbf{X}_2\})$ be a Gaussian noise perturbation of $\mathbf{X}_2$ with $c \to 0$. For a deep feedforward neural network with piecewise linear activations $f(\cdot)$ whose output vectors are L2-normalized, we have*

$$\exists\, \alpha, \exists\, \beta > 0, \ H(\mathbf{Z}_1|\mathbf{Z}_2) \leq \alpha - \beta\, \mathbb{E}[\cos(\mathbf{Z}_1, \mathbf{Z}_2)],$$

where $\mathbf{Z}_1 = f(\mathbf{X}_1)$, $\mathbf{Z}_2 = f(\mathbf{X}_2)$.

Proof. A deep feedforward neural network with piecewise linear activations is piecewise linear (Montúfar et al., 2014). Thus, for each $\boldsymbol{x}_2 \in \mathbf{X}_2$, $\exists$ there exists $\epsilon > 0$, $\boldsymbol{A}$, and $\boldsymbol{b}$ such that $f(\cdot)$ is equivalent to a linear transformation $l(\boldsymbol{x}) : \boldsymbol{x} \to \boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ on $B_\epsilon(x_2) = \{\boldsymbol{x}_1 \in \mathbf{X}_1 | \ ||\boldsymbol{x}_1 - \boldsymbol{x}_2|| < \epsilon\}$. Hence we have $\boldsymbol{z}_1 \sim \mathcal{N}(\boldsymbol{z}_2, \boldsymbol{\Sigma})$ where a symmetric matrix $\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{z}_1 - \mathbf{z}_2)(\mathbf{z}_1 - \mathbf{z}_2)^T] = \boldsymbol{A}\mathbb{E}[(\boldsymbol{x}_1 - \boldsymbol{x}_2)(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T]\boldsymbol{A}^T = c\boldsymbol{A}\boldsymbol{A}^T$ and $||\boldsymbol{\Sigma}|| \ll 1$ for $c \to 0$. By Theorem 4.4, $H(\mathbf{Z}_1|\mathbf{Z}_2) \leq \frac{d}{2} \log 2\pi\lambda_1 + \frac{1}{\lambda_d} - \frac{1}{\lambda_1}\mathbb{E}[\cos(\mathbf{Z}_1, \mathbf{Z}_2)]$ where $0 < \lambda_d \leq ... \leq \lambda_1$ are the eigenvalues of $\boldsymbol{\Sigma}$. □

As explained a few times, Theorem 5.1 holds only for weak Gaussian noise perturbation. Nonetheless, we use the cosine similarity loss for all the experiments.

6 EXPERIMENTS

Main algorithm: Our method is described in Algorithm 1. It is a straightforward implementation ~~of~~ designed by considering mutual information maximization as a difference of entropies, where the von Neumann entropy term is maximized and the conditional entropy term is minimized. The training and evaluation details are described in Appendix C.

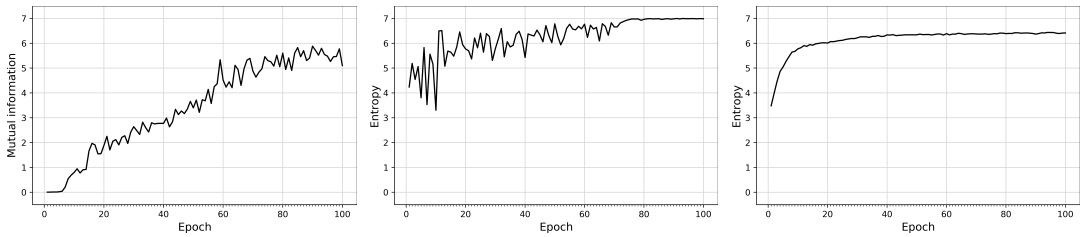
Variance analysis: In Section 1, the high variance problem of variational mutual information estimators was explained. To see if our method suffers less from such a high variance problem, we have investigated three different learning methods. The first is contrastive learning (SimCLR), the second is a difference of entropies implementation where the entropy is calculated as InfoNCE self-information ($I_{NCE}(Z_1; Z_1)$ is estimated as the Shannon entropy) and the conditional entropy term is calculated as the cosine similarity, and the third is a difference of entropies implementation where the entropy is calculated as von Neumann entropy and the conditional entropy term is calculated as the cosine similarity. The results are shown in Figure 2. For the contrastive loss in (a), the mutual information estimate steadily increases as the learning progresses. The estimated mutual information shows a considerable amount of variation. We have used batch size of 128 for the InfoNCE estimation of mutual information ($I_{NCE}(Z_1; Z_2)$). In this case, the $log(n)$ bound is equal to 7 bits, and the mutual information estimate did not hit the 7 bit bound. For the DoE loss with Shannon entropy in (b), the entropy estimation shows a considerable amount of variance. The entropy estimation also hits the 7 bit bound around epoch 80. Actually, the Shannon DoE loss fails to learn effective representations, and the linear evaluation fails with the Top-1 accuracy result of only 26.6%. For our method of DoE loss with von Neumann entropy in (c), the entropy estimation does not show any visible variance problem. The entropy steadily increases as the learning progresses. The von Neumann entropy is not limited by the 7 bit bound, but it happened to stay below 7 bits in our experiment. Similar results for the loss itself can be found in Appendix B.

Sensitivity to batch size: Figure 3 shows the trends of linear evaluation performance of our method for batch sizes between 8 and 128. In most of the cases, the performance improvement saturates around the batch size of 32. Batch size of 128 looks sufficiently large for the approximation of the empirical density operator for training while marginal improvement can be observed for larger batch sizes. Although a sufficiently large batch size is required for batch normalization due to inaccurate batch statistics estimation at small batch sizes (Wu & He, 2018), the von Neumann entropy loss works quite well for the small batch sizes.

Downstream task performance: Table 1a compares linear evaluation Top-1 accuracies between SimCLR, SimSiam, and ours for early epochs, small batch sizes, and a variety of datasets. When reproducing the results of SimCLR and SimSiam, we have followed all the details described in their works. As shown, our method consistently outperforms the other contrastive and non-contrastive methods. Table 1b compares linear evaluation Top-1 accuracies with ImageNet. The results for SimCLR and SimSiam are from (Chen et al., 2020) and (Chen & He, 2020), respectively. Our method ~~certainly~~ outperforms the contrastive method (SimCLR) and shows comparable results for the non-contrastive method (SimSiam). Due to the limited computation resource, we were able to train our model only upto 200 epochs and batch size upto 128. Transfer learning results are shown in Table 2, and we achieve the best performance for the half of the benchmark tasks in there (COCO tasks).

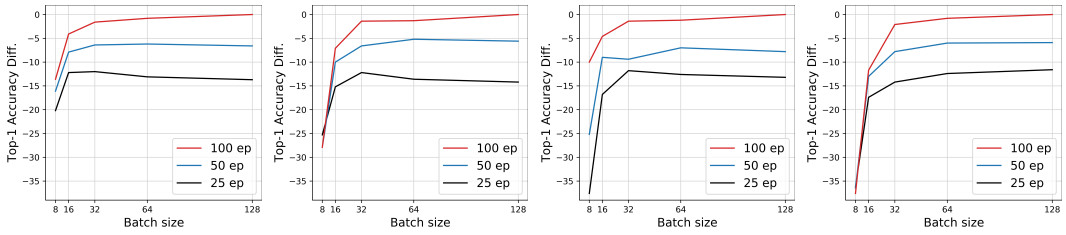
Algorithm 1: Main algorithm, PyTorch-like

Inputs: encoder f, projector g, Aug
Hyperparameters: batch size n, coefficient β
for X **in** loader **do**
  zerograd(f, g)
  $aug_1 \sim$ Aug; $aug_2 \sim$ Aug
  X_1, X_2 = $aug_1(X)$, $aug_2(X)$
  H_1, H_2 = $f(X_1)$, $f(X_2)$
  U_1, U_2 = normalize($g(H_1)$, dim=1), normalize($g(H_2)$, dim=1)
  eig_val = symeig(matmul(U_1.T,U_1)/n)[0][-n:]
  $Loss_1$ = (eig_val*log(eig_val)).sum() # maximization of the von Neumann entropy
  $Loss_2$ = $-$cosine_similarity(U_1, U_2).mean() # minimization of the conditional entropy
  Loss = $Loss_1$ + β*$Loss_2$ # maximization of the mutual information
  Loss.backward()
  update(f, g)
end for



(a) SimCLR (b) DoE (Shannon) (c) DoE (von Neumann)

Figure 2: Comparison of SimCLR (contrastive learning), Shannon DoE (learning with Shannon entropy and cosine similarity), and von Neumann DoE (learning with von Neumann entropy and cosine similarity). As the training is carried out, the loss function's information theoretic part is assessed. (a) For SimCLR, the mutual information is estimated with InfoNCE. (b) For Shannon DoE, the Shannon entropy part is estimated with InfoNCE self-information ($I_{NCE}(Z_1; Z_1)$). (c) For von Neumann DoE, the von Neumann entropy part is directly evaluated. (a), (b), and (c) achieved 85.0%, 26.6%, and 88.3% of Top-1 linear evaluation accuracy for CIFAR-10, respectively.



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet-10 (d) ImageNet-100

Figure 3: Analysis on the effect of batch sizes for empirical density operator for optimization

COCO dataset is more complex than VOC ($\times 4$ more categories, $\times 7$ more training samples, $\times 3$ more sample boxes per images), and it turns out that our loss works better than any other benchmark methods for the complex dataset. This is an evidence that ~~the principled approach can~~ our von Neumann entropy loss has the ability to outperform other existing non-contrastive methods. Considering that we did not try to improve the results at the cost of tuning, we believe our method certainly provides advantages over the existing methods. All the training details are described in Appendix C.

7 Discussion

Beyond mutual information: There are different views on self supervised learning beyond the InfoMax principle (Linsker, 1988). In Shannon information theory, MI is invariant under invertible transformation. Hence, MI alone does not guarantee learning of a useful structure of representation. Tschannen et al. (2019) provides empirical evidences that indicate the success of various self-supervised learning methods can be attributed not only to the maximization of mutual information but also to an effective geometry learned for the representation. Following the work, Wang & Isola (2020) explains contrastive learning as alignment (comparing positive pairs) and uniformity (comparing negative pairs). In our work, the von Neumann entropy takes linear correlations of representation into account. While one can think of its linearity as a limitation, we believe that minimization of negative von Neumann entropy $S(\mathbf{Z})$ and negative cosine similarity effectively guides the geometric characteristics of the learned representations. This can be another benefit of our method when compared to the original contrastive learning.

Limitation of $log(d)$: The von Neumann loss successfully removes InfoNCE's limitation of $O(log(n))$ and enables a learning with a small batch size. Instead, another limitation of $O(log(d))$

Table 1: Linear evaluation Top-1 accuracies (%). All are based on ResNet-50 pre-trained models.

(a) CIFAR-10/100 and ImageNet-10/100

Epoch	Batch size	CIFAR-10 16	32	64	128	CIFAR-100 16	32	64	128	ImageNet-10 16	32	64	128	ImageNet-100 16	32	64	128
25	SimCLR	73.6	73.8	73.5	68.7	43.2	44.4	40.3	39.3	**65.4**	63.4	63.6	65.4	63.1	64.7	66.7	65.9
	Simsiam	49.6	49.9	63.4	63.1	12.8	25.3	27.5	34.1	43.0	49.0	49.2	53.2	37.1	49.5	50.7	56.5
	Ours	**76.1**	**76.3**	**75.2**	**74.6**	**48.1**	**51.1**	**49.7**	**49.1**	**65.4**	**70.4**	**69.6**	**69.0**	**64.6**	**67.8**	**69.6**	**70.4**
50	SimCLR	78.1	79.7	79.0	79.7	47.7	49.5	49.1	50.1	70.8	70.0	70.0	69.0	68.3	71.1	72.1	72.8
	Simsiam	52.4	62.8	74.4	74.2	24.5	33.9	38.2	35.7	52.8	51.4	56.2	56.4	26.9	31.4	44.4	65.0
	Ours	**80.4**	**81.9**	**82.1**	**81.7**	**53.3**	**56.7**	**58.1**	**57.7**	**73.2**	**72.8**	**75.2**	**74.4**	**69.0**	**74.2**	**76.0**	**76.1**
100	SimCLR	81.8	82.8	83.6	85.0	52.2	55.8	57.2	59.3	77.4	77.6	77.4	78.4	68.6	77.6	77.9	78.4
	Simsiam	63.8	69.4	79.3	80.7	30.4	32.8	40.4	47.2	59.8	63.2	62.0	66.0	24.7	29.7	54.5	68.1
	Ours	**84.2**	**87.2**	**88.2**	**88.3**	**56.2**	**61.9**	**62.0**	**63.3**	**77.6**	**80.8**	**81.0**	**82.2**	**70.3**	**79.9**	**81.2**	**82.0**

(b) ImageNet

Epoch		Top-1 acc	Batch size
100	SimCLR	62.8	256
	SimSiam	67.3	128
		68.1	**256**
	Ours	64.4	64
		67.1	128
200	SimCLR	64.3	256
	SimSiam	**70.0**	**256**
	Ours	69.1	128

Table 2: Transfer learning. All are based on ResNet-50 models pre-trained for 200-epoch in ImageNet.

Pretrain	VOC 07 detection AP_{50}	AP	AP_{75}	VOC 07+12 detection AP_{50}	AP	AP_{75}	COCO detection AP_{50}	AP	AP_{75}	COCO instance seg. AP^{mask}_{50}	AP^{mask}	AP^{mask}_{75}
Scratch (repro. in (Chen & He, 2020))	35.9	16.8	13.0	60.2	33.8	33.1	44.0	26.4	27.8	46.9	29.3	30.8
Supervised (repro. in (Chen & He, 2020))	74.4	42.4	42.7	81.3	53.5	58.8	58.2	38.2	41.2	54.7	33.3	35.2
SimCLR (repro. in (Chen & He, 2020))	75.9	46.8	50.1	81.8	55.5	61.4	57.7	37.9	40.9	54.6	33.3	35.3
MoCo v2 (repro. in (Chen & He, 2020))	77.1	48.5	52.5	82.3	57.0	63.3	58.8	39.2	42.5	55.5	34.3	36.6
BYOL (repro. in (Chen & He, 2020))	77.1	47.0	49.9	81.4	55.3	61.1	57.8	37.9	40.9	54.3	33.2	35.0
SwAV (repro. in (Chen & He, 2020))	75.5	46.5	49.6	81.5	55.4	61.4	57.6	37.6	40.3	54.2	33.1	35.1
SimSiam (from (Chen & He, 2020))	**77.3**	**48.5**	**52.5**	**82.4**	**57.0**	**63.7**	59.3	39.2	42.1	56.0	34.4	36.7
Ours	76.4	45.6	47.7	80.9	52.6	57.7	**60.1**	**40.3**	**43.3**	**56.5**	**34.9**	**37.0**

becomes relevant because the von Neumann entropy in Definition 3.1 is limited in that its maximum value is $log(d)$, where d is the size of the representation vector $\mathbf{Z}$. Increasing d, however, does not incur a significant increase in computation and memory requirements when compared to increasing the mini-batch size n. The experiment results also support that our choice of $d = 256$ or $d = 512$ does not prevent a successful learning.

Quantum representation: Consider a coin flip. Shannon entropy is a measure of the expected surprise upon obtaining an observation where the observation can be either a tail or a head, but nothing else. The surprise is larger when a lower probability event occurs. Quantum entropy is also a measure of information but it is more complex in the sense that it is defined over the *state* where a state can be a tail ($[1, 0]^T$ in vector representation), a head ($[0, 1]^T$), or any probabilistic representation of the two (e.g., $[\sqrt{0.7}, \sqrt{0.3}]^T$ which represents a state of 70% chance of tail and 30% chance of head).[1] In quantum theory, a probability distribution itself is considered as a state and therefore quantum theory inherently has an additional level of uncertainty built in its representation and theory. Considering that the representation in deep learning is very difficult to understand analytically and that the ideas of many practical deep learning techniques stem from the concepts of superposition and entanglement, it might be helpful if we can train a neural network to have its representations follow the rules of quantum information theory. Ideally, it would be enlightening to be able to learn such a quantum representation, but our loss is limited in that the tightness of the $H(\mathbf{Z})$ bound is not guaranteed to be achievable and in that the conditional entropy $H(\mathbf{Z}_1|\mathbf{Z}_2)$ is not transformed into a quantum conditional entropy. Furthermore, the tightness of the cosine similarity bound is not guaranteed to be achievable, either. Nonetheless, quantum theory turned out to be quite useful and effective for handling practical problems such as $O(log(n))$ limitation and for improving learning stability in unsupervised representation learning.

[1] In quantum physics, an observation becomes available only after a measurement is made. Quantum information theory is defined over states, not observations, and it can be considered as a generalization of the Shannon information theory.

8 CONCLUSION

We proposed a ~~principled approach that uses difference of entropies as a mutual information bound.~~ non-contrastive loss based on a difference of entropies with a von Neumann entropy bound. We employed a mathematical framework of quantum information theory as an extension of Shannon information theory. The von Neumann entropy provides a tractable lower bound of Shannon entropy in the high dimensional vector space of $\mathbb{R}^d$. Additionally, negative cosine similarity was proven to be an upper bound for the conditional entropy. The proof of cosine similarity maximization is limited because it is derived for weak Gaussian noise augmentation only. By combining the two entropy losses, we have ~~obtained the first successful~~ proposed a non-contrastive ~~mutual information~~ loss for visual unsupervised representation learning. Our method is ~~principled,~~ simple, and it performs well even for a single-device training ~~due to the small~~ because it does not require a large batch size.

NOTE: length of this work will be shortened later by removing strikethrough lines.

Ethics Statement Our work proposes fundamental theories on deep representation learning. Hence, we do not expect any ethical issue. Depending on the applications, however, improved representation learning methods may or may not cause potential ethical problems including inappropriate use of deep fake.

Reproducibility Statement We have included concrete proofs for the theorems in Section 4 and Section 5, PyTorch-like pseudo codes for the algorithms in Section 6, and detailed learning hyperparameters in Appendix C. Fully reproducible code will be made available in github.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33, 2020.

Nicolas J Cerf and Chris Adami. Negative entropy and information in quantum mechanics. *Physical Review Letters*, 79(26):5194, 1997.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

Brian Cheung, Alex Terekhov, Yubei Chen, Pulkit Agrawal, and Bruno Olshausen. Superposition of many models into one. *arXiv preprint arXiv:1902.05522*, 2019.

Daeyoung Choi and Wonjong Rhee. Utilizing class information for deep network representation shaping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3396–3403, 2019.

Paul Adrien Maurice Dirac. A new notation for quantum mechanics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 35, pp. 416–418. Cambridge University Press, 1939.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941, 2016.

Andrew M Gleason. Measures on the closed subspaces of a hilbert space. *Journal of mathematics and mechanics*, pp. 885–893, 1957.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Olivier Henaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, Ali Eslami, and Aaron Van Den Oord. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Yichen Huang. Computing quantum discord is np-complete. *New journal of physics*, 16(3):033027, 2014.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *arXiv preprint arXiv:1402.1869*, 2014.

Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Yves-Laurent Kom Samo. Inductive mutual information estimation: A convex maximum-entropy copula approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 2242–2250. PMLR, 2021.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

Alessandro Sordoni, Jian-Yun Nie, and Yoshua Bengio. Modeling term dependencies with quantum language models for ir. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 653–662, 2013.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. *arXiv preprint arXiv:2102.06810*, 2021.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Mark M Wilde. *Quantum information theory*. Cambridge University Press, 2013.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. End-to-end quantum-like language models with application to question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

A A BRIEF INTRODUCTION TO QUANTUM THEORY

A classic bit can be either 0 or 1. In quantum theory (Nielsen & Chuang, 2002; Wilde, 2013), a *qubit* is a quantum extension of the classic bit, and it can be in state $|0\rangle$, state $|1\rangle$, or any linear combination (superposition state) of the two as $|\psi\rangle = a\,|0\rangle + b\,|1\rangle$, where $|a|^2 + |b|^2 = 1$. Even though deep learning is not directly related to quantum physics, theories based on this extension can be a useful tool for handling deep representations.

Dirac notation and basic concepts Dirac notation is used in quantum theory (Dirac, 1939). ~~It is also called the *bra-ket* notation because of the use of bracket (*bra*-c-*ket*) symbols $<$ and $>$.~~ For a state $|\psi\rangle$, ψ should be understood as the name or label of the state. Because linear algebra provides the mathematical foundation of quantum theory, vector notation is adopted. For instance, in the simple example of $|\psi\rangle = a\,|0\rangle + b\,|1\rangle$, $|\psi\rangle$ can be expressed as $|\psi\rangle = [a, b]^T$ where the interpretation should be state $|\psi\rangle$ can be 0 with probability $|a|^2$ and 1 with probability $|b|^2$ (therefore $|a|^2 + |b|^2 = 1$). Here, the *ket* vector $|\psi\rangle$ is the Dirac notation for a column vector in a Hilbert space $\mathcal{H}$. To represent a row vector, the *bra* vector $\langle\psi|$ is used, as in $\langle\psi| = [a, b]$. An inner product or *braket* is represented as $\langle\psi|\phi\rangle$ and an outer product or *ketbra* is represented as $|\psi\rangle\langle\phi|$.

A state can be either *pure* or *mixed*. In the simple example, $|0\rangle = [1, 0]^T$ and $|1\rangle = [0, 1]^T$ form the *computational basis states*, and they are pure states. Any superposition of the two, $|\psi\rangle = a\,|0\rangle + b\,|1\rangle$, is also a pure state because it corresponds to a single vector with a probabilistic distribution over the basis states. By contrast, a mixed state is a probabilistic mixture of a set of pure states. Note that a pure state already has a probabilistic interpretation over the basis states and a mixed state has an additional level of probabilistic interpretation over a set of such pure states. In this case, we are considering a state that is not completely known but is an ensemble of pure states $\{|\psi_i\rangle\}$ with respective probabilities $\{p_i\}$. The full information of a mixed state cannot be represented as a vector, and the notion of the density operator (also called density matrix) is required.

Definition A.1 (Density operator (Nielsen & Chuang, 2002))**.** *A density operator is defined as below.*

$$\rho \equiv \sum_i p_i\,|\psi_i\rangle\langle\psi_i|\,.$$

Density operator ρ satisfies $\rho \geq 0$ and $tr(\rho) = 1$. In addition, $\rho = \rho^2$ and $rank(\rho) = 1$ are satisfied for pure states and $tr(\rho^2) < 1$ is satisfied for mixed states. The density operator provides a convenient way to describe the uncertainty or probability distribution of a quantum system. According to Gleason's theorem (Gleason, 1957), the probability of a state $|\psi_i\rangle$ in the system with ρ is given by $tr(\rho\,|\psi_i\rangle\langle\psi_i|)$.

A *composite quantum state* of n qubits can be represented as a vector of size 2^n (e.g., a single-qubit state is represented as a vector of size two). For example, a quantum state of two separable single-qubit states can be represented as

$$|\psi\rangle \otimes |\phi\rangle = |\psi\rangle\,|\phi\rangle = |\psi\phi\rangle = [a, b]^T \otimes [c, d]^T = [ac, ad, bc, bd]^T$$

in which $|ac|^2, |ad|^2, |bc|^2$, and $|bd|^2$ represent the probability of $|\psi\phi\rangle$ being $|00\rangle, |01\rangle, |10\rangle$, and $|11\rangle$, respectively. In d-dimensional quantum system, a quantum state is on the unit hypersphere in a Hilbert space $\mathcal{H}$. Note that the hypersphere constraint on representation plays an important role for connecting cosine similarity and positive pair loss part of contrastive learning.

An *entangled state* is a state that cannot be represented as a product of two independent states. For example,

$$|\psi\phi\rangle = \frac{1}{\sqrt{2}}\,|00\rangle + \frac{1}{\sqrt{2}}\,|11\rangle = \frac{1}{\sqrt{2}}[1, 0, 0, 1]^T$$

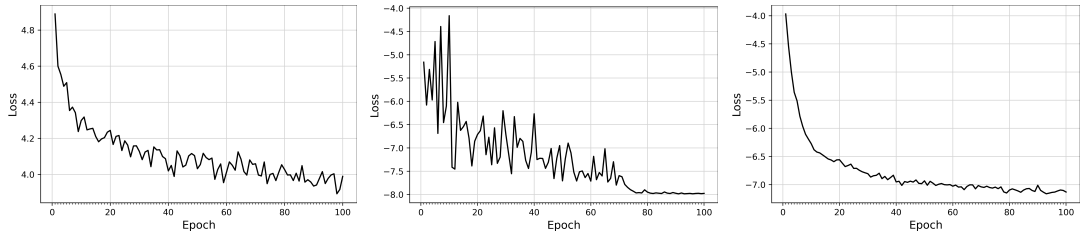
cannot be represented as a product of two single-qubit states; therefore it is an entangled state. In this example, note that $|\psi\rangle$ is always equal to $|\phi\rangle$ (with 50% chance, $|\psi\rangle = |\phi\rangle = |0\rangle$ and with 50% chance, $|\psi\rangle = |\phi\rangle = |1\rangle$). However,

$$|\psi\phi\rangle = \frac{1}{\sqrt{2}}\,|00\rangle + \frac{1}{\sqrt{2}}\,|01\rangle = \frac{1}{\sqrt{2}}[1, 1, 0, 0]^T = |0\rangle \otimes \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$$

is not an entangled state. For an entangled state, each qubit's state cannot be described independently.

B LOSS CURVE COMPARISONS

Similar to Figure 2, we have plotted the progress of loss values for the three learning methods. As in Figure 2, the loss is stably decreased by our method.



(a) SimCLR (b) DoE (Shannon) (c) DoE (von Neumann)

Figure 4: Comparison of SimCLR (contrastive learning), Shannon DoE (learning with Shannon entropy and cosine similarity), and von Neumann DoE (learning with von Neumann entropy and cosine similarity). As the training is carried out, the loss value is plotted for the three types of learning.

C TRAINING AND EVALUATION DETAILS

Encoder pre-training We pre-train the encoding network, ResNet-50 (He et al., 2016), using ImageNet (Russakovsky et al., 2015) training set without labels. The representation is the output of the second last layer (i.e. the global average pooling layer of 2048 dimensions). The projection network has three fully connected linear layers, two hidden layers of dimension 2048 and an output layer of dimension 512, with batch normalization (Ioffe & Szegedy, 2015). The default optimizer is SGD with momentum and global weight decay excluding biases and batch normalization parameters. The learning rate is linearly scaled with batch size (lr = base learning rate $\times$ batch size $/256$) and is scheduled by cosine learning rate decay with 10-epoch warm-up and without restarts (Loshchilov & Hutter, 2016; Goyal et al., 2017). Table 3 summarizes the details of hyperparameters. Here, we have adopted the hyperparameter settings from the previous works. For batch size and learning rate, we have used the same learning rate setting that is used for reproducing 100-epoch BYOL results in (Chen & He, 2020). For the momentum coefficient, we have used 0.9 because it is commonly adopted for imageNet training (Goyal et al., 2017). Because empirical loss for the von Neumann entropy quickly saturates while cosine similarity loss does not, we linearly increased β from 1 to 10 during training. When pre-training with tiny images (CIFAR-10 and CIFAR-100) (Krizhevsky et al., 2009), we made a few modifications to the encoding network and data augmentation, as described in Table 4.

Table 3: Hyperparameters.

	ImageNet				CIFAR-10	
	100 ep		200 ep		100 ep	800 ep
Batch size	64	128	64	128		
Weight decay	5.0e-06	7.5e-06	2.5e-06	5.0e-06	5.0e-05	1.0e-04
Base learning rate	0.45	0.45	0.45	0.45	0.5	0.25
Learning rate	base learning rate $\times$ batch size / 256					
Scheduling	10-epoch warm-up & cosine decay					
Momentum coefficient	0.9					
β	Linear scaling from 1 to 10					

As a key component of the invariant mapping, random data augmentation must inject sufficient randomness while guaranteeing that all distortions from one image share the same semantic content,

Table 4: Hyperparameters for CIFAR-10 and CIFAR-100 (modified from the hyperparameters of ImageNet).

		ImageNet	CIFAR-10,CIFAR-100
Data augmentation *Chenet al.* (2020)	RandomResizedCrop	224x224	32x32
	ColorJitter	p=0.8,b=0.8,c=0.8,s=0.8,h=0.2	p=0.8,b=0.4,c=0.4,s=0.4,h=0.1
	GaussianBlur	p=0.5	p=0
Encoding network *Chenet al.* (2020)	First Layer	7x7 Conv of stride 2	3x3 Conv of stride 1
	Second Layer	Maxpool	Identity
Projection network	# of hidden layers	2	1
	Dimension of output layer	512	256

which is supposed to be retained in the representation. Various combinations of image augmentations in visual self-supervised learning have been studied in (Chen et al., 2020; Grill et al., 2020). Following the works, we use similar data augmentation schemes as summarized in Table 5. Ours follows the same ones as in SimCLR, and it works consistently well with the other data augmentation sets as well.

Table 5: Summary of data augmentation policies (for ImageNet).

		Ours, SimCLR	SimSiam	BYOL
RandomResizedCrop	Probability	1.0	1.0	1.0
RandomHorizontalFlip	Probability	0.5	0.5	0.5
ColorJitter	Probability	0.8	0.8	0.8
	Brightness	0.8	0.4	0.4
	Contrast	0.8	0.4	0.4
	Saturation	0.8	0.4	0.2
	Hue	0.2	0.1	0.1
RandomGrayscale	Probability	0.2	0.2	0.2
GaussianBlur	Probability	0.5	0.5	1.0 / 0.1
	Kernel size	23	23	23
	Sigma range	0.1, 2.0	0.1, 2.0	0.1, 2.0
Solarization	Probability	0.0	0.0	0.0 / 0.2

Because Ours works for small batch sizes, special optimizers such as LARS (You et al., 2017) and multi-GPU data parallelisms for large batch training are not necessary. A single RTX 3090 GPU was used as the default device for pre-training the encoding network as it can handle a batch size of 128 in ImageNet. An 100-epoch pre-training takes a week, and can be accelerated by multiple GPUs.

Linear evaluation A linear classifier probe (Alain & Bengio, 2016) is a general method for measuring the quality of representation by training an independent linear classifier with labels on top of the frozen model's representations. We freeze the pre-trained encoding network and train a linear classifier with the labeled training set using the SGD optimizer with a learning rate of 0.3, weight decay of 1e-6, momentum of 0.995, and batch size of 256. The learning rate is scheduled for 100 epochs by cosine learning rate decay without warm-up and restarts. Performance is measured using the validation set Top-1 accuracy (%).

In Table 1a in Section 6, All the encoding networks are pre-trained in each training set with a base learning rate of 0.25, cosine learning rate decay with 10-epoch warm-up. ImageNet-10 and ImageNet-100 are randomly chosen subsets of 10 classes of ImageNet and 100 classes of ImageNet, respectively. The class names of ImageNet-10/100 are summarized in Table 6.

Transfer learning We also evaluate our method by transferring the model to various tasks. We use the detectron2 (Wu et al., 2019) released under the Apache 2.0 license and follow the MoCo's public codes (He et al., 2020) under the CC-BY-NC 4.0 license, which fine-tunes the pre-trained model to VOC object detection (Everingham et al., 2010) and COCO object detection and instance

Table 6: List of ImageNet-10/100 classes

ImageNet-10									
n02002556	n02168699	n02526121	n02930766	n03814639	n03843555	n04179913	n04591713	n07615774	n07717410
ImageNet-100									
n01484850	n01494475	n01532829	n01560419	n01632458	n01667114	n01689811	n01698640	n01770393	n01796340
n01828970	n01843383	n01855032	n01873310	n01978455	n01981276	n01990800	n02007558	n02086910	n02088238
n02090721	n02093647	n02096051	n02096294	n02098286	n02111500	n02111889	n02117135	n02123597	n02138441
n02167151	n02219486	n02321529	n02363005	n02483708	n02486261	n02492660	n02494079	n02500267	n02783161
n02787622	n02802426	n02814860	n02817516	n02883205	n02906734	n02917067	n02978881	n02992529	n03014705
n03063599	n03127747	n03255030	n03259280	n03344393	n03404251	n03417042	n03478589	n03482405	n03529860
n03642806	n03676483	n03706229	n03761084	n03769881	n03792782	n03803284	n03804744	n03873416	n03982430
n03992509	n04044716	n04070727	n04086273	n04141975	n04146614	n04153751	n04162706	n04179913	n04204238
n04252225	n04254120	n04355933	n04435653	n04476259	n04517823	n04525305	n04584207	n04613696	n07711569
n07714990	n07716906	n07718472	n07718747	n07754684	n09288635	n09472597	n12144580	n12620546	n13054560

segmentation tasks (Lin et al., 2014). VOC 07 detection and VOC 07+12 detection tasks fine-tune Faster R-CNN with C4-backbone (Ren et al., 2015; Wu et al., 2019) in VOC 07 trainval and VOC 07 trainval + VOC 12 train respectively and evaluate in VOC 07 test. COCO detection and instance segmentation tasks fine-tune ($1\times$ schedule) Mask R-CNN with C4-backbone (He et al., 2017; Wu et al., 2019) in COCO 17 train and evaluate in COCO 17 val. In this evaluation, we use $8\times$ RTX 3090 GPUs, the default number of GPUs for the tasks.