Information Flow in Self-Supervised Learning

Zhiquan Tan [1] **Jingqin Yang** [2] **Weiran Huang** [3 4] **Yang Yuan** [2 4 5] **Yifan Zhang** [2]

Abstract

In this paper, we conduct a comprehensive analysis of two dual-branch (Siamese architecture) self-supervised learning approaches, namely Barlow Twins and spectral contrastive learning, through the lens of matrix mutual information. We prove that the loss functions of these methods implicitly optimize both matrix mutual information and matrix joint entropy. This insight prompts us to further explore the category of single-branch algorithms, specifically MAE and U-MAE, for which mutual information and joint entropy become the entropy. Building on this intuition, we introduce the Matrix Variational Masked Auto-Encoder (M-MAE), a novel method that leverages the matrix-based estimation of entropy as a regularizer and subsumes U-MAE as a special case. The empirical evaluations underscore the effectiveness of M-MAE compared with the state-of-the-art methods, including a 3.9% improvement in linear probing ViT-Base, and a 1% improvement in fine-tuning ViT-Large, both on ImageNet.

1. Introduction

Self-supervised learning (SSL) has demonstrated remarkable advancements across various tasks, including image classification and segmentation, often surpassing the performance of supervised learning approaches (Chen et al., 2020; Caron et al., 2021; Li et al., 2021; Zbontar et al., 2021; Bardes et al., 2021). Broadly, SSL methods can be categorized into three types: contrastive learning, feature decorrelation-based learning, and masked image modeling.

One prominent approach in contrastive self-supervised learning is SimCLR (Chen et al., 2020), which employs the InfoNCE loss (Oord et al., 2018) to facilitate the learning process. Interestingly, Oord et al. (2018) show that InfoNCE loss can serve as a surrogate loss for the mutual information between two augmented views. Unlike contrastive learning which needs to include large amounts of negative samples to "contrast", another line of work usually operates without explicitly contrasting with negative samples which are usually called feature decorrelation-based learning (Garrido et al., 2022), e.g., BYOL (Grill et al., 2020), SimSiam (Chen & He, 2021), Barlow Twins (Zbontar et al., 2021), VICReg (Bardes et al., 2021), etc. These methods have garnered attention from researchers seeking to explore alternative avenues for SSL beyond contrastive approaches.

On a different path, the masked autoencoder (MAE) (He et al., 2022) introduces a different way to tackle self-supervised learning. Unlike contrastive and feature decorrelation-based methods that learn useful representations by exploiting the invariance between augmented views, MAE employs a masking strategy to have the model deduce the masked patches from visible patches. Therefore, the representation of MAE carries valuable information for downstream tasks.

At first glance, these three types of self-supervised learning methods may seem distinct, but researchers have made progress in understanding their connections. Garrido et al. (2022) establish a duality between contrastive and feature decorrelation-based methods, shedding light on their fundamental connections and complementarity. Additionally, Balestriero & LeCun (2022) unveil the links between popular feature decorrelation-based SSL methods and dimension reduction methods commonly employed in traditional unsupervised learning. These findings contribute to our understanding of the theoretical underpinnings and potential applications of feature decorrelation-based SSL techniques. However, compared to connections between contrastive and feature decorrelation-based methods, the relationship between MAE and contrastive or feature decorrelation-based methods remains largely unknown. To the best of our knowledge, (Zhang et al., 2022b) is the only paper that relates MAE to the alignment term in contrastive learning.

Though progress has been made in understanding the existing self-supervised learning methods, the tools used

[1]Department of Mathematical Sciences, Tsinghua University, Beijing, China [2]IIIS, Tsinghua University, Beijing, China [3]MIFA Lab, Qing Yuan Research Institute, SEIEE, Shanghai Jiao Tong University, Shanghai, China [4]Shanghai AI Laboratory, Shanghai, China [5]Shanghai Qizhi Institute, Shanghai, China. Correspondence to: Yang Yuan <yuanyang@tsinghua.edu.cn>, Yifan Zhang <zhangyif21@mails.tsinghua.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

in the literature are diverse. As contrastive and feature decorrelation-based learning usually use two augmented views of the same image, one prominent approach is analyzing the mutual information between two views (Oord et al., 2018; Shwartz-Ziv et al., 2023; Shwartz-Ziv & LeCun, 2023). A unified toolbox to understand and improve self-supervised methods is needed. Recently, (Bach, 2022; Skean et al., 2023) have considered generalizing the traditional information-theoretic quantities to the matrix regime. Interestingly, we find these quantities can be powerful tools in understanding and improving existing self-supervised methods regardless of whether they are contrastive, feature decorrelation-based, or masking-based (He et al., 2022).

Taking the matrix information theoretic perspective, we analyze some prominent contrastive and feature decorrelation-based losses and prove that both Barlow Twins and spectral contrastive learning (HaoChen et al., 2021) are maximizing mutual information and joint entropy, see Theorem 4.4 and Theorem 4.14. These claims are crucial for analyzing contrastive and feature decorrelation-based methods, offering a cohesive and elegant understanding. More interestingly, the same analytical framework extends to MAE as well, wherein the concepts of mutual information and joint entropy gracefully degenerate to entropy. Propelled by this observation, we augment the MAE loss with a matrix-based estimation of entropy, giving rise to our new method, Matrix variational Masked Auto-Encoder (M-MAE), which subsumes U-MAE as a special case, see Theorem 5.2.

Empirically, M-MAE stands out with commendable performance. Specifically, it has achieved a 3.9% improvement in linear probing ViT-Base, and a 1% improvement in finetuning ViT-Large, both on ImageNet. This empirical result not only underscores the efficacy of M-MAE but also accentuates the potential of matrix information theory in ushering advancements in self-supervised learning paradigms.

In summary, our contributions can be listed as follows:

- We use matrix information-theoretic tools like matrix mutual information and joint entropies to understand existing contrastive and feature decorrelation-based self-supervised methods.
- We introduce a novel method, M-MAE, which is rooted in matrix information theory, and subsumes U-MAE as a special case.
- Our proposed M-MAE has demonstrated remarkable empirical performance, showcasing a notable improvement in self-supervised learning benchmarks.

2. Related Work

Self-supervised learning. Contrastive and feature decorrelation based methods have emerged as powerful approaches for unsupervised representation learning. By leveraging diverse views or augmentations of input data, they aim to capture meaningful and informative representations that can generalize across different tasks and domains (Chen et al., 2020; Hjelm et al., 2018; Wu et al., 2018; Tian et al., 2019; Chen & He, 2021; Gao et al., 2021; Bachman et al., 2019; Oord et al., 2018; Ye et al., 2019; Henaff, 2020; Misra & Maaten, 2020; Caron et al., 2020; HaoChen et al., 2021; Caron et al., 2021; Li et al., 2021; Zbontar et al., 2021; Tsai et al., 2021; Bardes et al., 2021; Tian et al., 2020; Robinson et al., 2021; Dubois et al., 2022).

Inspired by the widely adopted Masked Language Modeling (MLM) paradigm in NLP, such as BERT (Devlin et al., 2018), Masked Image Modeling (MIM) (Zhang et al., 2022a) has gained attention as a visual representation learning approach. Notably, several MIM methods, including iBOT (Zhou et al., 2021), SimMIM (Xie et al., 2022), and MAE (He et al., 2022), have demonstrated promising results in this domain.

Matrix information theory. Recently, there have been attempts to generalize information theory to measure the relationships between matrices (Bach, 2022; Skean et al., 2023; Zhang et al., 2023a;b). The idea is to apply the traditional information-theoretic quantities on the spectrum of matrices. (Zhang et al., 2023a) discuss the relationship between matrix entropy and effective rank. They also discuss the relationship between matrix KL divergence, total coding rate, and matrix entropy, and propose loss to improve the feature decorrelation-based method. (Liu et al., 2022) use total coding rate to understand the feature decorrelation-based methods.

Theoretical understanding of self-supervised learning. The practical achievements of contrastive learning have ignited a surge of theoretical investigations into the understanding how contrastive loss works (Arora et al., 2019; HaoChen et al., 2021; 2022; Tosh et al., 2020; 2021; Lee et al., 2020; Wang et al., 2022; Nozawa & Sato, 2021; Huang et al., 2021; Tian, 2022; Hu et al., 2022; Tan et al., 2023). (Wang & Isola, 2020) provide an insightful analysis of the optimal solutions of the InfoNCE loss, providing insights into the alignment term and uniformity term that constitute the loss, thus contributing to a deeper understanding of self-supervised learning. (HaoChen et al., 2021; Wang et al., 2022; Tan et al., 2023) explore contrastive self-supervised learning methods from a spectral graph perspective. Several theoretical investigations have delved into the realm of feature decorrelation based methods within the domain of self-supervised learning, as evidenced by a collection of notable studies (Wen & Li, 2022; Tian et al., 2021; Garrido et al., 2022; Balestriero & LeCun, 2022; Tsai et al., 2021; Pokle et al., 2022; Tao et al., 2022; Lee et al., 2021).

Compared to contrastive and feature decorrelation based methods, the theoretical understanding of masked image modeling is still in an early stage. Cao et al. (2022) use the viewpoint of the integral kernel to understand MAE. Zhang et al. (2022b) use the idea of a masked graph to relate MAE with the alignment loss in contrastive learning. Recently, Kong et al. (2023) show MAE effectively detects and identifies a specific group of latent variables using a hierarchical model.

3. Background

3.1. Matrix information-theoretic quantities

In this subsection, we assume **all the mentioned matrices are positive semi-definite** and follow the constraint that all their **diagonal elements are** 1.

We shall first provide the definition of (matrix) entropy as follows:

Definition 3.1 (Matrix-based α-order (Rényi) entropy (Skean et al., 2023))**.** Suppose matrix $\mathbf{K}_1 \in \mathbb{R}^{n\times n}$ and α is a positive real number. The α-order (Rényi) entropy for matrix $\mathbf{K}_1$ is defined as follows:

$$\mathrm{H}_\alpha\left(\mathbf{K}_1\right) = \frac{1}{1-\alpha}\log\left[\mathrm{tr}\left(\left(\frac{1}{n}\mathbf{K}_1\right)^\alpha\right)\right],$$

where $\mathbf{K}_1^\alpha$ is the matrix power.

The case of $\alpha = 1$ is defined as the von Neumann (matrix) entropy (Von Neumann, 2013), i.e.

$$\mathrm{H}_1\left(\mathbf{K}_1\right) = -\,\mathrm{tr}\left(\frac{1}{n}\mathbf{K}_1\log\frac{1}{n}\mathbf{K}_1\right),$$

where log is the matrix logarithm.

Using the definition of matrix entropy, we can define matrix mutual information and joint entropy as follows.

Definition 3.2 (Matrix-based mutual information (Skean et al., 2023))**.** Suppose matrix $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{n\times n}$ and α is a positive real number. The α-order (Rényi) mutual information for matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ is defined as follows:

$$\mathrm{I}_\alpha(\mathbf{K}_1;\mathbf{K}_2) = \mathrm{H}_\alpha(\mathbf{K}_1) + \mathrm{H}_\alpha(\mathbf{K}_2) - \mathrm{H}_\alpha(\mathbf{K}_1\odot\mathbf{K}_2).$$

Definition 3.3 (Matrix-based joint entropy (Skean et al., 2023))**.** Suppose matrix $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{n\times n}$ and α is a positive real number. The α-order (Rényi) joint-entropy for matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ is defined as follows:

$$\mathrm{H}_\alpha(\mathbf{K}_1,\mathbf{K}_2) = \mathbf{H}_\alpha(\mathbf{K}_1\odot\mathbf{K}_2),$$

where $\odot$ is the (matrix) Hadamard product.

3.2. Canonical self-supervised learning losses

We shall recap some canonical losses used in self-supervised learning. As we roughly characterize self-supervised learning into contrastive learning, feature decorrelation-based learning, and masked image modeling. We shall introduce the canonical losses used in these areas sequentially.

In contrastive and feature decorrelation-based learning, people usually adopt the Siamese architecture (dual networks), namely using two parameterized networks: the online network f_θ and the target network f_ϕ. To create different perspectives of a batch of B data points $\{\mathbf{x}_i\}_{i=1}^B$, we randomly select an augmentation $\mathcal{T}$ from a predefined set τ and use it to transform each data point, resulting in new representations $\mathbf{z}_i^{(1)} = f_\theta(\mathcal{T}(\mathbf{x}_i)) \in \mathbb{R}^d$ and $\mathbf{z}_i^{(2)} = f_\phi(\mathbf{x}_i) \in \mathbb{R}^d$ generated by the online and target networks, respectively. We then combine these representations into matrices $\mathbf{Z}_1 = [\mathbf{z}_1^{(1)},\ldots,\mathbf{z}_B^{(1)}]$ and $\mathbf{Z}_2 = [\mathbf{z}_1^{(2)},\ldots,\mathbf{z}_B^{(2)}]$, we assume $\|\mathbf{z}_j^{(k)}\|_2 = 1$ ($k = 1.2$ and $j = 1,\cdots,B$). Denote the (batch normalized) vectors for each dimension i ($1 \le i \le d$) of the online and target networks as $\bar{\mathbf{z}}_i^{(1)}$ and $\bar{\mathbf{z}}_i^{(2)}$, i.e. coordinate-wise

$$\bar{\mathbf{z}}_i^{(k)}(j) = \frac{\mathbf{z}_j^{(k)}(i)}{\sqrt{\sum_{j=1}^B(\mathbf{z}_j^{(k)}(i))^2}}$$

($k = 1.2$ and $i = 1,\cdots,d$, and $j = 1,\cdots,B$). We also define $\bar{\mathbf{Z}}_1 = [\bar{\mathbf{z}}_1^{(1)}\cdots\bar{\mathbf{z}}_d^{(1)}]^\top$ and $\bar{\mathbf{Z}}_2 = [\bar{\mathbf{z}}_1^{(2)}\cdots\bar{\mathbf{z}}_d^{(2)}]^\top$, where $\bar{\mathbf{z}}_i^{(k)} = [\bar{\mathbf{z}}_i^{(k)}(1)\cdots\bar{\mathbf{z}}_i^{(k)}(B)]^\top$ ($k = 1, 2$).

The idea of contrastive learning is to make the representation of similar objects align and dissimilar objects apart. One of the widely adopted losses in contrastive learning is InfoNCE loss (Chen et al., 2020), which is defined as follows:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2}\Big(\sum_{i=1}^B\log\frac{\exp\left((\mathbf{z}_i^{(1)})^\top\mathbf{z}_i^{(2)}\right)}{\sum_{j=1}^B\exp\left((\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(2)}\right)} + \sum_{i=1}^B\log\frac{\exp\left((\mathbf{z}_i^{(2)})^\top\mathbf{z}_i^{(1)}\right)}{\sum_{j=1}^B\exp\left((\mathbf{z}_i^{(2)})^\top\mathbf{z}_j^{(1)}\right)}\Big). \quad (1)$$

As the InfoNCE loss may be difficult to analyze theoretically, HaoChen et al. (2021) then propose spectral contrastive loss as a good surrogate for InfoNCE. The loss is defined as follows:

$$\mathcal{L}_{SC} = \sum_{i=1}^B \|\,\mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)}\,\|_2^2 + \lambda\sum_{i\neq j}((\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(2)})^2, \quad (2)$$

where λ is a hyperparameter. (Here, we slightly generalize the initial loss a bit, the initial paper sets $\lambda = 1$.)

The idea of feature decorrelation-based learning is to learn useful representation by decorrelating features and not explicitly distinguish negative samples. Some notable losses

involve VICReg (Bardes et al., 2021), and Barlow Twins (Zbontar et al., 2021). The Barlow Twins loss is given as follows:

$$\mathcal{L}_{BT} = \sum_{i=1}^{d} (1 - \mathcal{C}_{ii})^2 + \lambda \sum_{i=1}^{d} \sum_{j \neq i} {\mathcal{C}_{ij}}^2, \tag{3}$$

where λ is a hyperparameter and $\mathcal{C}_{ij} = (\bar{\mathbf{z}}_i^{(1)})^\top \bar{\mathbf{z}}_j^{(2)}$ is the cross-correlation coefficient.

The idea of masked image modeling is to learn useful representations by generating the representation from partially visible patches and predicting the rest of the image from the representation, thus useful information in the image remains in the representation. We shall briefly introduce MAE (He et al., 2022) as an example. In masked image modeling, people usually adopt only one branch and do not use Siamese architecture. Given a batch of images $\{\mathbf{x}_i\}_{i=1}^B$, we shall first partition each of the images into n disjoint patches $\mathbf{x}_i = \mathbf{x}_i(j)$ $(1 \leq j \leq n)$. Then B random mask vectors $\mathbf{m}_i \in \{0,1\}^n$ will be generated, and denote the two images generated by these masks as

$$\mathbf{x}_i^{(1)} = \mathbf{x}_i \odot \mathbf{m}_i \quad \text{and} \quad \mathbf{x}_i^{(2)} = \mathbf{x}_i \odot (1 - \mathbf{m}_i). \tag{4}$$

The model consists of two modules: an encoder f and a decoder g. The encoder transform each view $\mathbf{x}_i^{(1)}$ into a representation $\mathbf{z}_i = f(\mathbf{x}_i^{(1)})$. The loss function is $\sum_{i=1}^B \|g(\mathbf{z}_i) - \mathbf{x}^{(2)}\|_2^2$. We also denote the representations in a batch as $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_B]$.

Finally, we will present the U-MAE loss (Zhang et al., 2022b) as:

$$\mathcal{L}_{\text{U-MAE}} = \mathcal{L}_{\text{MAE}} + \gamma \sum_{i \neq j} (\mathbf{z}_i^\top \mathbf{z}_j)^2,$$

where γ is a hyper-parameter.

The goal of this paper is to use a matrix information maximization viewpoint to understand the seemingly different losses in contrastive and feature decorrelation-based methods. We would like also to use matrix information-theoretic tools to improve MAE. We only analyze 4 popular losses: spectral contrastive, Barlow Twins, MAE, and U-MAE. All the proofs can be found in Appendix A. More experiments can be found in Appendix C.

4. Applying matrix information theory to contrastive and feature decorrelation-based methods

As we have discussed in the preliminary session, in contrastive and feature decorrelation-based methods, a common practice is to use two branches (Siamese architecture) namely an online network and a target network to learn useful representations. However, the relationship of the two branches during the training process is mysterious. In this section, we shall use matrix information quantities to unveil the complicated relationship in Siamese architectures.

4.1. Measuring the mutual information

One interesting derivation in (Oord et al., 2018) is that it can be shown that

$$\mathcal{L}_{\text{InfoNCE}} \geq -\operatorname{I}(\mathbf{Z}^{(1)}; \mathbf{Z}^{(2)}) + \log B, \tag{5}$$

where $\mathbf{Z}^{(i)}$ denotes the sampled distribution of the representation.

Though InfoNCE loss is a promising surrogate for estimating the mutual information between the two branches in self-supervised learning. (Sordoni et al., 2021) doubt its effectiveness when facing high-dimensional inputs, where the mutual information may be larger than $\log B$, making the bound vacuous. Then a natural question arises: Can we calculate the mutual information exactly? Unfortunately, it is hard to calculate the mutual information reliably and effectively. Thus we want to see the effect of changing our strategy by using the *matrix* mutual information instead of the traditional one.

As the matrix mutual information has a closed-form expression with only requires a few mild conditions on the input matrices (please refer to section 3.1), one question remains: How to choose the matrices used in the (matrix) mutual information? We find the (batch normalized) sample covariance matrix and batch gram matrix of l_2 normalized representations serve as good candidates. The reason is that by using normalization, the covariance and gram matrices naturally satisfy the requirements that: All the diagonals equal to 1, the matrix is positive semi-definite and it is easy to estimate from data samples.

Notably, the covariance and garm matrix can be seen to have an informal "duality" (Garrido et al., 2022). Specifically, the sample covariance matrix can be expressed as $\mathbf{Z}\mathbf{Z}^\top \in \mathbb{R}^{d \times d}$ and the batch-sample Gram matrix can be expressed as $\mathbf{Z}^\top \mathbf{Z} \in \mathbb{R}^{B \times B}$. The closeness of these two matrices makes us call B and d has duality. As matrix information theory can not only deal with samples from batches but also can exploit the relationship among batches. This makes this theory well-suited for analyzing self-supervised learning methods.

Notably, spectral contrastive loss (Eqn. 2) is a good surrogate loss for InfoNCE loss and calculates the loss involving the batch gram matrix. Another famous loss used in feature decorrelation-based methods is the Barlow Twins (Eqn. 3), which involves the batch-normalized sample covariance matrix. Therefore, these two losses will be our main focus for

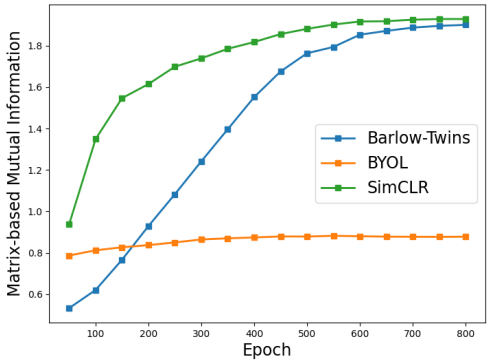


Figure 1. Visualization of matrix-based mutual information on CIFAR10 for Barlow-Twins, BYOL, and SimCLR.

theoretical investigation.

As traditional information theory provides a bound Eqn. (5), thus we are interested in investigating whether there is a matrix information type bound. In the following, we will show that spectral contrastive loss and Barlow Twins loss have (matrix) mutual information bound. Specifically, for ease of theoretical analysis, we first consider setting the α in entropy to be 2.

To prove the bound, we will first present a proposition that relates the mutual information with the Frobenius norm.

Proposition 4.1. $\mathrm{I}_2(\mathbf{K}_1;\mathbf{K}_2) = 2\log d - \log \frac{||\mathbf{K}_1||_F^2||\mathbf{K}_2||_F^2}{||\mathbf{K}_1\odot\mathbf{K}_2||_F^2}$, *where* d *is the size of matrix* $\mathbf{K}_1$.

We will also need a technical proposition that relates the cross-correlation and auto-correlation.

Lemma 4.2. *Suppose* $\mathbf{a}$, $\mathbf{b}$, $\mathbf{a}'$ *and* $\mathbf{b}'$ *are* l_2 *normalized, then* $|\mathbf{a}^\top\mathbf{b}| \leq |\mathbf{a}^\top\mathbf{b}'| + \|\mathbf{b}-\mathbf{b}'\| = |\mathbf{a}^\top\mathbf{b}'| + \sqrt{2(1-\mathbf{b}^\top\mathbf{b}')}$.

Using the previous two propositions, we can derive the following bounds that relate the matrix mutual information and the loss value.

Theorem 4.3. *1. For the spectral contrastive loss, we have*

$$\mathrm{I}_2(\mathbf{Z}_1^\top\mathbf{Z}_1;\mathbf{Z}_2^\top\mathbf{Z}_2) \geq \log B - 2\log(1+(2+\frac{2}{B\lambda})\mathcal{L}_{SC}).$$

2. For the Barlow Twins' loss, we have

$$\mathrm{I}_2(\bar{\mathbf{Z}}_1\bar{\mathbf{Z}}_1^\top;\bar{\mathbf{Z}}_2\bar{\mathbf{Z}}_2^\top) \geq \log d - 2\log(1+\frac{2}{d\lambda}\mathcal{L}_{BT}+4\sqrt{d\mathcal{L}_{BT}}).$$

Proof. We will only present the proof for spectral contrastive loss as Barlow Twins loss is similar.

By Proposition 4.1, we know that $\mathrm{I}_2(\mathbf{Z}_1^\top\mathbf{Z}_1;\mathbf{Z}_2^\top\mathbf{Z}_2) = 2\log B - \log\frac{||\mathbf{Z}_1^\top\mathbf{Z}_1||_F^2||\mathbf{Z}_2^\top\mathbf{Z}_2||_F^2}{||\mathbf{Z}_1^\top\mathbf{Z}_1\odot\mathbf{Z}_2^\top\mathbf{Z}_2||_F^2} \geq 2\log B - \log\frac{||\mathbf{Z}_1^\top\mathbf{Z}_1||_F^2||\mathbf{Z}_2^\top\mathbf{Z}_2||_F^2}{B} = \log B - \log\frac{||\mathbf{Z}_1^\top\mathbf{Z}_1||_F^2}{B}\frac{||\mathbf{Z}_2^\top\mathbf{Z}_2||_F^2}{B}$.

On the other hand, $\frac{||\mathbf{Z}_1^\top\mathbf{Z}_1||_F^2}{B} = 1 + \frac{\sum_{i\neq j}((\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(1)})^2}{B}$.

Using Lemma 4.2, we know that $((\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(1)})^2 \leq (|(\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(2)}| + \|\mathbf{z}_j^{(1)}-\mathbf{z}_j^{(2)}\|)^2 \leq 2(|(\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(2)}|^2 + \|\mathbf{z}_j^{(1)}-\mathbf{z}_j^{(2)}\|^2)$.

Therefore, $\sum_{i\neq j}((\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(1)})^2 \leq 2(\sum_{i\neq j}|(\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(2)}|^2 + \sum_{i\neq j}\|\mathbf{z}_j^{(1)}-\mathbf{z}_j^{(2)}\|^2) < 2(\sum_{i\neq j}|(\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(2)}|^2 + B\sum_j\|\mathbf{z}_j^{(1)}-\mathbf{z}_j^{(2)}\|^2) \leq 2(\frac{1}{\lambda}+B)\mathcal{L}_{SC}$.

Combining all the above, the conclusion follows. □

What about the mutual information when $\alpha \neq 2$. For example $\alpha = 1$? We then plot the mutual information of covariance matrices between branches in Figure 1. We can find out that the mutual information increases during training, which is similar to the case of $\alpha = 2$ proved by Theorem 4.3. More interestingly, the mutual information of SimCLR and Barlow Twins meet at the end of the training, strongly emphasizing the duality of these algorithms. The empirical findings motivate us to consider the case of general $\alpha > 0$.

Unfortunately, it is hard for us to provide bounds similar to Theorem 4.3 for general $\alpha > 0$. But interestingly, we find the following interesting theorem.

Theorem 4.4. *When* $\alpha > 0$*, Barlow Twins and spectral contrastive learning losses maximize the matrix mutual information when their loss value is* 0.

The proof of theorem 4.4 relies on the following upper bound (Proposition 4.5). The key idea is when the loss value is 0, the mutual information can be explicitly calculated and meets the upper bound.

Proposition 4.5. *Suppose* $\mathbf{K}_1$ *and* $\mathbf{K}_2$ *are* $d\times d$ *positive semi-definite matrices with the constraint that each of its diagonals is* 1*. Then* $\mathrm{I}_\alpha(\mathbf{K}_1;\mathbf{K}_2) \leq \log d$.

By combining Theorems 4.3 and 4.4, we can conclude the following corollary.

Corollary 4.6. *When* $\alpha = 2$*, the bounds given by Theorem 4.3 is tight when loss values are* 0.

From the above theorems, we know that when minimizing the spectral contrastive loss and Barlow Twins loss, the mutual information follows a trajectory towards its maximum. This can be seen as mitigating the slight drawback of bound Eqn. (5) in that it only provides an inequality and does not discuss the optimal point.

4.2. Measuring the (joint) entropy

After discussing the application of matrix mutual information in self-supervised learning. We wonder how another import quantity (joint entropy) evolves during the process.

We can show that the matrix joint entropy can indeed reflect the dimensions of representations in Siamese architectures through the following Proposition 4.7.

Proposition 4.7. *The joint entropy lower bounds the representation rank in two branches by having the inequality as follows:*

$$\begin{aligned}&\mathrm{H}_1(\mathbf{K}_1,\mathbf{K}_2)\le\log(\mathrm{rank}(\mathbf{K}_1\odot\mathbf{K}_2))\\&\le\log\mathrm{rank}(\mathbf{K}_1)+\log\mathrm{rank}(\mathbf{K}_2).\end{aligned}$$

$$\begin{aligned}&\max\{\mathrm{H}_\alpha(\mathbf{K}_1),\mathrm{H}_\alpha(\mathbf{K}_2)\}\\&\le\mathrm{H}_\alpha(\mathbf{K}_1,\mathbf{K}_2)\le\mathrm{H}_\alpha(\mathbf{K}_1)+\mathrm{H}_\alpha(\mathbf{K}_2).\end{aligned}$$

This proposition shows that the bigger the joint entropy between the two branches is, the less likely that the representation (rank) collapse. Interestingly, similar results can be proven for (traditional) entropy surrogates (Yu et al., 2020), which we will briefly introduce as follows.

We shall introduce a matrix-based surrogate for entropy as follows:

Definition 4.8. Suppose B samples $\mathbf{Z}=[\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_B]\in\mathbb{R}^{d\times B}$ are i.i.d. samples from a distribution $p(z)$. Then the total coding rate (TCR) (Yu et al., 2020) of $p(z)$ is defined as follows:

$$\mathrm{TCR}_\mu(\mathbf{Z})=\log\det(\mu\mathbf{I}_d+\mathbf{Z}\mathbf{Z}^\top),\tag{6}$$

where μ is a non-negative hyperparameter.

For notation simplicity, we shall also write $\mathrm{TCR}_\mu(\mathbf{Z})$ as $\mathrm{TCR}_\mu(\mathbf{Z}\mathbf{Z}^\top)$. Notably, there is a close relationship between TCR and matrix entropy, which is presented in the following theorem through the lens of matrix KL divergence 4.9. The key is utilizing the asymmetries of the matrix KL divergence (Zhang et al., 2023a).

Definition 4.9 (Matrix KL divergence (Bach, 2022))**.** Suppose matrices $\mathbf{K}_1,\mathbf{K}_2\in\mathbb{R}^{n\times n}$ which $\mathbf{K}_1(i,i)=\mathbf{K}_2(i,i)=1$ for every $i=1,\cdots,n$. Then the Kullback-Leibler (KL) divergence between two positive semi-definite matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ is defined as

$$\mathrm{KL}\left(\mathbf{K}_1\,||\,\mathbf{K}_2\right)=\mathrm{tr}\left[\mathbf{K}_1\left(\log\mathbf{K}_1-\log\mathbf{K}_2\right)\right].$$

Proposition 4.10. *Suppose* $\mathbf{K}$ *is a* $d\times d$ *matrix with the constraint that each of its diagonals is* 1*. Then the following equalities holds:*

$$\begin{aligned}&\mathrm{H}_1(\mathbf{K})=\log d-\frac{1}{d}\mathrm{KL}(\mathbf{K},\mathbf{I}_d),\\&\mathrm{TCR}_\mu(\mathbf{K})=d\log(1+\mu)-\mathrm{KL}(\mathbf{I}_d,\frac{1}{1+\mu}(\mu\mathbf{I}_d+\mathbf{K})).\end{aligned}\tag{7}$$

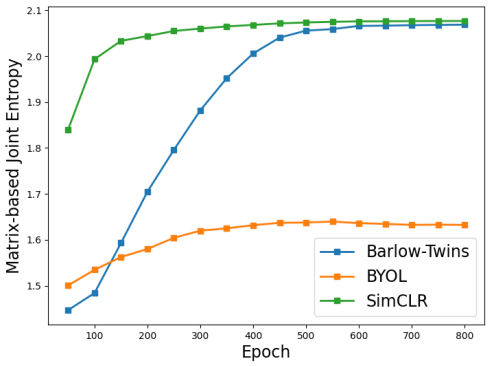


Figure 2. Visualization of matrix-based joint entropy on CIFAR10 for Barlow-Twins, BYOL and SimCLR.

As TCR can be treated as a good surrogate for entropy, we can obtain the following bound for its joint entropy version.

Proposition 4.11. *The (joint) total coding rate upper bounds the rate in two branches by having the inequality as follows:*

$$\mathrm{TCR}_{\mu^2+2\mu}(\mathbf{K}_1\odot\mathbf{K}_2)\ge\mathrm{TCR}_\mu(\mathbf{K}_1)+\mathrm{TCR}_\mu(\mathbf{K}_2).\tag{8}$$

Combining Propositions 4.10, 4.7, and 4.11, it is clear that the bigger the entropy is for each branch the bigger the joint entropy. Thus by combining the conclusion from the above two theorems, it is evident that the joint (matrix or TCR) entropy strongly reflects the extent of collapse during training.

We will then show a bound relating to the matrix joint entropy and the loss values. This remarkable conclusion is proved when the Renyi entropy order $\alpha=2$.

We shall first present a proposition that relates the joint entropy with the Frobenius norm.

Proposition 4.12. *Suppose* $\mathbf{K}_1,\mathbf{K}_2\in\mathbb{R}^{d\times d}$*. Then* $\mathrm{H}_2(\mathbf{K}_1,\mathbf{K}_2)=2\log d-\log\,||\,\mathbf{K}_1\odot\mathbf{K}_2\,||_F^2$*, where* F *is the Frobenius norm.*

We can use Lemma 4.2 and Proposition 4.12 to bind the matrix joint entropy with the loss values.

Theorem 4.13. *1. In the spectral contrastive loss, we have*

$$\mathrm{H}_2(\mathbf{Z}_1^\top\mathbf{Z}_1,\mathbf{Z}_2^\top\mathbf{Z}_2)\ge\log B-\log(1+(2+\frac{2}{B\lambda})\mathcal{L}_{SC}).$$

2. In the Barlow Twins loss, we have

$$\mathrm{H}_2(\bar{\mathbf{Z}}_1\bar{\mathbf{Z}}_1^\top,\bar{\mathbf{Z}}_2\bar{\mathbf{Z}}_2^\top)\ge\log d-\log(1+\frac{2}{d\lambda}\mathcal{L}_{BT}+4\sqrt{d\mathcal{L}_{BT}}).$$

What about the joint entropy when $\alpha=1$ behaves empirically? We shall then plot the joint entropy of covariance

matrices between branches in Figure 2. We can find out that the joint entropy increases during training. More interestingly, the joint entropy of SimCLR and Barlow Twins meet at the end of training, strongly reflects a duality of these algorithms.

Motivated by the above bound, one may wonder what will happen to the joint entropy when the loss is optimized to 0 for general $\alpha > 0$.

Then we can show the following theorem for general $\alpha > 0$.

Theorem 4.14. *When $\alpha > 0$, Barlow Twins and spectral contrastive learning losses maximize the matrix joint entropy when their loss value is 0.*

The key to proving theorem 4.14 lies in the following proposition that finds the optimal point of entropy.

Proposition 4.15.

$$\mathbf{I}_d = \operatorname{argmax} \mathrm{H}_\alpha(\mathbf{K}), \text{and } \mathbf{I}_d = \operatorname{argmax} \mathrm{TCR}_\mu(\mathbf{K}), \quad (9)$$

where the maximization is over $d \times d$ positive semi-definite matrices with the constraint that each of its diagonals is 1.

By combining Theorems 4.13 and 4.14, we can conclude the following corollary.

Corollary 4.16. *When $\alpha = 2$, the bounds given by Theorem 4.13 is tight when loss values are 0.*

Similarly, one may also prove that

Theorem 4.17. $\mathrm{TCR}_\mu(\bar{\mathbf{Z}}_1\bar{\mathbf{Z}}_1^\top \odot \bar{\mathbf{Z}}_2\bar{\mathbf{Z}}_2^\top) \geq d\log(\mu+1) - \frac{1}{2\mu^2}(\frac{2}{\lambda}\mathcal{L}_{BT} + 4(d-1)\sqrt{d\mathcal{L}_{BT}})$.

Remark: The bound given by theorem 4.17 is also tight when the loss value is 0. A similar bound can also be proven batch-wise.

From the above theorems, we know that when minimizing the spectral contrastive loss and Barlow Twins loss, the matrix joint entropy follows a trajectory towards its maximum.

Our theoretical results may show that various contrastive and feature-decorrelation-based methods have similar implicit information maximization processes, thus explaining why they get comparable performance after sufficient training.

5. Applying matrix information theory to masked image modeling

As we have previously discussed the central role of mutual information and joint entropy in the contrastive and feature decorrelation-based methods (Which are all Siamese architecture-based due to the use of augmented views of images). We wonder if can we apply this matrix information theory to improve self-supervised methods using only one network, not the Siamese architecture. To the best of our knowledge, the famous vision self-supervised method using only one network is the masked autoencoder type (MAE) (He et al., 2022).

From a (traditional) information-theoretic point of view, when the two branches merge into one branch the mutual information $\mathrm{I}(\mathbf{X};\mathbf{X})$ and the joint entropy $\mathrm{H}(\mathbf{X},\mathbf{X})$ both equal to the Shannon entropy $\mathrm{H}(\mathbf{X})$. By Propositions 4.10, 4.7, and 4.11, one may see that the joint entropy maximization is closely related to each branch's maximization. Additionally, by the conclusion of Theorems 4.13 and 4.14, one may expect a higher entropy during contrastive and feature decorrelation-based methods. Thus we would like to use the matrix-based entropy in MAE training.

Moreover, matrix entropy can be shown to be very close to a quantity called effective rank. And Zhang et al. (2022b); Garrido et al. (2023) show that the effective rank is a critical quantity for better representation. The definition of effective rank is formally stated in Definition 5.1 and it is easy to show when the matrix is positive semi-definite and has all its diagonal being 1 the effective rank is the exponential of the matrix entropy (Zhang et al., 2023a).

Definition 5.1 (Effective Rank (Roy & Vetterli, 2007))**.** For a non-all-zero matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, the effective rank, denoted as $\mathrm{erank}(\mathbf{A})$, is defined as

$$\mathrm{erank}(\mathbf{A}) \triangleq \exp\left(\mathrm{H}\left(p_1, p_2, \ldots, p_n\right)\right), \quad (10)$$

where $p_i = \frac{\sigma_i}{\sum_{k=1}^n \sigma_k}$, $\{\sigma_i \mid i = 1, \cdots, n\}$ represents the singular values of $\mathbf{A}$, and H denotes the Shannon entropy.

Thus it is natural to add the matrix entropy to the MAE loss to give a new self-supervised learning method. As the numerical instability of calculating matrix entropy is larger than its proxy TCR during training, we shall use the TCR loss (definition 4.8), which is a matrix-based estimator for entropy (Yu et al., 2020).

Recall that we assume each representation z_i is l_2 normalized. If we take the latent distribution of Z as the uniform distribution on the unit hyper-sphere S^{d-1}, we shall get the following loss for self-supervised learning.

$$\mathcal{L}_{\text{M-MAE}} \triangleq \mathcal{L}_{\text{MAE}} - \lambda \cdot \mathrm{TCR}_\mu(\mathbf{Z}), \quad (11)$$

where λ is a loss-balancing hyperparameter.

The name of matrix variational masked auto-encoder (M-MAE) is due to the reason that we can link this new loss to a traditional unsupervised learning method variational auto-encoder (VAE) (Doersch, 2016).

Recall the loss for traditional variational auto-encoder which is given as follows.

$$\mathcal{L}_{\text{VAE}} \triangleq \mathbb{E}_{\mathbf{z}}[-\log q(\mathbf{x}|\mathbf{z}) + \mathrm{KL}(p(\mathbf{z}|\mathbf{x})\|q(\mathbf{z}))],$$
$$\text{where } \mathbf{z} \sim p(\mathbf{z}|\mathbf{x}).$$

The loss contains two terms, the first term $-\log q(\mathbf{x}|\mathbf{z})$ is a reconstruction loss that measures the decoding error. The second term is a discriminative term, which measures the divergence of the encoder distribution $p(\mathbf{z}|\mathbf{x})$ with the latent distribution $q(\mathbf{z})$.

We will first show why MAE loss resembles the first term in VAE, i.e. $\mathbb{E}_{\mathbf{z}}[-\log q(\mathbf{x}|\mathbf{z})]$. In the context of masked image modeling, we usually use MSE loss in place of the log-likelihood. For any input image $\mathbf{x}$, the process of randomly generating a masked vector m and obtaining $\mathbf{z} = f(\mathbf{x} \odot \mathbf{m})$ can be seen as modeling the generating process of $\mathbf{z}|\mathbf{x}$. The decoding process $\mathbf{x}|\mathbf{z}$ can be modeled by concatenating $g(\mathbf{z})$ and $\mathbf{x} \odot \mathbf{m}$ by the (random) position induced by m. Thus the reconstruction loss will be $\|\text{concat}(g(\mathbf{z}), \mathbf{x} \odot \mathbf{m}) - \mathbf{x}\|_2^2 = \|g(\mathbf{z}) - \mathbf{x} \odot (1 - \mathbf{m})\|_2^2$. For a batch of images $\{\mathbf{x}_i\}_{i=1}^B$, this recovers the MAE loss.

We will then show how matrix entropy resembles the second term in VAE, i.e. $\mathbb{E}_{\mathbf{z}}[\text{KL}(p(\mathbf{z}|\mathbf{x})\|q(\mathbf{z}))]$. This is clear by noticing that $q(\mathbf{z})$ is a latent distribution on the unit hyper-sphere S^{d-1} and we naturally choose it as uniform distribution. By taking the covariance matrix of $p(\mathbf{z}|\mathbf{x})$ and $q(\mathbf{z})$ and using the matrix KL divergence (definition 4.9), this term becomes $\text{KL}(\mathbf{Z}\mathbf{Z}^\top \| \mathbf{I}_d)$. By Theorem 4.10, this closely relates to the TCR (and matrix entropy)

Finally, we will present the link of our M-MAE loss to a state-of-the-art one U-MAE (Zhang et al., 2022b).

Theorem 5.2. *U-MAE is a second-order approximation of our proposed M-MAE.*

Proof. The key point is noticing that representations are l_2 normalized and the fact that $\|\mathbf{Z}^\top\mathbf{Z}\|_F^2 = \text{tr}((\mathbf{Z}^\top\mathbf{Z})^2)$.

Using Taylor expansion, we will have:

$$\begin{aligned}
\mathcal{L}_{\text{M-MAE}} &= \mathcal{L}_{\text{MAE}} - \lambda \cdot \log\det(\mathbf{I}_d + \frac{1}{\mu}\mathbf{Z}\mathbf{Z}^\top) + \text{Const.} \\
&= \mathcal{L}_{\text{MAE}} - \lambda \cdot \log\det(\mathbf{I}_B + \frac{1}{\mu}\mathbf{Z}^\top\mathbf{Z}) + \text{Const.} \\
&= \mathcal{L}_{\text{MAE}} - \lambda \cdot \text{tr}\log(\mathbf{I}_B + \frac{1}{\mu}\mathbf{Z}^\top\mathbf{Z}) + \text{Const.} \\
&= \mathcal{L}_{\text{MAE}} - \lambda \cdot \text{tr}(\frac{1}{\mu}\mathbf{Z}^\top\mathbf{Z} - \frac{1}{2\mu^2}(\mathbf{Z}^\top\mathbf{Z})^2 + \cdots) \\
&= \mathcal{L}_{\text{U-MAE}} + \text{Higher-order-terms} + \text{Const.}
\end{aligned}$$

□

Remark: A proof similar to theorem 4.17 will also give a bound that relates M-MAE and U-MAE.

6. Experiments

In this section, we empirically evaluate our Matrix Variational Masked Auto-Encoder (M-MAE) with TCR loss, placing special emphasis on its performance in comparison to the U-MAE model with Square uniformity loss as a baseline. This experiment aims to shed light on the benefits that matrix information-theoretic tools can bring to methods based on masked image modeling.

6.1. Experimental setup

Datasets: ImageNet-1K. We utilize the ImageNet-1K dataset (Deng et al., 2009), which is one of the most comprehensive datasets for image classification. It contains over 1 million images spread across 1000 classes and in self-supervised learning experiments the labels are dropped, providing a robust platform for evaluating our method's generalization capabilities.

Model architectures. We adopt Vision Transformers (ViT) such as ViT-Base and ViT-Large for our models. We closely follow the precedent settings by the U-MAE (Zhang et al., 2022b) paper, as Theorem 5.2 shows the closeness of this method to our M-MAE loss.

Hyperparameters. For a fair comparison, we adopt U-MAE's original hyperparameters: a mask ratio of 0.75 and a uniformity term coefficient λ of 0.01 by default. Both models are pre-trained for 200 epochs on ImageNet-1K with a batch size of 1024, and weight decay is similarly configured as 0.05 to ensure parity in the experimental conditions. For ViT-Base, we set the TCR coefficients $\mu = 1$, and for ViT-Large, we set $\mu = 3$.

6.2. Evaluation results

Evaluation metrics. From Table 1, it's evident that the M-MAE loss outperforms both MAE and U-MAE in terms of linear evaluation and fine-tuning accuracy. Specifically, for ViT-Base, M-MAE achieves a linear probing accuracy of 62.4%, which is a substantial improvement over MAE's 55.4% and U-MAE's 58.5%. Similarly, in the context of ViT-Large, M-MAE achieves an accuracy of 66.0%, again surpassing both MAE and U-MAE. In terms of fine-tuning performance, M-MAE also exhibits superiority, achieving 83.1% and 84.3% accuracy for ViT-Base and ViT-Large respectively. Notably, a 1% increase in accuracy at ViT-Large is very significant. These results empirically validate the theoretical advantages of incorporating matrix information-theoretic tools into masked image modeling, as encapsulated by the TCR loss term in the M-MAE loss function.

7. Conclusion

In conclusion, this study delves into self-supervised learning (SSL), examining contrastive, feature decorrelation-based learning, and masked image modeling through the lens of matrix information theory. Our exploration reveals that many SSL methods are maximizing matrix information-theoretic quantities like matrix mutual information and matrix joint entropy on Siamese architectures.

Table 1. Linear evaluation accuracy (%) and fine-tuning accuracy (%) of pretrained models by MAE loss, U-MAE loss, and M-MAE loss with different ViT backbones on ImageNet-1K. The uniformity regularizer TCR loss in the M-MAE loss significantly improves the linear evaluation performance and fine-tuning performance of the MAE loss.

Downstream Task	Method	ViT-Base	ViT-Large
Linear Probing	MAE	55.4	62.2
	U-MAE	<u>58.5</u>	<u>65.8</u>
	M-MAE	**62.4**	**66.0**
Fine-tuning	MAE	82.9	<u>83.3</u>
	U-MAE	<u>83.0</u>	83.2
	M-MAE	**83.1**	**84.3**

Motivated by the theoretical findings, we also introduce a novel method, the matrix variational masked auto-encoder (M-MAE), enhancing masked image modeling by adding matrix-based estimators for entropy. Empirical results show the effectiveness of the introduced M-MAE loss.

Acknowledgment

Yang Yuan is supported by the Ministry of Science and Technology of the People's Republic of China, the 2030 Innovation Megaprojects "Program on New Generation Artificial Intelligence" (Grant No. 2021AAA0150000).

Weiran Huang is supported by the 2023 CCF-Baidu Open Fund and Microsoft Research Asia.

We would also like to express our sincere gratitude to the reviewers of ICML 2024 for their insightful and constructive feedback. Their valuable comments have greatly contributed to improving the quality of our work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, 2019. 2

Bach, F. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 2022. 2, 6

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019. 2

Balestriero, R. and LeCun, Y. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *arXiv preprint arXiv:2205.11508*, 2022. 1, 2

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 1, 2, 4

Cao, S., Xu, P., and Clifton, D. A. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022. 3

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. 2

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. 1, 2

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. 1, 2, 3

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. 1, 2

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 8

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

Doersch, C. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 7

Dubois, Y., Hashimoto, T., Ermon, S., and Liang, P. Improving self-supervised learning by characterizing idealized representations. *arXiv preprint arXiv:2209.06235*, 2022. 2

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021. 2

Garrido, Q., Chen, Y., Bardes, A., Najman, L., and Lecun, Y. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*, 2022. 1, 2, 4, 15

Garrido, Q., Balestriero, R., Najman, L., and Lecun, Y. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning*, pp. 10929–10974. PMLR, 2023. 7

Giraldo, L. G. S., Rao, M., and Principe, J. C. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014. 13

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 Iformation Processing Systems*, 33:21271–21284, 2020. 1

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021. 2, 3

HaoChen, J. Z., Wei, C., Kumar, A., and Ma, T. Beyond separability: Analyzing the linear transferability of contrastive representations to related subpopulations. *Advances in Neural Information Processing Systems*, 2022. 2

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022. 1, 2, 4, 7

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020. 2

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. In *International Conference on Learning Representations*, 2018. 2

Hoyos-Osorio, J. K. and Sanchez-Giraldo, L. G. The representation jensen-shannon divergence. *arXiv preprint arXiv:2305.16446*, 2023. 17

Hu, T., Liu, Z., Zhou, F., Wang, W., and Huang, W. Your contrastive learning is secretly doing stochastic neighbor embedding. *arXiv preprint arXiv:2205.14814*, 2022. 2

Huang, W., Yi, M., and Zhao, X. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021. 2

Kong, L., Ma, M. Q., Chen, G., Xing, E. P., Chi, Y., Morency, L.-P., and Zhang, K. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7918–7928, 2023. 3

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020. 2

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*, 34:309–323, 2021. 2

Li, Y., Pogodin, R., Sutherland, D. J., and Gretton, A. Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems*, 34:15543–15556, 2021. 1, 2

Liu, X., Wang, Z., Li, Y.-L., and Wang, S. Self-supervised learning via maximum entropy coding. *Advances in Neural Information Processing Systems*, 35:34091–34105, 2022. 2

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. 2

Nozawa, K. and Sato, I. Understanding negative samples in instance discriminative self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:5784–5797, 2021. 2

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2, 4

Pokle, A., Tian, J., Li, Y., and Risteski, A. Contrasting the landscape of contrastive and non-contrastive learning. *arXiv preprint arXiv:2203.15702*, 2022. 2

Robinson, J. D., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. In *ICLR*, 2021. 2

Roy, O. and Vetterli, M. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007. 7

Shwartz-Ziv, R. and LeCun, Y. To compress or not to compress–self-supervised learning and information theory: A review. *arXiv preprint arXiv:2304.09355*, 2023. 2

Shwartz-Ziv, R., Balestriero, R., Kawaguchi, K., Rudner, T. G., and LeCun, Y. An information-theoretic perspective on variance-invariance-covariance regularization. *arXiv preprint arXiv:2303.00633*, 2023. 2

Skean, O., Osorio, J. K. H., Brockmeier, A. J., and Giraldo, L. G. S. Dime: Maximizing mutual information by a difference of matrix-based entropies. *arXiv preprint arXiv:2301.08164*, 2023. 2, 3

Sordoni, A., Dziri, N., Schulz, H., Gordon, G., Bachman, P., and Des Combes, R. T. Decomposed mutual information estimation for contrastive representation learning. In *International Conference on Machine Learning*, pp. 9859–9869. PMLR, 2021. 4

Tan, Z., Zhang, Y., Yang, J., and Yuan, Y. Contrastive learning is spectral clustering on similarity graph. *arXiv preprint arXiv:2303.15103*, 2023. 2

Tao, C., Wang, H., Zhu, X., Dong, J., Song, S., Huang, G., and Dai, J. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14431–14440, 2022. 2

Tian, Y. Deep contrastive learning is provably (almost) principal component analysis. *arXiv preprint arXiv:2201.12680*, 2022. 2

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 2

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020. 2

Tian, Y., Chen, X., and Ganguli, S. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021. 2

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive estimation reveals topic posterior information to linear models. *arXiv:2003.02234*, 2020. 2

Tosh, C., Krishnamurthy, A., and Hsu, D. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021. 2

Tsai, Y.-H. H., Bai, S., Morency, L.-P., and Salakhutdinov, R. A note on connecting barlow twins with negative-sample-free contrastive learning. *arXiv preprint arXiv:2104.13712*, 2021. 2

Von Neumann, J. *Mathematische grundlagen der quantenmechanik*, volume 38. Springer-Verlag, 2013. 3

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020. 2

Wang, Y., Zhang, Q., Wang, Y., Yang, J., and Lin, Z. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv preprint arXiv:2203.13457*, 2022. 2

Wen, Z. and Li, Y. The mechanism of prediction head in non-contrastive self-supervised learning. *arXiv preprint arXiv:2205.06226*, 2022. 2

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018. 2

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022. 2

Ye, M., Zhang, X., Yuen, P. C., and Chang, S.-F. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6210–6219, 2019. 2

Yu, Y., Chan, K. H. R., You, C., Song, C., and Ma, Y. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33:9422–9434, 2020. 6, 7

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021. 1, 2, 4

Zhang, C., Zhang, C., Song, J., Yi, J. S. K., Zhang, K., and Kweon, I. S. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173*, 2022a. 2

Zhang, Q., Wang, Y., and Wang, Y. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35: 27127–27139, 2022b. 1, 3, 4, 7, 8

Zhang, Y., Tan, Z., Yang, J., Huang, W., and Yuan, Y. Matrix information theory for self-supervised learning. *arXiv preprint arXiv:2305.17326*, 2023a. 2, 6, 7, 14

Zhang, Y., Yang, J., Tan, Z., and Yuan, Y. Relationmatch: Matching in-batch relationships for semi-supervised learning. *arXiv preprint arXiv:2305.10397*, 2023b. 2

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2

A. Appendix for proofs

Proposition A.1. $\mathrm{I}_2(\mathbf{K}_1;\mathbf{K}_2) = 2\log d - \log \frac{||\mathbf{K}_1||_F^2||\mathbf{K}_2||_F^2}{||\mathbf{K}_1\odot\mathbf{K}_2||_F^2}$, *where* d *is the size of matrix* $\mathbf{K}_1$.

Proof. The proof is straightforward by using the definition of matrix mutual information when $\alpha = 2$ and the fact that when $\mathbf{K}$ is symmetric $\mathrm{tr}(\mathbf{K}^2) = \mathrm{tr}(\mathbf{K}^\top\mathbf{K}) =|| \mathbf{K} ||_F^2$. □

Lemma A.2. *Suppose* $\mathbf{a}$, $\mathbf{b}$, $\mathbf{a}'$ *and* $\mathbf{b}'$ *are* l_2 *normalized, then* $|\mathbf{a}^\top\mathbf{b}| \le |\mathbf{a}^\top\mathbf{b}'| + \|\mathbf{b}-\mathbf{b}'\| = |\mathbf{a}^\top\mathbf{b}'| + \sqrt{2(1-\mathbf{b}^\top\mathbf{b}')}$.

Proof. Note $|\mathbf{a}^\top\mathbf{b}| = |\mathbf{a}^\top\mathbf{b}' + \mathbf{a}^\top(\mathbf{b}-\mathbf{b}')| \le |\mathbf{a}^\top\mathbf{b}'| + \|\mathbf{a}\|\|\mathbf{b}-\mathbf{b}'\| = |\mathbf{a}^\top\mathbf{b}'| + \sqrt{2(1-\mathbf{b}^\top\mathbf{b}')}$. □

Proposition A.3. *Suppose* $\mathbf{K}_1$ *and* $\mathbf{K}_2$ *are* $d\times d$ *positive semi-definite matrices with the constraint that each of its diagonals is* 1. *Then* $\mathrm{I}_\alpha(\mathbf{K}_1;\mathbf{K}_2) \le \log d$.

Proof. The proof is straightforward by using the inequalities introduced in (Giraldo et al., 2014) as follows. $\mathrm{I}_\alpha(\mathbf{K}_1;\mathbf{K}_2) = \mathrm{H}_\alpha(\mathbf{K}_1) + \mathrm{H}_\alpha(\mathbf{K}_2) - \mathrm{H}_\alpha(\mathbf{K}_1\odot\mathbf{K}_2) \le \mathrm{H}_\alpha(\mathbf{K}_1) \le \log d$. □

Theorem A.4. *When* $\alpha > 0$, *Barlow Twins and spectral contrastive learning losses maximize the matrix mutual information when their loss value is* 0.

Proof. Denote the (batch normalized) vectors for each dimension i ($1 \le i \le d$) of the online and target networks as $\bar{\mathbf{z}}_i^{(1)}$ and $\bar{\mathbf{z}}_i^{(2)}$.

Take $\mathbf{K}_1 = [\bar{\mathbf{z}}_1^{(1)}\cdots\bar{\mathbf{z}}_d^{(1)}]^\top[\bar{\mathbf{z}}_1^{(1)}\cdots\bar{\mathbf{z}}_d^{(1)}]$ and $\mathbf{K}_2 = [\bar{\mathbf{z}}_1^{(2)}\cdots\bar{\mathbf{z}}_d^{(2)}]^\top[\bar{\mathbf{z}}_1^{(2)}\cdots\bar{\mathbf{z}}_d^{(2)}]$.

When the loss value is 0, Barlow Twins loss has $\bar{\mathbf{z}}_i^{(1)} = \bar{\mathbf{z}}_i^{(2)}$ for each $i \in \{1,\cdots,d\}$ and $(\bar{\mathbf{z}}_i^{(1)})^\top\bar{\mathbf{z}}_j^{(2)} = 0$ for each $i \ne j$. Then for each $i \ne j$, $(\bar{\mathbf{z}}_i^{(1)})^\top\bar{\mathbf{z}}_j^{(1)} = (\bar{\mathbf{z}}_i^{(1)})^\top\bar{\mathbf{z}}_j^{(2)} = 0$. Similarly, $(\bar{\mathbf{z}}_i^{(2)})^\top\bar{\mathbf{z}}_j^{(2)} = 0$. Then $\mathbf{K}_1 = \mathbf{K}_2 = \mathbf{I}_d$. By noticing $\mathrm{H}_\alpha(\mathbf{I}_d,\mathbf{I}_d) = \log d$. Then the matrix mutual information is maximized.

When performing spectral contrastive learning, the loss is $\sum_{i=1}^B || \mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)} ||_2^2 + \lambda\sum_{i\ne j}((\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(2)})^2$. Take $\mathbf{K}_1 = \mathbf{Z}_1^\top\mathbf{Z}_1$ and $\mathbf{K}_2 = \mathbf{Z}_2^\top\mathbf{Z}_2$, the results follows similarly. Thus concludes the proof. □

Proposition A.5. *The joint entropy lower bounds the representation rank in two branches by having the inequality as follows:*

$$\mathrm{H}_1(\mathbf{K}_1,\mathbf{K}_2) \le \log(\mathrm{rank}(\mathbf{K}_1\odot\mathbf{K}_2)) \le \log\mathrm{rank}(\mathbf{K}_1) + \log\mathrm{rank}(\mathbf{K}_2).$$

$$\max\{\mathrm{H}_\alpha(\mathbf{K}_1),\mathrm{H}_\alpha(\mathbf{K}_2)\} \le \mathrm{H}_\alpha(\mathbf{K}_1,\mathbf{K}_2) \le \mathrm{H}_\alpha(\mathbf{K}_1) + \mathrm{H}_\alpha(\mathbf{K}_2).$$

Proof. The first inequality comes from the fact that effective rank lower bounds the rank. The second inequality comes from the rank inequality of Hadamard product. The third and fourth inequalities follow from (Giraldo et al., 2014). □

Proposition A.6. *Suppose* $\mathbf{K}$ *is a* $d\times d$ *matrix with the constraint that each of its diagonals is* 1. *Then the following equalities holds:*

$$\begin{aligned}\mathrm{H}_1(\mathbf{K}) &= \log d - \frac{1}{d}\mathrm{KL}(\mathbf{K},\mathbf{I}_d),\\ \mathrm{TCR}_\mu(\mathbf{K}) &= d\log(1+\mu) - \mathrm{KL}(\mathbf{I}_d,\frac{1}{1+\mu}(\mu\mathbf{I}_d+\mathbf{K})).\end{aligned} \tag{12}$$

Proof. The proof is from directly using the definition of matrix KL divergence. □

Proposition A.7. *The (joint) total coding rate upperbounds the rate in two branches by having the inequality as follows:*

$$\mathrm{TCR}_{\mu^2+2\mu}(\mathbf{K}_1\odot\mathbf{K}_2) \ge \mathrm{TCR}_\mu(\mathbf{K}_1) + \mathrm{TCR}_\mu(\mathbf{K}_2). \tag{13}$$

Proof. The inequality comes from the determinant inequality of Hadamard products and the fact that $(\mathbf{K}_1+\mu\mathbf{I})\odot(\mathbf{K}_1+\mu\mathbf{I}) = \mathbf{K}_1\odot\mathbf{K}_2 + (\mu^2+2\mu)\mathbf{I}$. □

Theorem A.8. *1. In the spectral contrastive loss, we have*

$$\mathrm{H}_2(\mathbf{Z}_1^\top \mathbf{Z}_1, \mathbf{Z}_2^\top \mathbf{Z}_2) \geq \log B - \log(1 + (2 + \frac{2}{B\lambda})\mathcal{L}_{SC}).$$

2. In the Barlow Twins loss, we have

$$\mathrm{H}_2(\bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^\top, \bar{\mathbf{Z}}_2 \bar{\mathbf{Z}}_2^\top) \geq \log d - \log(1 + \frac{2}{d\lambda}\mathcal{L}_{BT} + 4\sqrt{d\mathcal{L}_{BT}}).$$

Proof. We will only present the proof for spectral contrastive loss as Barlow Twins loss is similar.

By Proposition 4.12, we know that $\mathrm{H}_2(\mathbf{Z}_1^\top \mathbf{Z}_1, \mathbf{Z}_2^\top \mathbf{Z}_2) = 2\log B - \log ||\mathbf{Z}_1^\top \mathbf{Z}_1 \odot \mathbf{Z}_2^\top \mathbf{Z}_2||_F^2 = \log B - \log \frac{||\mathbf{Z}_1^\top \mathbf{Z}_1 \odot \mathbf{Z}_2^\top \mathbf{Z}_2||_F^2}{B}$.

On the other hand, $\frac{||\mathbf{Z}_1^\top \mathbf{Z}_1 \odot \mathbf{Z}_2^\top \mathbf{Z}_2||_F^2}{B} = 1 + \frac{\sum_{i\neq j}((\mathbf{z}_i^{(1)})^\top \mathbf{z}_j^{(1)})^2 (\mathbf{z}_i^{(2)})^\top \mathbf{z}_j^{(2)})^2}{B}$.

Using Lemma 4.2, we know that $((\mathbf{z}_i^{(1)})^\top \mathbf{z}_j^{(1)})^2 \leq (|(\mathbf{z}_i^{(1)})^\top \mathbf{z}_j^{(2)}| + \|\mathbf{z}_j^{(1)} - \mathbf{z}_j^{(2)}\|)^2 \leq 2(|(\mathbf{z}_i^{(1)})^\top \mathbf{z}_j^{(2)}|^2 + \|\mathbf{z}_j^{(1)} - \mathbf{z}_j^{(2)}\|^2)$.

Therefore, $\sum_{i\neq j}((\mathbf{z}_i^{(1)})^\top \mathbf{z}_j^{(1)})^2 (\mathbf{z}_i^{(2)})^\top \mathbf{z}_j^{(2)})^2 \leq \sum_{i\neq j}((\mathbf{z}_i^{(1)})^\top \mathbf{z}_j^{(1)})^2 \leq 2(\frac{1}{\lambda} + B)\mathcal{L}_{SC}$.

Combining all the above, the conclusion follows.

□

Proposition A.9.

$$\mathbf{I}_d = \operatorname{argmax} \mathrm{H}_\alpha(\mathbf{K}), \text{and } \mathbf{I}_d = \operatorname{argmax} \mathrm{TCR}_\mu(\mathbf{K}), \tag{14}$$

where the maximization is over $d \times d$ positive semi-definite matrices with the constraint that each of its diagonals is 1.

Proof. Specifically, matrix entropy is Shannon entropy on the spectrum and the uniform distribution on the spectrum maximizes the entropy. Consider the spectrum will also give the result for TCR. Another proof directly using matrix KL divergence can be seen in (Zhang et al., 2023a). □

Theorem A.10. *When $\alpha > 0$, Barlow twins and spectral contrastive learning losses maximize the matrix joint entropy when their loss value is* 0.

Proof. Denote the (along batch normalized) vectors for each dimension i ($1 \leq i \leq d$) of the online and target networks as $\bar{\mathbf{z}}_i^{(1)}$ and $\bar{\mathbf{z}}_i^{(2)}$. Take $\mathbf{K}_1 = [\bar{\mathbf{z}}_1^{(1)} \cdots \bar{\mathbf{z}}_d^{(1)}]^\top [\bar{\mathbf{z}}_1^{(1)} \cdots \bar{\mathbf{z}}_d^{(1)}]$ and $\mathbf{K}_2 = [\bar{\mathbf{z}}_1^{(2)} \cdots \bar{\mathbf{z}}_d^{(2)}]^\top [\bar{\mathbf{z}}_1^{(2)} \cdots \bar{\mathbf{z}}_d^{(2)}]$. When the loss value is 0, Barlow Twins loss has $\bar{\mathbf{z}}_i^{(1)} = \bar{\mathbf{z}}_i^{(2)}$ for each $i \in \{1, \cdots, d\}$ and $(\bar{\mathbf{z}}_i^{(1)})^\top \bar{\mathbf{z}}_j^{(2)} = 0$ for each $i \neq j$. Then for each $i \neq j$, $(\bar{\mathbf{z}}_i^{(1)})^\top \bar{\mathbf{z}}_j^{(1)} = (\bar{\mathbf{z}}_i^{(1)})^\top \bar{\mathbf{z}}_j^{(2)} = 0$. Similarly, $(\bar{\mathbf{z}}_i^{(2)})^\top \bar{\mathbf{z}}_j^{(2)} = 0$. Then $\mathbf{K}_1 = \mathbf{K}_2 = \mathbf{I}_d$. By noticing $\mathbf{I}_d \odot \mathbf{I}_d = \mathbf{I}_d$. Then the matrix joint entropy is maximized by noticing Proposition 4.15.

When performing spectral contrastive learning, the loss is $\sum_{i=1}^B || \mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)} ||_2^2 + \lambda \sum_{i\neq j}((\mathbf{z}_i^{(1)})^\top \mathbf{z}_j^{(2)})^2$. Take $\mathbf{K}_1 = \mathbf{Z}_1^\top \mathbf{Z}_1$ and $\mathbf{K}_2 = \mathbf{Z}_2^\top \mathbf{Z}_2$, the results follows similarly. □

Theorem A.11. $\mathrm{TCR}_\mu(\bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^\top \odot \bar{\mathbf{Z}}_2 \bar{\mathbf{Z}}_2^\top) \geq d\log(\mu + 1) - \frac{1}{2\mu^2}(\frac{2}{\lambda}\mathcal{L}_{BT} + 4(d-1)\sqrt{d\mathcal{L}_{BT}})$.

Proof.

Lemma A.12. $\forall\, x, a \geq 0$, *we have* $\log(1+x) \geq \log(1+a) - \frac{1}{2}(x-a)^2 + \frac{1}{1+a}(x-a)$.

Proof. The proof of the lemma is direct from taking the derivative and finding that $x = a$ is the minimal point. □

Denote λ_i as the eigenvalues of $\bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^\top \odot \bar{\mathbf{Z}}_2 \bar{\mathbf{Z}}_2^\top$, we know that $\lambda_i \geq 0$ and $\sum_{i=1}^d \lambda_i = d$ and $\sum_{i=1}^d \lambda_i^2 = \|\bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^\top \odot \bar{\mathbf{Z}}_2 \bar{\mathbf{Z}}_2^\top\|_F^2$. Then take $a = \frac{1}{\mu}$ in the above lemma, we will get the following: $\mathrm{TCR}_\mu(\bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^\top \odot \bar{\mathbf{Z}}_2 \bar{\mathbf{Z}}_2^\top) = \log\det(\mu \mathbf{I}_d + \bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^\top \odot \bar{\mathbf{Z}}_2 \bar{\mathbf{Z}}_2^\top) = \sum_{i=1}^d \log(\mu + \lambda_i) = d\log(\mu) + \sum_{i=1}^d \log(1 + \frac{\lambda_i}{\mu}) \geq d\log(\mu) + \sum_{i=1}^d (\log(1 + \frac{1}{\mu}) + \frac{1}{1+\frac{1}{\mu}}(\frac{\lambda_i}{\mu} - \frac{1}{\mu}) - \frac{1}{2}(\frac{\lambda_i}{\mu} - \frac{1}{\mu})^2) = d\log(\mu+1) + \frac{1}{2\mu^2}d - \frac{1}{2\mu^2}\sum_{i=1}^d \lambda_i^2 = d\log(\mu+1) + \frac{1}{2\mu^2}d - \frac{1}{2\mu^2}\|\bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^\top \odot \bar{\mathbf{Z}}_2 \bar{\mathbf{Z}}_2^\top\|_F^2$. If we denote $\mathbf{K}_1 = \bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^\top$ and

$\mathbf{K}_2 = \bar{\mathbf{Z}}_2\bar{\mathbf{Z}}_2^\top$, then $\mathrm{TCR}_\mu(\mathbf{K}_1 \odot \mathbf{K}_2) \geq d\log(\mu+1) - \frac{1}{2\mu^2}\sum_{i\neq j}\mathbf{K}_1^2(i,j)\mathbf{K}_2^2(i,j) \geq d\log(\mu+1) - \frac{1}{2\mu^2}\sum_{i\neq j}\mathbf{K}_1^2(i,j) \geq d\log(\mu+1) - \frac{1}{2\mu^2}\sum_{i\neq j}2(\mathcal{C}_{i,j}^2 + (2-2\mathcal{C}_{j,j})) \geq d\log(\mu+1) - \frac{1}{2\mu^2}(\frac{2}{\lambda}\mathcal{L}_{BT} + 4(d-1)\sqrt{d\mathcal{L}_{BT}})$. □

Remark: Following the proof of our Theorem 4.6 and Theorem 4.14 and Theorem 4.17, our theoretical results can be generalized to sample contrastive and dimension contrastive methods defined in (Garrido et al., 2022). As pointed out by (Garrido et al., 2022), sample and dimension contrastive methods contain many famous self-supervised methods (Proposition 3.2 of (Garrido et al., 2022)).

Below are other ways of proving the case of $\alpha = 2$.

Theorem A.13. *When $\alpha = 2$, Barlow Twins and spectral contrastive learning losses maximize the matrix mutual information when their loss value is 0.*

We shall first present a lemma as follows:

Lemma A.14. *Given two positive integers n, m. Denote two sequences $\mathbf{x} = (x_1, \cdots, x_m)$ and $\mathbf{y} = (y_1, \cdots, y_m)$. Then $\mathbf{x} = \mathbf{y} = 0$ is the unique solution to the following optimization problem:*

$$\min_{0\leq\mathbf{x}\leq 1, 0\leq\mathbf{y}\leq 1} \frac{(n+\sum_{i=1}^m x_i)(n+\sum_{i=1}^m y_i)}{n+\sum_{i=1}^m x_i y_i}.$$

Proof. Notice that

$$\frac{(n+\sum_{i=1}^m x_i)(n+\sum_{i=1}^m y_i)}{n+\sum_{i=1}^m x_i y_i} - n = \frac{n(\sum_{i=1}^m x_i + \sum_{i=1}^m y_i) - n\sum_{i=1}^m x_i y_i + (\sum_{i=1}^m x_i)(\sum_{i=1}^m y_i)}{n+\sum_{i=1}^m x_i y_i}$$

Note $x_i \geq x_i^2$ and $y_i \geq y_i^2$. Then we shall get inequality as follows:

$$\sum_{i=1}^m x_i + \sum_{i=1}^m y_i \geq 2\sqrt{(\sum_{i=1}^m x_i)(\sum_{i=1}^m y_i)} \geq 2\sqrt{(\sum_{i=1}^m x_i^2)(\sum_{i=1}^m y_i^2)} \geq 2\sum_{i=1}^m x_i y_i.$$

Thus the above optimization problem gets a minimum of n, with $\mathbf{x} = \mathbf{y} = 0$ the unique solution. □

Proof. Denote the (batch normalized) vectors for each dimension i ($1 \leq i \leq d$) of the online and target networks as $\bar{\mathbf{z}}_i^{(1)}$ and $\bar{\mathbf{z}}_i^{(2)}$.

Take $\mathbf{K}_1 = [\bar{\mathbf{z}}_1^{(1)} \cdots \bar{\mathbf{z}}_d^{(1)}]^\top[\bar{\mathbf{z}}_1^{(1)} \cdots \bar{\mathbf{z}}_d^{(1)}]$ and $\mathbf{K}_2 = [\bar{\mathbf{z}}_1^{(2)} \cdots \bar{\mathbf{z}}_d^{(2)}]^\top[\bar{\mathbf{z}}_1^{(2)} \cdots \bar{\mathbf{z}}_d^{(2)}]$.

From Proposition 4.1, it is clear that the mutual information $\mathbf{I}_2(\mathbf{K}_1; \mathbf{K}_2)$ is maximized iff $\frac{||\mathbf{K}_1||_F^2||\mathbf{K}_2||_F^2}{||\mathbf{K}_1\odot\mathbf{K}_2||_F^2}$ is minimized. Take $((\bar{\mathbf{z}}_i^{(1)})^\top\bar{\mathbf{z}}_j^{(1)})^2$ and $((\bar{\mathbf{z}}_i^{(2)})^\top\bar{\mathbf{z}}_j^{(2)})^2$ as elements of $\mathbf{x}$ and $\mathbf{y}$ in Lemma A.14, then we can see the maximal mutual information is attained iff $(\bar{\mathbf{z}}_i^{(1)})^\top\bar{\mathbf{z}}_j^{(1)} = 0$ and $(\bar{\mathbf{z}}_i^{(2)})^\top\bar{\mathbf{z}}_j^{(2)} = 0$.

When the loss value is 0, Barlow Twins loss has $\bar{\mathbf{z}}_i^{(1)} = \bar{\mathbf{z}}_i^{(2)}$ for each $i \in \{1, \cdots, d\}$ and $(\bar{\mathbf{z}}_i^{(1)})^\top\bar{\mathbf{z}}_j^{(2)} = 0$ for each $i \neq j$. Then for each $i \neq j$, $(\bar{\mathbf{z}}_i^{(1)})^\top\bar{\mathbf{z}}_j^{(1)} = (\bar{\mathbf{z}}_i^{(1)})^\top\bar{\mathbf{z}}_j^{(2)} = 0$. Similarly, $(\bar{\mathbf{z}}_i^{(2)})^\top\bar{\mathbf{z}}_j^{(2)} = 0$. Then the matrix mutual information is maximized.

When performing spectral contrastive learning, the loss is $\sum_{i=1}^B || \mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)} ||_2^2 + \lambda\sum_{i\neq j}((\mathbf{z}_i^{(1)})^\top\mathbf{z}_j^{(2)})^2$. Take $\mathbf{K}_1 = \mathbf{Z}_1^\top\mathbf{Z}_1$ and $\mathbf{K}_2 = \mathbf{Z}_2^\top\mathbf{Z}_2$, the results follows similarly. Thus concludes the proof. □

Theorem A.15. *When $\alpha = 2$, Barlow Twins and spectral contrastive learning losses maximize the matrix joint entropy when their loss value is 0.*

Proof. Denote the (along batch normalized) vectors for each dimension i ($1 \leq i \leq d$) of the online and target networks as $\bar{\mathbf{z}}_i^{(1)}$ and $\bar{\mathbf{z}}_i^{(2)}$. Take $\mathbf{K}_1 = [\bar{\mathbf{z}}_1^{(1)} \cdots \bar{\mathbf{z}}_d^{(1)}]^\top[\bar{\mathbf{z}}_1^{(1)} \cdots \bar{\mathbf{z}}_d^{(1)}]$ and $\mathbf{K}_2 = [\bar{\mathbf{z}}_1^{(2)} \cdots \bar{\mathbf{z}}_d^{(2)}]^\top[\bar{\mathbf{z}}_1^{(2)} \cdots \bar{\mathbf{z}}_d^{(2)}]$. From Proposition 4.12, it is clear that the joint entropy $\mathrm{H}_2(\mathbf{K}_1, \mathbf{K}_2)$ is maximized iff $||\ \mathbf{K}_1 \odot \mathbf{K}_2\ ||_F^2$ is minimized. Note from the definition

of Frobenius norm, $|| \mathbf{K}_1 \odot \mathbf{K}_2 ||_F^2 = \sum_{i,j}((\mathbf{K}_1 \odot \mathbf{K}_2)(i,j))^2 = \sum_{i,j}(\mathbf{K}_1(i,j)\mathbf{K}_2(i,j))^2$. When the loss value is 0, Barlow Twins loss has $\bar{\mathbf{z}}_i^{(1)} = \bar{\mathbf{z}}_i^{(2)}$ for each $i \in \{1, \cdots, d\}$ and $(\bar{\mathbf{z}}_i^{(1)})^\top \bar{\mathbf{z}}_j^{(2)} = 0$ for each $i \neq j$. Then for each $i \neq j$, $(\bar{\mathbf{z}}_i^{(1)})^\top \bar{\mathbf{z}}_j^{(1)} = (\bar{\mathbf{z}}_i^{(1)})^\top \bar{\mathbf{z}}_j^{(2)} = 0$. Similarly, $(\bar{\mathbf{z}}_i^{(2)})^\top \bar{\mathbf{z}}_j^{(2)} = 0$. When performing spectral contrastive learning, the loss is $\sum_{i=1}^B || \mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)} ||_2^2 + \lambda \sum_{i \neq j}((\mathbf{z}_i^{(1)})^\top \mathbf{z}_j^{(2)})^2$. Take $\mathbf{K}_1 = \mathbf{Z}_1^\top \mathbf{Z}_1$ and $\mathbf{K}_2 = \mathbf{Z}_2^\top \mathbf{Z}_2$, the results follows similarly. □

B. Ablation study

Table 2. Linear probing accuracy (%) of M-MAE for ViT-Base with varying μ coefficients.

μ Coefficient	0.1	0.5	0.75	1	1.25	1.5	3
Accuracy	58.61	59.38	59.87	**62.40**	59.54	57.76	50.46

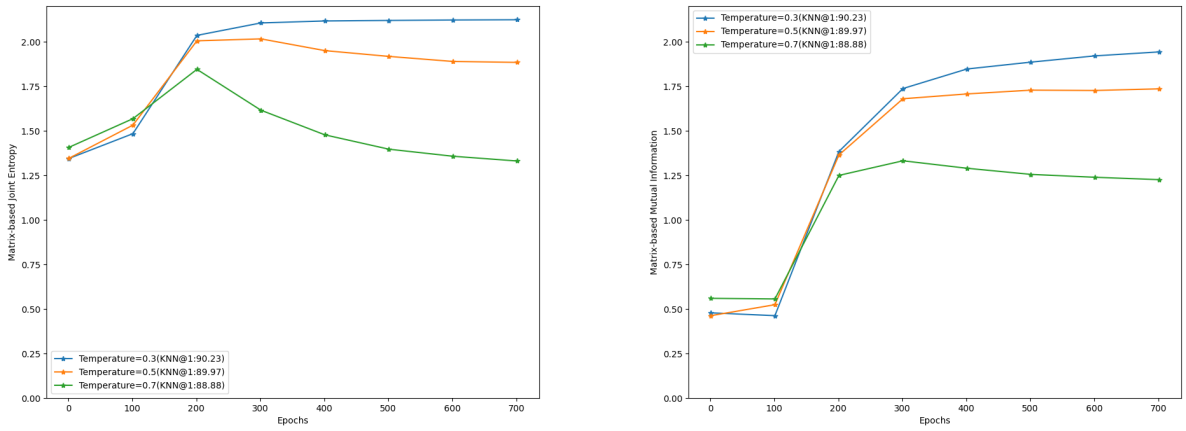
To investigate the robustness of our approach to variations in hyperparameters, we perform an ablation study focusing on the coefficients μ in the TCR loss. The results for different μ values are summarized as in Table 2.

As observed in Table 2, the M-MAE model exhibits a peak performance at $\mu = 1$ for ViT-Base. Deviating from this value leads to a gradual degradation in performance, illustrating the importance of careful hyperparameter tuning for maximizing the benefits of the TCR loss.

C. More experiments

C.1. The tendency under different temperatures

One of the important hyper-parameters in SimCLR is the temperature in InfoNCE loss. We plot the matrix mutual information and matrix joint entropy during the pretraining of SimCLR on CIFAR-10 with different temperatures. We set temperatures as 0.3, 0.5, 0.7. From the Figure 3, we can observe that the increase of matrix mutual information or matrix joint entropy during training ties closely with the final accuracy. As temperature = 0.3 outperforms 0.5 and 0.7 in KNN accuracy, it also has the biggest matrix mutual information and matrix joint entropy value.



(a) Matrix joint entropy.

(b) Matrix mutual information

Figure 3. Tendency of matrix information quantities under different temperatures. The experiments are conducted on CIFAR-10 using SimCLR.

C.2. Longer training on masked modeling

We have conducted experiments on CIFAR-100. The hyper-parameters are similar to U-MAE, and we set $\mu = 1$ and pretrain CIFAR-100 for 1000 epochs. M-MAE may use hyperparameters that are not identical to U-MAE to fully reflect its potential.

However, due to time constraints, we were unable to extensively search for these hyperparameters. We believe that with more reasonable hyperparameters, M-MAE can achieve even better results. As shown in the Table 3, our method performs remarkably well, even without an exhaustive hyperparameter search.

Table 3. Results on CIFAR-100 under various masked modeling pretraining algorithms.

Method	linear probe@1	linear probe@5	finetune@1	finetune@5
M-MAE (vit-base)	**60.9**	**88.1**	83.8	**97.0**
U-MAE (vit-base)	59.7	86.8	84.1	96.5
MAE (vit-base)	59.2	86.8	**84.5**	96.6

We plot the effective rank of learned representations under algorithms MAE, U-MAE, and M-MAE in Figure 4. We find that M-MAE has the biggest effective rank among the algorithms U-MAE has its effective rank bigger than MAE, and the effective ranks of M-MAE have an increasing trend during training. This aligns with our theorem which shows U-MAE can be seen as a second-order approximation of our M-MAE method.

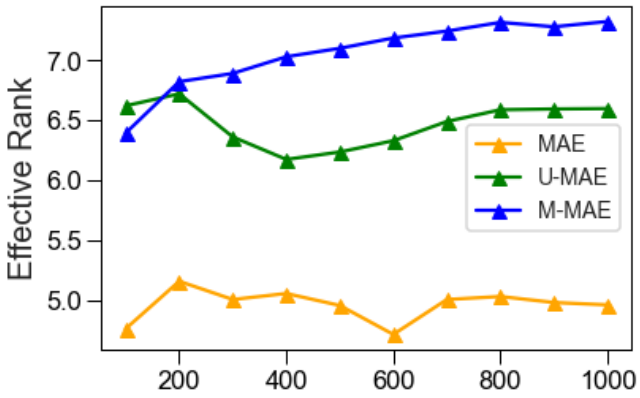


Figure 4. The effective rank during pre-training.

C.3. Measuring the difference between Siamese branches

As we have discussed the total or shared information in the Siamese architectures, we haven't used the matrix information-theoretic tools to analyze the **differences** in the two branches.

From information theory, we know that KL divergence is a special case of f-divergence defined as follows:

Definition C.1. For two probability distributions $\mathbf{P}$ and $\mathbf{Q}$, where $\mathbf{P}$ is absolutely continuous with respect to $\mathbf{Q}$. Suppose $\mathbf{P}$ and $\mathbf{Q}$ has density $p(x)$ and $q(x)$ respectively. Then for a convex function f is defined on non-negative numbers which is right-continuous at 0 and satisfies $f(1) = 0$. The f-divergence is defined as:

$$D_f(\mathbf{P} \,||\, \mathbf{Q}) = \int f\left(\frac{p(x)}{q(x)}\right) q(x)dx. \tag{15}$$

When $f(x) = x \log x$ will recover the KL divergence. Then a natural question arises: are there other f divergences that can be easily generalized to matrices? Note by taking $f(x) = -(x+1)\log\frac{x+1}{2} + x\log x$, we shall retrieve JS divergence. Recently, (Hoyos-Osorio & Sanchez-Giraldo, 2023) generalized JS divergence to the matrix regime.

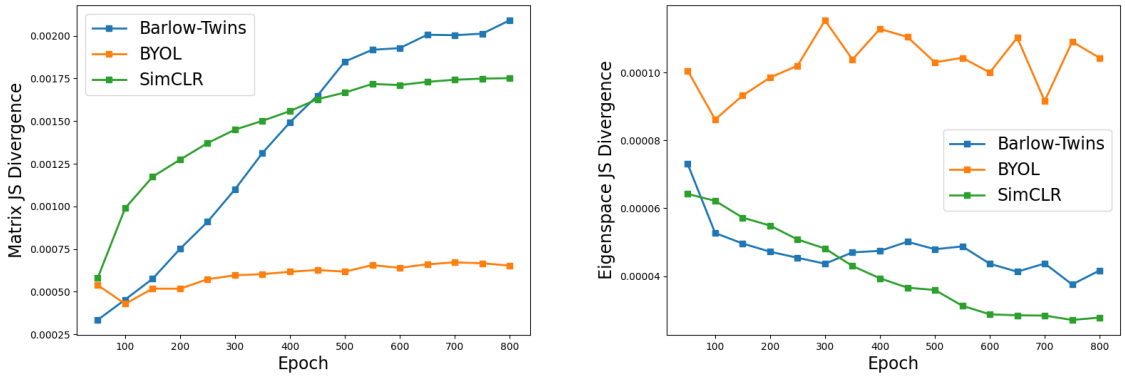
Definition C.2 (Matrix JS divergence (Hoyos-Osorio & Sanchez-Giraldo, 2023))**.** Suppose matrix $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{n\times n}$ which $\mathbf{K}_1(i,i) = \mathbf{K}_2(i,i) = 1$ for every $i = 1, \cdots, n$. The Jensen-Shannon (JS) divergence between these two matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ is defined as

$$\mathrm{JS}\left(\mathbf{K}_1 \,||\, \mathbf{K}_2\right) = \mathbf{H}_1\left(\frac{\mathbf{K}_1 + \mathbf{K}_2}{2}\right) - \frac{\mathbf{H}_1(\mathbf{K}_1) + \mathbf{H}_1(\mathbf{K}_2)}{2}.$$

One may think the matrix KL divergence is a good candidate, but this quantity has some severe problems making it not a good choice. One problem is that the matrix KL divergence is not symmetric. Another problem is that the matrix KL

divergence is not bounded, and sometimes may even be undefined. Recall these drawbacks are similar to that of KL divergence in traditional information theory. In traditional information theory, JS divergence successfully overcomes these drawbacks, thus we may use the matrix JS divergence to measure the differences between branches. As matrix JS divergence considers the interactions between branches, we shall also include the JS divergence between eigenspace distributions as another difference measure.

Specifically, the online and target batch normalized feature correlation matrices can be calculated by $\mathbf{K}_1 = \bar{\mathbf{Z}}_1\bar{\mathbf{Z}}_1^\top$ and $\mathbf{K}_2 = \bar{\mathbf{Z}}_2\bar{\mathbf{Z}}_2^\top$. Denote $\mathbf{p}_1$ and $\mathbf{p}_2$ the online and target (normalized) eigen distribution respectively. We plot the matrix JS divergence $\mathrm{JS}(\mathbf{K_1}, \mathbf{K_2})$ between branches in Figure 5a. It is evident that throughout the whole training, the JS divergence is a small value, indicating a small gap between the branches. More interestingly, the JS divergence increases during training, which means that an effect of "symmetry-breaking" may exist in self-supervised learning. Additionally, we plot the plain JS divergence $\mathrm{JS}(\mathbf{p_1}, \mathbf{p_2})$ between branches in Figure 5b. It is evident that $\mathrm{JS}(\mathbf{p_1}, \mathbf{p_2})$ is very small, even compared to $\mathrm{JS}(\mathbf{K_1}, \mathbf{K_2})$. Thus we hypothesize that the "symmetry-breaking" phenomenon is mainly due to the interactions between Siamese branches.



(a) Matrix JS Divergence.

(b) Eigenspace JS Divergence.

Figure 5. Visualization of matrix JS divergence and eigenspace JS divergence on CIFAR10 for Barlow-Twins, BYOL and SimCLR.