WHAT MAKES FOR GOOD REPRESENTATIONS FOR CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning between different views of the data achieves outstanding success in the field of self-supervised representation learning and the learned representations are useful in various downstream tasks. Since all supervision information for one view comes from the other view, contrastive learning tends to obtain the minimal sufficient representation which contains the shared information and eliminates the non-shared information between views. Considering the diversity of the downstream tasks, it can not be guaranteed that all task-relevant information is shared between views. Therefore, we assume the task-relevant information that is not shared between views can not be ignored and theoretically prove that the minimal sufficient representation in contrastive learning is not sufficient for the downstream tasks, which causes performance degradation. This reveals a new problem that the contrastive learning models have the risk of over-fitting to the shared information between views. To alleviate this problem, we propose to increase the mutual information between the representation and input as regularization to approximately introduce more task-relevant information since we can not utilize any downstream task information during training. Extensive experiments verify the rationality of our analysis and the effectiveness of our method. It significantly improves the performance of several classic contrastive learning models in downstream tasks.

1 INTRODUCTION

Recently, contrastive learning (Chen et al., 2020; Grill et al., 2020a) between different views of the data achieves outstanding success in the field of self-supervised representation learning. The learned representations are broadly useful for various downstream tasks in practice, such as classification and instance segmentation (He et al., 2020). In contrastive learning, the representation that contains all shared information between views is defined as *sufficient representation*, while the representation that contains only the shared and eliminates the non-shared information is defined as *minimal sufficient representation* (Tian et al., 2020b; Tsai et al., 2021). Contrastive learning maximizes the mutual information between the representations of different views, thereby obtaining the sufficient representation. Furthermore, since all supervision information for one view comes from the other view (Federici et al., 2020), the non-shared information is often ignored, so that the minimal sufficient representation is approximately obtained.

Tian et al. (2020b) find that the optimal views for contrastive learning depend on the downstream tasks. In other words, even if the given views are optimal for some downstream tasks, they may not be suitable for other tasks because some task-relevant information is not shared between them. This is intuitive since the downstream tasks are changeable and so as the required information. In this work, we assume that the non-shared task-relevant information can not be ignored, and theoretically prove that the minimal sufficient representation contains less task-relevant information than other sufficient representations and has an non-ignorable gap with the optimal representation, which causes performance degradation. Concretely, we consider two types of the downstream task, i.e., classification and regression tasks, and prove that the lowest achievable error of the minimal sufficient representation is higher than other sufficient representations.

According to our analysis, when some task-relevant information is not shared between views, the learned representations in contrastive learning are not sufficient for downstream tasks. ***We therefore***

find a new problem that the contrastive learning models have the risk of over-fitting to the shared information between views. To this end, we need to introduce more non-shared task-relevant information to the representations. Since we can not utilize any downstream task information when training the contrastive learning models, it is impossible to achieve this directly. As an alternative, ***we propose an objective term which increases the mutual information between the representation and input to approximately introduce more task-relevant information***. We provide extensive empirical experiments to verify the rationality of our analysis and the effectiveness of our method. Concretely, our method can effectively introduce more non-shared task-relevant information and prevent the contrastive learning models from over-fitting to the shared information between views, thereby improve performance. As an exploration, our method can also prevent the models from over-fitting to the label information and get better transfer performance in supervised learning.

2 THEORETICAL ANALYSIS AND MODEL

In this section, we first introduce the contrastive learning framework and theoretically analyze the disadvantages of minimal sufficient representation in contrastive learning, and then propose our method to approximately introduce more task-related information to the representations.

2.1 CONTRASTIVE LEARNING

Contrastive learning (Hjelm et al., 2018; Chen et al., 2020) is a general framework for unsupervised representation learning which maximizes the mutual information between the representations of two random variables $\mathbf{v}_1$ and $\mathbf{v}_2$ with the joint distribution $p(\mathbf{v}_1, \mathbf{v}_2)$

$$\max_{f_1, f_2} I(\mathbf{z}_1, \mathbf{z}_2) \tag{1}$$

where $\mathbf{z}_i = f_i(\mathbf{v}_i), i = 1, 2$ are also random variables and $f_i, i = 1, 2$ are encoding functions. In practice, $\mathbf{v}_1$ and $\mathbf{v}_2$ are two views of the data $\mathbf{x}$, such as local patches and the whole image (Hjelm et al., 2018), different augmentations of the same image (Wu et al., 2018; Bachman et al., 2019; He et al., 2020; Chen et al., 2020), different image channels (Tian et al., 2020a), or video and text pairs (Sun et al., 2019; Miech et al., 2020). When $\mathbf{v}_1$ and $\mathbf{v}_2$ have the same marginal distributions ($p(\mathbf{v}_1) = p(\mathbf{v}_2)$), the function f_1 and f_2 can be the same ($f_1 = f_2$).

In contrastive learning, the variable $\mathbf{v}_2$ provides supervision information for $\mathbf{v}_1$ and plays the same role as the label $\mathbf{y}$ in the supervised learning, and vice versa (Federici et al., 2020). Similar to the information bottleneck theory (Tishby & Zaslavsky, 2015; Achille & Soatto, 2018) in the supervised learning, we can define the sufficient representation and minimal sufficient representation of $\mathbf{v}_1$ (or $\mathbf{v}_2$) for $\mathbf{v}_2$ (or $\mathbf{v}_1$) in contrastive learning (Tian et al., 2020b; Tsai et al., 2021).

Definition 1. *(Sufficient Representation in Contrastive Learning) The representation $\mathbf{z}_1^{suf}$ of $\mathbf{v}_1$ is sufficient for $\mathbf{v}_2$ if and only if $I(\mathbf{z}_1^{suf}, \mathbf{v}_2) = I(\mathbf{v}_1, \mathbf{v}_2)$.*

The sufficient representation $\mathbf{z}_1^{suf}$ of $\mathbf{v}_1$ keeps all the information in $\mathbf{v}_1$ about $\mathbf{v}_2$. In other words, $\mathbf{z}_1^{suf}$ contains all the shared information between $\mathbf{v}_1$ and $\mathbf{v}_2$, i.e., $I(\mathbf{v}_1, \mathbf{v}_2 | \mathbf{z}_1^{suf}) = 0$. Symmetrically, the sufficient representation $\mathbf{z}_2^{suf}$ of $\mathbf{v}_2$ for $\mathbf{v}_1$ satisfies $I(\mathbf{v}_1, \mathbf{z}_2^{suf}) = I(\mathbf{v}_1, \mathbf{v}_2)$.

Definition 2. *(Minimal Sufficient Representation in Contrastive Learning) The sufficient representation $\mathbf{z}_1^{min}$ of $\mathbf{v}_1$ is minimal if and only if $I(\mathbf{z}_1^{min}, \mathbf{v}_1) \leq I(\mathbf{z}_1^{suf}, \mathbf{v}_1)$, $\forall \mathbf{z}_1^{suf}$ that is sufficient.*

Among all sufficient representations, the minimal sufficient representation $\mathbf{z}_1^{min}$ contains the least information about $\mathbf{v}_1$. Further, it is usually assumed that $\mathbf{z}_1^{min}$ only contains the shared information between $\mathbf{v}_1$ and $\mathbf{v}_2$ and eliminates other non-shared information, i.e., $I(\mathbf{z}_1^{min}, \mathbf{v}_1 | \mathbf{v}_2) = 0$. Note that for a specific input instance, the representation can extract the patterns corresponding to its information from this input to obtain the feature of the instance.

Applying the Data Processing Inequality (Cover & Thomas, 2006) to the Markov chain $\mathbf{v}_1 \to \mathbf{v}_2 \to \mathbf{z}_2$ and $\mathbf{z}_2 \to \mathbf{v}_1 \to \mathbf{z}_1$, we have

$$I(\mathbf{v}_1, \mathbf{v}_2) \geq I(\mathbf{v}_1, \mathbf{z}_2) \geq I(\mathbf{z}_1, \mathbf{z}_2) \tag{2}$$

i.e., $I(\mathbf{v}_1, \mathbf{v}_2)$ is the upper bound of $I(\mathbf{z}_1, \mathbf{z}_2)$. Considering that $I(\mathbf{v}_1, \mathbf{v}_2)$ remains unchanged during the optimization process, contrastive learning optimizes the functions f_1 and f_2 so that $I(\mathbf{z}_1, \mathbf{z}_2)$

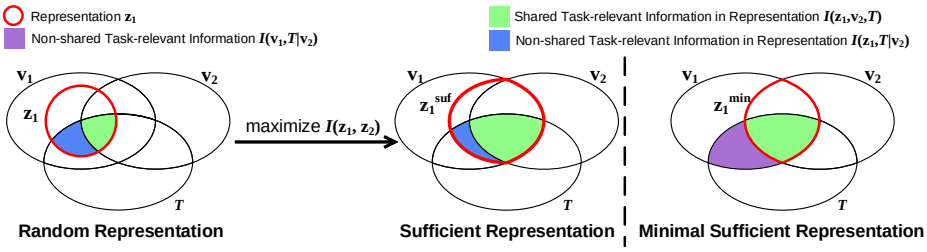


Figure 1: Information diagrams of different representations in contrastive learning. We consider the situation where the non-shared task-relevant information $I(\mathbf{v}_1, T|\mathbf{v}_2)$ cannot be ignored. Contrastive learning makes the representations extracting the shared information between views to obtain the sufficient representation which tends to be minimal. The minimal sufficient representation contains less task-relevant information from the input than other sufficient representations.

approximates $I(\mathbf{v}_1, \mathbf{v}_2)$. When these functions have enough capacity and are well learned based on sufficient data, we can assume $I(\mathbf{z}_1, \mathbf{z}_2) = I(\mathbf{v}_1, \mathbf{v}_2)$, which means the learned representations in contrastive learning are sufficient and tend to be minimal since all supervision information comes from the other view. Therefore, the shared information controls the properties of the representations.

The learned representations in contrastive learning are typically used in various downstream tasks, so we introduce a random variable T to represent the information required for a downstream task which can be classification, regression or clustering task. Tian et al. (2020b) find that the optimal views for contrastive learning depend on the downstream tasks under the assumption of minimal sufficient representation. This discovery is intuitive since various downstream tasks need different information that is unknown during training and it is difficult for the given views to share all the information required for these tasks. For example, when one view is a video stream and the other view is an audio stream, the shared information is sufficient for identity recognition task, but not for object tracking task. Some task-relevant information may not lie in the shared information between views, i.e., $I(\mathbf{v}_1, T|\mathbf{v}_2)$ can not be ignored. Eliminating non-shared information has the risk of damaging the performance in the downstream tasks of the learned representations.

2.2 Disadvantages of minimal sufficient representation

The minimal sufficient representation intuitively is not a good choice for downstream tasks, because it completely eliminates the non-shared information between views which may be important for some downstream tasks. We formalize this problem and theoretically prove that in contrastive learning, the minimal sufficient representation is expected to perform worse in downstream tasks than other sufficient representations. All proofs for the below theorems are provided in Appendix A.

Considering the symmetry between $\mathbf{v}_1$ and $\mathbf{v}_2$, without loss of generality, we take $\mathbf{v}_2$ as the supervision signal for $\mathbf{v}_1$ and take $\mathbf{v}_1$ as the input of a task. It is generally believed that the more task-relevant information contained in the presentations, the better performance can be obtained (Feder & Merhav, 1994; Cover & Thomas, 2006). Therefore, we examine the task-relevant information contained in the representations.

Theorem 1. *(Task-Relevant Information in Representations) In contrastive learning, for a downstream task T, the minimal sufficient representation $\mathbf{z}_1^{min}$ contains less task-relevant information from the input $\mathbf{v}_1$ than other sufficient representation $\mathbf{z}_1^{suf}$, and $I(\mathbf{z}_1^{min}, T)$ has a gap of $I(\mathbf{v}_1, T|\mathbf{v}_2)$ with the upper bound $I(\mathbf{v}_1, T)$. Formally, we have*

$$I(\mathbf{v}_1, T) = I(\mathbf{z}_1^{min}, T) + I(\mathbf{v}_1, T|\mathbf{v}_2) \geq I(\mathbf{z}_1^{suf}, T) = I(\mathbf{z}_1^{min}, T) + I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) \geq I(\mathbf{z}_1^{min}, T) \quad (3)$$

Theorem 1 indicates that $\mathbf{z}_1^{suf}$ may have better performance in task T than $\mathbf{z}_1^{min}$ because it contains more task-relevant information. When non-shared task-relevant information $I(\mathbf{v}_1, T|\mathbf{v}_2)$ is significant, $\mathbf{z}_1^{min}$ has poor performance because it loses a lot of useful information. See Figure 1 for the demonstration using information diagrams. To make this observation more concrete, we examine two types of the downstream task: classification tasks and regression tasks, and provide theoretical analysis on the generalization error of the representations.

When the downstream task is a classification task and T is a categorical variable, we consider the Bayes error rate (Fukunaga, 2013) which is the lowest achievable error for any classifier learned from the representations. Concretely, let P_e be the Bayes error rate of arbitrary learned representation $\mathbf{z}_1$ and $\widehat{T}$ be the prediction for T based on $\mathbf{z}_1$, we have $P_e = 1 - \mathbb{E}_{p(\mathbf{z}_1)}[\max_{t \in T} p(\widehat{T} = t|\mathbf{z}_1)]$ and $0 \leq P_e \leq 1 - 1/|T|$ where $|T|$ is the cardinality of T. According to the value range of P_e, we define a threshold function $\Gamma(x) = \min\{\max\{x, 0\}, 1 - 1/|T|\}$ to prevent overflow.

Theorem 2. *(Bayes Error Rate of Representations) For arbitrary learned representation $\mathbf{z}_1$, its Bayes error rate $P_e = \Gamma(\bar{P}_e)$ with*

$$\bar{P}_e \leq 1 - \exp[-(H(T) - I(\mathbf{z}_1, T|\mathbf{v}_2) - I(\mathbf{z}_1, \mathbf{v}_2, T))] \tag{4}$$

Specifically, for sufficient representation $\mathbf{z}_1^{suf}$, its Bayes error rate $P_e^{suf} = \Gamma(\bar{P}_e^{suf})$ with

$$\bar{P}_e^{suf} \leq 1 - \exp[-(H(T) - I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) - I(\mathbf{v}_1, \mathbf{v}_2, T))] \tag{5}$$

for minimal sufficient representation $\mathbf{z}_1^{min}$, its Bayes error rate $P_e^{min} = \Gamma(\bar{P}_e^{min})$ with

$$\bar{P}_e^{min} \leq 1 - \exp[-(H(T) - I(\mathbf{v}_1, \mathbf{v}_2, T))] \tag{6}$$

Since $I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) \geq 0$, Theorem 2 indicates for classification task T, the upper bound of P_e^{min} is larger than P_e^{suf}. In other words, $\mathbf{z}_1^{min}$ is expected to obtain a higher classification error rate in the task T than $\mathbf{z}_1^{suf}$. According to the Equation (5), considering that $H(T)$ and $I(\mathbf{v}_1, \mathbf{v}_2, T)$ are not related to the representations, increasing $I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2)$ can reduce the Bayes error rate in task T. When $I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) = I(\mathbf{v}_1, T|\mathbf{v}_2)$, $\mathbf{z}_1^{suf}$ contains all the useful information for task T in $\mathbf{v}_1$.

When the downstream task is a regression task and T is a continuous variable, let $\widetilde{T}$ be the prediction for T based on arbitrary learned representation $\mathbf{z}_1$, we consider the smallest achievable expected squared prediction error $R_e = \min_{\widetilde{T}} \mathbb{E}[(T - \widetilde{T}(\mathbf{z}_1))^2] = \mathbb{E}[\varepsilon^2]$ with $\varepsilon(T, \mathbf{z}_1) = T - \mathbb{E}[T|\mathbf{z}_1]$.

Theorem 3. *(Minimum Expected Squared Prediction Error of Representations) For arbitrary learned representation $\mathbf{z}_1$, when the conditional distribution $p(\varepsilon|\mathbf{z}_1)$ is uniform, Laplacian or Gaussian distribution, the minimum expected squared prediction error R_e satisfies*

$$R_e = \alpha \cdot \exp[2 \cdot (H(T) - I(\mathbf{z}_1, T|\mathbf{v}_2) - I(\mathbf{z}_1, \mathbf{v}_2, T))] \tag{7}$$

Specifically, for sufficient representation $\mathbf{z}_1^{suf}$, its minimum expected prediction error R_e^{suf} satisfies

$$R_e^{suf} = \alpha \cdot \exp[2 \cdot (H(T) - I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) - I(\mathbf{v}_1, \mathbf{v}_2, T))] \tag{8}$$

for minimal sufficient representation $\mathbf{z}_1^{min}$, its minimum expected prediction error R_e^{min} satisfies

$$R_e^{min} = \alpha \cdot \exp[2 \cdot (H(T) - I(\mathbf{v}_1, \mathbf{v}_2, T))] \tag{9}$$

where the constant coefficient α depends on the conditional distribution $p(\varepsilon|\mathbf{z}_1)$.

The assumption about estimation error ε in Theorem 3 is reasonable because ε is analogous to the 'noise' with the mean of 0 which is generally assumed to come from simple distributions (e.g., Gaussian distribution) in statistical learning theory. Similar to the classification tasks, Theorem 3 indicates that for regression tasks, $\mathbf{z}_1^{suf}$ can achieve lower expected squared prediction error than $\mathbf{z}_1^{min}$ and increasing $I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2)$ can improve the performance in downstream regression tasks.

Theorem 2 and Theorem 3 analyze the disadvantages of the minimal sufficient representation $\mathbf{z}_1^{min}$ in classification tasks and regression tasks respectively. The essential reason is that $\mathbf{z}_1^{min}$ has less task-relevant information than other sufficient representation $\mathbf{z}_1^{suf}$ and has a non-ignorable gap $I(\mathbf{v}_1, T|\mathbf{v}_2)$ with the optimal representation, as shown in Theorem 1.

2.3 Extracting non-shared task-relevant information

According to the above theoretical analysis, in contrastive learning, the minimal sufficient representation is not sufficient for downstream tasks due to the lack of some non-shared task-relevant information. Moreover, contrastive learning tends to learn the minimal sufficient representation,

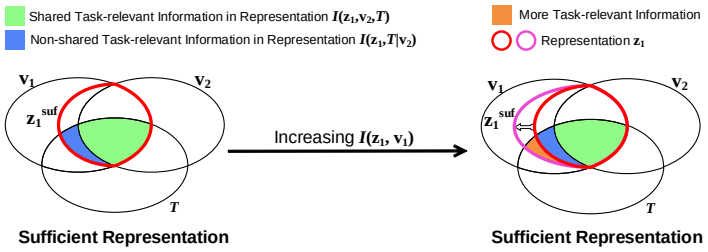


Figure 2: Demonstration of our motivation using information diagrams. Based on the sufficient representation learned by the contrastive learning models, increasing $I(\mathbf{z}_1, \mathbf{v}_1)$ approximately introduces more non-shared task-relevant information.

thereby having the risk of over-fitting to the shared information between views. To this end, we propose to extract more non-shared task-relevant information in $\mathbf{v}_1$, i.e., increasing $I(\mathbf{z}_1, T|\mathbf{v}_2)$. However, we can not utilize any downstream task information during training, so it is impossible to increase $I(\mathbf{z}_1, T|\mathbf{v}_2)$ directly. We consider increasing $I(\mathbf{z}_1, \mathbf{v}_1)$ as an alternative because the increased information from $\mathbf{v}_1$ in $\mathbf{z}_1$ may be relevant to some downstream tasks, and this motivation is demonstrated in Figure 2. In addition, increasing $I(\mathbf{z}_1, \mathbf{v}_1)$ also helps to extract the shared information between views at the beginning of the optimization process. Concretely, considering the symmetry between $\mathbf{v}_1$ and $\mathbf{v}_2$, our optimization objective is

$$\max_{f_1, f_2} I(\mathbf{z}_1, \mathbf{z}_2) + \lambda_1 I(\mathbf{z}_1, \mathbf{v}_1) + \lambda_2 I(\mathbf{z}_2, \mathbf{v}_2) \tag{10}$$

which consists of the original optimization objective (1) in contrastive learning and the regularization terms we proposed. The coefficients λ_1 and λ_2 are used to control the amount of increasing $I(\mathbf{z}_1, \mathbf{v}_1)$ and $I(\mathbf{z}_2, \mathbf{v}_2)$ respectively. For optimizing $I(\mathbf{z}_1, \mathbf{z}_2)$, we adopt the commonly used implementations in contrastive learning models (Chen et al., 2020; Grill et al., 2020b; Zbontar et al., 2021) which are usually the lower bound estimate of mutual information. For optimizing $I(\mathbf{z}_i, \mathbf{v}_i), i = 1, 2$, we consider two different implementations.

Implementation I Since $I(\mathbf{z}, \mathbf{v}) = H(\mathbf{v}) - H(\mathbf{v}|\mathbf{z})$ and $H(\mathbf{v})$ is not related with $\mathbf{z}$, we can equivalently decrease the conditional entropy $H(\mathbf{v}|\mathbf{z}) = -\mathbb{E}_{p(\mathbf{z},\mathbf{v})}[\ln p(\mathbf{v}|\mathbf{z})]$. Concretely, we use the representation $\mathbf{z}$ to reconstruct the original input $\mathbf{v}$, as done in auto-encoder models (Vincent et al., 2010). Decreasing the entropy of reconstruction encourages the representation $\mathbf{z}$ to contain more information about the original input $\mathbf{v}$. However, the conditional distribution $p(\mathbf{v}|\mathbf{z})$ is intractable in practice, so we use the distribution $q(\mathbf{v}|\mathbf{z})$ as an approximation, and have

$$\mathbb{E}_{p(\mathbf{z},\mathbf{v})}[\ln p(\mathbf{v}|\mathbf{z})] - \mathbb{E}_{p(\mathbf{z},\mathbf{v})}[\ln q(\mathbf{v}|\mathbf{z})] = \mathbb{E}_{p(\mathbf{z})}[D_{\mathrm{KL}}(p(\mathbf{v}|\mathbf{z})\|q(\mathbf{v}|\mathbf{z}))] \geq 0 \tag{11}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ represents the Kullback–Leibler divergence. Therefore, $\mathbb{E}_{p(\mathbf{z},\mathbf{v})}[\ln q(\mathbf{v}|\mathbf{z})]$ is the lower bound of $\mathbb{E}_{p(\mathbf{z},\mathbf{v})}[\ln p(\mathbf{v}|\mathbf{z})]$, and we can increase $\mathbb{E}_{p(\mathbf{z},\mathbf{v})}[\ln q(\mathbf{v}|\mathbf{z})]$ as an alternative objective. According to the type of input $\mathbf{v}$ (e.g., images, text or audio), $q(\mathbf{v}|\mathbf{z})$ can be any appropriate distribution with known probability density function, such as Bernoulli distribution, Gaussian distribution or Laplace distribution, and its parameters are the functions of $\mathbf{z}$. For example, when $q(\mathbf{v}|\mathbf{z})$ is the Gaussian distribution $\mathcal{N}(\mathbf{v}; \boldsymbol{\mu}(\mathbf{z}), \sigma^2 \boldsymbol{I})$ with given variance σ^2 and deterministic mean function $\boldsymbol{\mu}(\cdot)$ which is usually parameterized by neural networks, we have

$$\mathbb{E}_{p(\mathbf{z},\mathbf{v})}[\ln q(\mathbf{v}|\mathbf{z})] \propto -\mathbb{E}_{p(\mathbf{z},\mathbf{v})}[\|\mathbf{v} - \boldsymbol{\mu}(\mathbf{z})\|_2^2] + c \tag{12}$$

where c is a constant to representation $\mathbf{z}$. The final optimization objective becomes

$$\max_{f_1, f_2, \boldsymbol{\mu}} I(\mathbf{z}_1, \mathbf{z}_2) - \lambda_1 \mathbb{E}_{p(\mathbf{z}_1,\mathbf{v}_1)}[\|\mathbf{v}_1 - \boldsymbol{\mu}_1(\mathbf{z}_1)\|_2^2] - \lambda_2 \mathbb{E}_{p(\mathbf{z}_2,\mathbf{v}_2)}[\|\mathbf{v}_2 - \boldsymbol{\mu}_2(\mathbf{z}_2)\|_2^2] \tag{13}$$

Implementation II Although the above implementation is effective in practice, it needs to reconstruct the input which is challenging for high-dimensional input and increases the amount of model parameters. To this end, we propose another representation-level implementation as an optional alternative. We investigate various lower bound estimates of mutual information, such as the bound of Barber and Agakov (Barber & Agakov, 2003), the bound of Nguyen, Wainwright and Jordan

(Nguyen et al., 2010), MINE (Belghazi et al., 2018) and InfoNCE (Poole et al., 2019). We choose the InfoNCE lower bound and the detailed discussion is provided in Appendix D. Concretely, the InfoNCE lower bound is

$$\hat{I}_{NCE}(\mathbf{z}, \mathbf{v}) = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} \ln \frac{p(\mathbf{z}^k|\mathbf{v}^k)}{\frac{1}{N}\sum_{l=1}^{N} p(\mathbf{z}^l|\mathbf{v}^k)}\right] \tag{14}$$

where $(\mathbf{z}^k, \mathbf{v}^k), k = 1, \cdots, N$ are N copies of $(\mathbf{z}, \mathbf{v})$ and the expectation is over $\Pi_k p(\mathbf{z}^k, \mathbf{v}^k)$. In the first implementation, we map the input $\mathbf{v}$ to the representation $\mathbf{z}$ through a deterministic function f with $\mathbf{z} = f(\mathbf{v})$ and approximate the distribution $p(\mathbf{v}|\mathbf{z})$ of the reconstructed input. Differently, here we need to define $p(\mathbf{z}|\mathbf{v})$ to calculate the InfoNCE lower bound which means the representation $\mathbf{z}$ is no longer a deterministic output of input $\mathbf{v}$, so we use the reparameterization trick (Kingma & Welling, 2014) during training. For example, when we define $p(\mathbf{z}|\mathbf{v})$ as the Gaussian distribution $\mathcal{N}(\mathbf{z}; f(\mathbf{v}), \sigma^2 \boldsymbol{I})$ with given variance σ^2 and the function f is the same as in the first implementation, we have $\mathbf{z} = f(\mathbf{v}) + \epsilon\sigma, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\hat{I}_{NCE}$ is equivalent to

$$\tilde{I}_{NCE}(\mathbf{z}, \mathbf{v}) = \mathbb{E}\left[-\frac{1}{N}\sum_{k=1}^{N} \ln \sum_{l=1}^{N} \exp(-\rho\|\mathbf{z}^l - f(\mathbf{v}^k)\|_2^2)\right] \tag{15}$$

where ρ is a scale factor. The final optimization objective becomes

$$\max_{f_1, f_2} I(\mathbf{z}_1, \mathbf{z}_2) + \lambda_1 \tilde{I}_{NCE}(\mathbf{z}_1, \mathbf{v}_1) + \lambda_2 \tilde{I}_{NCE}(\mathbf{z}_2, \mathbf{v}_2) \tag{16}$$

Since the regularization term (15) is calculated at the representation-level, when we use the convolutional neural networks (e.g., ResNet (He et al., 2016)) to parameterize f, it can be applied to the output activation of multiple internal blocks.

It is worth noting that increasing $I(\mathbf{z}, \mathbf{v})$ is not conflict with the information bottleneck theory (Tishby & Zaslavsky, 2015). This theory tells us to compress the information from the input $\mathbf{v}$ in the representation $\mathbf{z}$ under the condition that the representation $\mathbf{z}$ is sufficient for the task T. However, according to our analysis, the learned representations in contrastive learning are not sufficient for the downstream tasks. Therefore, we need to make the information in the representations more sufficient for the downstream tasks and it is not time to compress it. On the other hand, we can not introduce too much information from the input $\mathbf{v}$ either, which may contain too much noise to increase the data demand. Here we use the coefficients λ_1 and λ_2 to control this.

3 Related work

Contrastive learning (Hjelm et al., 2018; He et al., 2020; Chen et al., 2020) is a successful unsupervised representation learning framework and the learned representations are useful in various downstream tasks (He et al., 2020). In contrastive learning, the views are constructed by exploiting the internal structures of unlabeled data and typically share the information in which we are interested. Recently, Tian et al. (2020b) find that the optimal views for contrastive learning depend on the downstream tasks under the assumption of minimal sufficient representation. In other words, the optimal views for the downstream task T_1 may not be suitable for the task T_2. The reason may be that some information relevant to T_2 is not shared between the views. In this work, we formalize this conjecture and provide theoretical analysis. According to our analysis, we find a new problem that contrastive learning may over-fit to the shared information between views, and thus propose to increase the mutual information between representation and input to alleviate this problem.

Conversely, some recent works (Federici et al., 2020; Tsai et al., 2021) propose to learn the minimal sufficient representation. They assume that either view alone is approximately redundant to the other view for the downstream tasks, i.e., almost all the information relevant to downstream tasks is shared between views. However, redundancy only makes sense for the fixed known task and the downstream tasks in contrastive learning are changeable and unknown, so this is an overly idealistic assumption and conflicts with the discovery in (Tian et al., 2020b).

In this work, we consider two implementations to increase the mutual information between the representation and input. Our first implementation refers to the auto-encoder models (Vincent et al., 2010; Kingma & Welling, 2014) which reconstruct the input to make the representation containing the key information about the data. Our second implementation relies on the high-dimensional mutual information estimate (Belghazi et al., 2018; Poole et al., 2019).

4 EXPERIMENTS

In this section, we first verify the effectiveness of increasing $I(\mathbf{z}, \mathbf{v})$ on various datasets, and then provide some analytical experiments. We choose two classic contrastive learning models as our baselines: SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020a). We denote our first implementation (13) as "RC" for "ReConstruction" and the second implementation (16) as "LBE" for "Lower Bound Estimate". For all experiments, we use random cropping, flip and random color distortion as the data augmentation, just as suggested by Chen et al. (2020). For "LBE", we set $\sigma = 0.1$ and $\rho = 0.05$, and apply it to the output activation of the last three blocks in ResNet.

4.1 EFFECTIVENESS OF INCREASING $I(\mathbf{z}, \mathbf{v})$

We consider different types of downstream task, including classification tasks and instance segmentation tasks. Due to limited space, we provide the results of classification tasks below and the results of instance segmentation tasks in Appendix B.1.

Experimental setup. We train the models on CIFAR10 (Krizhevsky et al., 2009) and STL-10 (Coates et al., 2011), and evaluate the learned representations on the source dataset and six transfer datasets: DTD (Cimpoi et al., 2014), MNIST (LeCun et al., 1998), FashionMNIST (Xiao et al., 2017), CUBirds (Wah et al., 2011), VGG Flower (Nilsback & Zisserman, 2008) and Traffic Signs (Houben et al., 2013). We follow the linear evaluation protocol where a linear classifier is trained on top of the frozen backbone. For contrastive learning, we use a ResNet18 backbone and the models are trained for 200 epochs with batch size 256 using Adam optimizer with learning rate 3e-4. For linear evaluation, the linear classifier is trained for 100 epochs with batch size 128 using SGD optimizer with learning rate 1e-2 and momentum 0.9. We drop the learning rate by a factor of 10 on epoch 60 and 80. We set $\lambda_1 = \lambda_2 = 1$ for SimCLR and $\lambda_1 = \lambda_2 = 0.1$ for BYOL.

Table 1: Downstream classification accuracy ($\%$) on the source dataset (CIFAR10 or STL-10) and six transfer datasets.

Model	CIFAR10	DTD	MNIST	FaMNIST	CUBirds	VGGFlower	TrafficSigns
SimCLR	85.76	29.52	97.03	88.36	8.87	42.81	92.41
SimCLR+RC (ours)	**85.78**	33.67	**97.99**	**90.31**	**10.89**	**54.16**	**95.84**
SimCLR+LBE (ours)	85.45	**34.52**	97.94	89.26	10.60	54.10	94.96
BYOL	85.64	31.22	97.15	88.92	8.84	40.90	92.17
BYOL+RC (ours)	**85.80**	**34.73**	**98.07**	**89.61**	9.68	48.75	94.19
BYOL+LBE (ours)	85.28	33.99	97.76	88.99	**9.96**	**54.10**	**95.09**
Model	STL-10	DTD	MNIST	FaMNIST	CUBirds	VGGFlower	TrafficSigns
SimCLR	78.74	39.41	95.00	87.31	8.34	49.41	80.25
SimCLR+RC (ours)	79.21	41.81	**97.48**	**89.98**	10.03	**60.46**	**94.73**
SimCLR+LBE (ours)	**80.17**	**42.07**	97.04	88.68	**10.11**	58.51	87.77
BYOL	80.83	40.05	94.45	87.23	8.54	49.41	77.54
BYOL+RC (ours)	**81.11**	42.02	**96.96**	**88.92**	9.63	55.71	**88.57**
BYOL+LBE (ours)	80.85	**42.55**	95.75	87.88	**10.55**	**59.39**	84.62

Results. Table 1 shows the results on CIFAR10 and STL-10, and the best result in each block is in bold. Increasing $I(\mathbf{z}, \mathbf{v})$ can introduce non-shared information and improve the classification accuracy, especially on transfer datasets. This means the shared information between views is not sufficient for some tasks, e.g., classification on DTD, VGG Flower and Traffic Signs where increasing $I(\mathbf{z}, \mathbf{v})$ achieves significant improvement. In other words, increasing $I(\mathbf{z}, \mathbf{v})$ can prevent the models from over-fitting to the shared information between views. Note that the introduced information is not guaranteed to be significantly effective for all tasks, e.g., classification on MNIST, FashionMNIST or CIFAR10 where increasing $I(\mathbf{z}, \mathbf{v})$ achieves slight improvement.

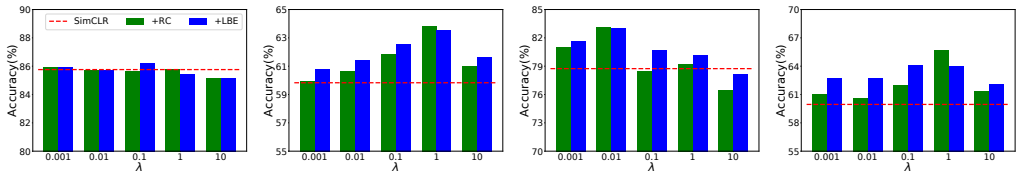
4.2 ANALYTICAL EXPERIMENTS

Next, we provide some analytical experiments to further understand our hypotheses, theoretical analysis and model. All experiments use the same training schedule as in the Section 4.1.

Table 2: Downstream classification accuracy (%) on CIFAR10 and six transfer datasets. † represents adding Gaussian noise to the representations.

Model	CIFAR10	DTD	MNIST	FaMNIST	CUBirds	VGGFlower	TrafficSigns
SimCLR	85.76	29.52	**97.03**	**88.36**	**8.87**	42.81	**92.41**
SimCLR+IP	**85.86**	**30.15**	96.71	88.18	8.66	**43.22**	92.13
SimCLR†	85.81	**31.70**	**97.08**	**88.85**	8.77	**44.41**	92.41
SimCLR+MIB	**86.20**	31.17	97.00	88.62	**9.01**	43.88	**93.01**

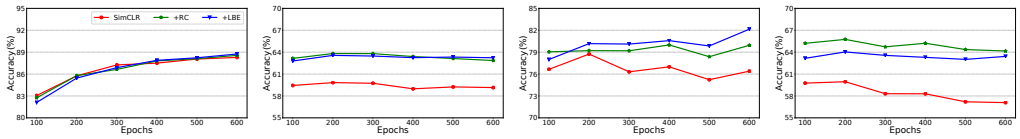
Performance of eliminating non-shared information. Some recent works (Federici et al., 2020; Tsai et al., 2021) propose to eliminate the non-shared information between views to get the minimal sufficient representation. To this end, Federici et al. (2020) minimize the regularization term $L_{MIB} = \frac{1}{2}\left(D_{\mathrm{KL}}(p(\mathbf{z}_1|\mathbf{v}_1)\|p(\mathbf{z}_2|\mathbf{v}_2)) + D_{\mathrm{KL}}(p(\mathbf{z}_2|\mathbf{v}_2)\|p(\mathbf{z}_1|\mathbf{v}_1))\right)$, and when $p(\mathbf{z}_1|\mathbf{v}_1)$ and $p(\mathbf{z}_2|\mathbf{v}_2)$ are modeled by $\mathcal{N}(\mathbf{z}_i; f_i(\mathbf{v}_i), \sigma^2\boldsymbol{I}), i = 1, 2$ with given variance σ^2, it can be rewritten as $L_{MIB} = \mathbb{E}_{p(\mathbf{v}_1,\mathbf{v}_2)}\left[\|f_1(\mathbf{v}_1) - f_2(\mathbf{v}_2)\|_2^2\right]$. Identically, Tsai et al. (2021) minimize the inverse predictive loss $L_{IP} = \mathbb{E}_{p(\mathbf{v}_1,\mathbf{v}_2)}\left[\|f_1(\mathbf{v}_1) - f_2(\mathbf{v}_2)\|_2^2\right]$. The detailed derivation are provided in Appendix C. We evaluate the effect of these two regularization terms on the classification tasks and choose their coefficient with best accuracy on the source dataset. The results are shown in Table 2 and the best result in each block is in bold. Although these two regularization terms have the same form, Federici et al. (2020) uses stochastic encoders which is equivalent to adding Gaussian noise, so we report the results of SimCLR with Gaussian noise, marked by †. As we can see, eliminating the non-shared information can not change the accuracy in downstream classification tasks much. This means that the sufficient representation learned in contrastive learning tends to be minimal and we don't need to further remove the non-shared information.



(a) CIFAR10 (b) Trans. from CIFAR10 (c) STL-10 (d) Trans. from STL-10

Figure 3: The classification accuracy on the source dataset (CIFAR10 or STL-10) and the averaged accuracy on six transfer datasets with varying hyper-parameter λ.

Changing the amount of increasing $I(\mathbf{z}, \mathbf{v})$. The hyper-parameters λ_1 and λ_2 control the amount of increasing $I(\mathbf{z}_1, \mathbf{v}_1)$ and $I(\mathbf{z}_2, \mathbf{v}_2)$ respectively. Therefore, we set $\lambda_1 = \lambda_2 = \lambda$ and evaluate the performance of different λ from $\{0.001, 0.01, 0.1, 1, 10\}$. We choose SimCLR as the baseline and the results are shown in Figure 3. We report the accuracy on the source dataset (CIFAR10 or STL-10) and the averaged accuracy on six transfer datasets. Oversized λ (e.g., 10) damages the optimization of the contrastive loss, resulting in a decrease in performance. But for other reasonable λ, increasing $I(\mathbf{z}, \mathbf{v})$ consistently improves the performance in downstream classification tasks. We can observe a non-monotonous reverse-U trend of accuracy with the change of λ, which means excessively increasing $I(\mathbf{z}, \mathbf{v})$ may introduce noise beside useful information.



(a) CIFAR10 (b) Trans. from CIFAR10 (c) STL-10 (d) Trans. from STL-10

Figure 4: The classification accuracy on the source dataset (CIFAR10 or STL-10) and the averaged accuracy on six transfer datasets with varying epochs.

Training with more epochs. In the above experiments, we train all models for 200 epochs. Here we further show the behavior of the contrastive learning models and increasing $I(\mathbf{z}, \mathbf{v})$ when training with more epochs. We choose SimCLR as the baseline and train all models for 100, 200, 300, 400, 500 and 600 epochs. The results are shown in Figure 4. With more training epochs, the learned representations are more approximate to the minimal sufficient representation and mainly contain the shared information between views and ignore the non-shared information. For the classification tasks on the transfer datasets, the shared information between views is not sufficient. Concretely, the accuracy on the transfer datasets decreases with more epochs and the learned representations over-fit to the shared information between views. Increasing $I(\mathbf{z}, \mathbf{v})$ can introduce non-shared information and obtain the significant improvement. For the classification tasks on the source datasets, the shared information between views is sufficient on CIFAR10 but not on STL-10. Concretely, the accuracy on CIFAR10 increases with more epochs and increasing $I(\mathbf{z}, \mathbf{v})$ can not make a difference. But the accuracy on STL-10 decreases with more epochs, and increasing $I(\mathbf{z}, \mathbf{v})$ can significantly improve the accuracy and does not decrease with more epochs. In fact, we use the *unlabeled* split for contrastive training on STL-10, so it is intuitive that the shared information between views is not sufficient for the classification task on the *train* and *test* split.

Table 3: Downstream classification accuracy (%) on the source dataset and six transfer datasets.

Model	CIFAR10	DTD	MNIST	FaMNIST	CUBirds	VGGFlower	TrafficSigns
Supervised	**93.25**	34.10	98.52	90.09	8.37	46.14	93.05
Supervised+RC (ours)	93.09	32.77	98.61	89.77	8.84	49.05	93.28
Supervised+LBE (ours)	93.18	**34.79**	**98.68**	**90.40**	**9.72**	**53.15**	**94.47**
Model	CIFAR100	DTD	MNIST	FaMNIST	CUBirds	VGGFlower	TrafficSigns
Supervised	71.92	36.06	98.48	88.97	11.51	64.21	96.54
Supervised+RC (ours)	**72.02**	34.79	**98.59**	89.35	10.94	65.34	96.67
Supervised+LBE (ours)	71.89	**36.33**	98.37	**89.42**	**11.89**	**65.64**	**96.91**

Increasing $I(\mathbf{z}, \mathbf{x})$ in supervised learning. According to the information bottleneck theory (Tishby & Zaslavsky, 2015), a model extracts the approximate minimal sufficient statistics of the input $\mathbf{x}$ with respect to the label $\mathbf{y}$ in supervised learning. In other words, the representation $\mathbf{z}$ only contains the information related to the label and eliminates other irrelevant information which is considered as noise. However, label-irrelevant information may be useful for other tasks, so we evaluate the effect of increasing $I(\mathbf{z}, \mathbf{x})$ in supervised learning. We train the ResNet18 backbone using cross-entropy classification loss on CIFAR10 and CIFAR100, and choose $\lambda_1 = \lambda_2 = \lambda$ from $\{0.001, 0.01, 0.1, 1\}$. The training and evaluate schedule is the same as in Section 4.1. The results are shown in Table 3 and the best result in each block is in bold. As we can see, increasing $I(\mathbf{z}, \mathbf{x})$ improves the performance on the transfer datasets and achieves comparable results on the source dataset, which means it can effectively alleviate the over-fitting on the label information. This discovery helps to obtain more general representations in the field of supervised pre-training.

5 Conclusions and Future Works

In this work, we explore the relationship between the learned representations and downstream tasks in contrastive learning. Although some recent works propose to learn the minimal sufficient representation, we theoretically and empirically verify that the minimal sufficient representation is not sufficient for downstream tasks because it loses non-shared task-relevant information. We find that contrastive learning tend to obtain the minimal sufficient representation, which means it may over-fit to the shared information between views. To this end, we propose the regularization term which increases the mutual information between the representation and input to approximately introduce more non-shared task-relevant information when the downstream tasks are unknown. Extensive experiments show that our method can effectively prevent the contrastive learning models from over-fitting to the shared information between views. For the future works, we suggest two directions. 1) We can consider the situation where the downstream tasks or even some downstream supervision information are given. Then we can design task-customized views and regularization terms which can directly introduce more task-relevant information. 2) In contrastive learning, one view plays the same role as the label in supervised learning for the other view, so we can extend our analysis and method to the supervised learning for further exploration.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, 32: 15535–15545, 2019.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pp. 201–208, 2003.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pp. 531–540. PMLR, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

Meir Feder and Neri Merhav. Relations between entropy and error probability. *IEEE Transactions on Information theory*, 40(1):259–266, 1994.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Benoît Frénay, Gauthier Doquire, and Michel Verleysen. Is mutual information adequate for feature selection in regression? *Neural Networks*, 48:1–7, 2013.

Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Pires, Zhaohan Guo, Mohammad Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Neural Information Processing Systems*, 2020a.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

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.

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

Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *The 2013 international joint conference on neural networks (IJCNN)*, pp. 1–8. Ieee, 2013.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9879–9889, 2020.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

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.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2015.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 2010.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. 2019. *URL https://github. com/facebookresearch/detectron2*, 2(3), 2019.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. *arXiv preprint arXiv:1708.03888*, 6:12, 2017.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. PMLR, 2021.

A PROOFS OF THEOREMS

In this section, we provide the proofs of the theorems in the main text. Since the random variable $\mathbf{z}_1 = f_1(\mathbf{v}_1)$ is the representation of random variable $\mathbf{v}_1$ where f_1 is an encoding function, we have

Assumption 1. *Random variable $\mathbf{z}_1$ is conditionally independent from any other variable $\mathbf{s}$ in the system once random variable $\mathbf{v}_1$ is observed, i.e., $I(\mathbf{z}_1, \mathbf{s}|\mathbf{v}_1) = 0, \forall \mathbf{s}$.*

This assumption is also adopted in Federici et al. (2020). When f_1 is a deterministic function, this assumption strictly holds. And when f_1 is a random function, the information in $\mathbf{z}_1$ consists of the information from $\mathbf{v}_1$ and the information introduced by the randomness of function f_1 which can be considered irrelevant to other variables in the system, so this assumption still holds. We first present two lemmas for subsequent proofs.

Lemma 1. *Let $\mathbf{z}_1^{suf}$ and $\mathbf{z}_1^{min}$ are a sufficient representation and the minimal sufficienr representation of view $\mathbf{v}_1$ for $\mathbf{v}_2$ in contrative learning respectively, we have*

$$I(\mathbf{z}_1^{min}, \mathbf{v}_2, T) = I(\mathbf{z}_1^{suf}, \mathbf{v}_2, T) = I(\mathbf{v}_1, \mathbf{v}_2, T) \tag{17}$$

$$I(\mathbf{z}_1^{min}, T|\mathbf{v}_2) = 0 \tag{18}$$

Proof. 1) From the Definition 1 and the Assumption 1, we have

$$\begin{aligned}
&I(\mathbf{v}_1, \mathbf{v}_2, T) - I(\mathbf{z}_1^{suf}, \mathbf{v}_2, T) \\
&= [I(\mathbf{v}_1, \mathbf{v}_2) - I(\mathbf{v}_1, \mathbf{v}_2|T)] - [I(\mathbf{z}_1^{suf}, \mathbf{v}_2) - I(\mathbf{z}_1^{suf}, \mathbf{v}_2|T)] \\
&= I(\mathbf{z}_1^{suf}, \mathbf{v}_2|T) - I(\mathbf{v}_1, \mathbf{v}_2|T) \\
&= [H(\mathbf{v}_2|T) - H(\mathbf{v}_2|\mathbf{z}_1^{suf}, T)] - [H(\mathbf{v}_2|T) - H(\mathbf{v}_2|\mathbf{v}_1, T)] \\
&= H(\mathbf{v}_2|\mathbf{v}_1, T) - H(\mathbf{v}_2|\mathbf{z}_1^{suf}, T) \\
&= [I(\mathbf{z}_1^{suf}, \mathbf{v}_2|\mathbf{v}_1, T) + H(\mathbf{v}_2|\mathbf{v}_1, \mathbf{z}_1^{suf}, T)] - [I(\mathbf{v}_1, \mathbf{v}_2|\mathbf{z}_1^{suf}, T) + H(\mathbf{v}_2|\mathbf{v}_1, \mathbf{z}_1^{suf}, T)] \\
&= I(\mathbf{z}_1^{suf}, \mathbf{v}_2|\mathbf{v}_1, T) - I(\mathbf{v}_1, \mathbf{v}_2|\mathbf{z}_1^{suf}, T) \\
&= I(\mathbf{z}_1^{suf}, \mathbf{v}_2|\mathbf{v}_1, T) = 0
\end{aligned}$$

Therefore, we have

$$I(\mathbf{z}_1^{suf}, \mathbf{v}_2, T) = I(\mathbf{v}_1, \mathbf{v}_2, T)$$

The above proof process only uses the sufficiency of $\mathbf{z}_1^{suf}$ for $\mathbf{v}_2$, so we have

$$I(\mathbf{z}_1^{min}, \mathbf{v}_2, T) = I(\mathbf{v}_1, \mathbf{v}_2, T)$$

2) From the Definition 2 and the Assumption 1, we have

$$I(\mathbf{z}_1^{min}, \mathbf{v}_1|\mathbf{v}_2) = 0 \qquad I(\mathbf{z}_1^{min}, T|\mathbf{v}_1) = 0$$

Applying these two equations, we have

$$\begin{aligned}
I(\mathbf{z}_1^{min}, T|\mathbf{v}_2) &= I(\mathbf{z}_1^{min}, T|\mathbf{v}_1, \mathbf{v}_2) + I(\mathbf{z}_1^{min}, T, \mathbf{v}_1|\mathbf{v}_2) \\
&= I(\mathbf{z}_1^{min}, T, \mathbf{v}_1|\mathbf{v}_2) \\
&= I(\mathbf{z}_1^{min}, \mathbf{v}_1|\mathbf{v}_2) - I(\mathbf{z}_1^{min}, \mathbf{v}_1|T, \mathbf{v}_2) = 0
\end{aligned}$$

□

We consider the conditional entropy of the task variable T given the representation $\mathbf{z}_1$.

Lemma 2. *For arbitrary learned representation $\mathbf{z}_1$, the conditional entropy $H(T|\mathbf{z}_1)$ of the task variable T given $\mathbf{z}_1$ satisfies*

$$H(T|\mathbf{z}_1) = H(T) - I(\mathbf{z}_1, T|\mathbf{v}_2) - I(\mathbf{z}_1, \mathbf{v}_2, T) \tag{19}$$

Specifically, for the sufficient representation $\mathbf{z}_1^{suf}$, the conditional entropy $H(T|\mathbf{z}_1^{suf})$ satisfies

$$H(T|\mathbf{z}_1^{suf}) = H(T) - I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) - I(\mathbf{v}_1, \mathbf{v}_2, T) \tag{20}$$

for the minimal sufficient representation $\mathbf{z}_1^{min}$, the conditional entropy $H(T|\mathbf{z}_1^{min})$ satisfies

$$H(T|\mathbf{z}_1^{min}) = H(T) - I(\mathbf{v}_1, \mathbf{v}_2, T) \tag{21}$$

Proof. We have

$$\begin{aligned} H(T|\mathbf{z}_1) &= H(T) - I(T, \mathbf{z}_1) \\ &= H(T) - [I(T, \mathbf{z}_1, \mathbf{v}_2) + I(T, \mathbf{z}_1|\mathbf{v}_2)] \\ &= H(T) - I(\mathbf{z}_1, T|\mathbf{v}_2) - I(\mathbf{z}_1, \mathbf{v}_2, T) \end{aligned}$$

Applying the Equation (17), the conditional entropy $H(T|\mathbf{z}_1^{suf})$ satisfies

$$\begin{aligned} H(T|\mathbf{z}_1^{suf}) &= H(T) - I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) - I(\mathbf{z}_1^{suf}, \mathbf{v}_2, T) \\ &= H(T) - I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) - I(\mathbf{v}_1, \mathbf{v}_2, T) \end{aligned}$$

Further, applying the Equation (18), the conditional entropy $H(T|\mathbf{z}_1^{min})$ satisfies

$$\begin{aligned} H(T|\mathbf{z}_1^{min}) &= H(T) - I(\mathbf{z}_1^{min}, T|\mathbf{v}_2) - I(\mathbf{v}_1, \mathbf{v}_2, T) \\ &= H(T) - I(\mathbf{v}_1, \mathbf{v}_2, T) \end{aligned}$$

□

Finally, we formally give the proofs of Theorem 1, 2 and 3.

The proof of Theorem 1.

Proof. We decompose the Theorem 1 into three equations and prove them in turn.

1) $I(\mathbf{v}_1, T) = I(\mathbf{z}_1^{min}, T) + I(\mathbf{v}_1, T|\mathbf{v}_2)$.

$$\begin{aligned} I(\mathbf{v}_1, T) &= I(\mathbf{v}_1, T, \mathbf{v}_2) + I(\mathbf{v}_1, T|\mathbf{v}_2) \\ &= I(\mathbf{z}_1^{min}, T, \mathbf{v}_2) + I(\mathbf{v}_1, T|\mathbf{v}_2) \\ &= I(\mathbf{z}_1^{min}, T) - I(\mathbf{z}_1^{min}, T|\mathbf{v}_2) + I(\mathbf{v}_1, T|\mathbf{v}_2) \\ &= I(\mathbf{z}_1^{min}, T) + I(\mathbf{v}_1, T|\mathbf{v}_2) \end{aligned}$$

2) $I(\mathbf{z}_1^{suf}, T) = I(\mathbf{z}_1^{min}, T) + I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2)$.

$$\begin{aligned} I(\mathbf{z}_1^{suf}, T) &= I(\mathbf{z}_1^{suf}, T, \mathbf{v}_2) + I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) \\ &= I(\mathbf{z}_1^{min}, T, \mathbf{v}_2) + I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) \\ &= I(\mathbf{z}_1^{min}, T) - I(\mathbf{z}_1^{min}, T|\mathbf{v}_2) + I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) \\ &= I(\mathbf{z}_1^{min}, T) + I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) \end{aligned}$$

3) $I(\mathbf{v}_1, T|\mathbf{v}_2) \geq I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) \geq 0$.

Applying the Data Processing Inequality (Cover & Thomas, 2006) to the Markov chain $T \to \mathbf{v}_1 \to \mathbf{z}_1^{suf}$, we have $I(\mathbf{v}_1, T) \geq I(\mathbf{z}_1^{suf}, T)$, so

$$\begin{aligned} I(\mathbf{v}_1, T|\mathbf{v}_2) &= I(\mathbf{v}_1, T) - I(\mathbf{v}_1, T, \mathbf{v}_2) \\ &= I(\mathbf{v}_1, T) - I(\mathbf{z}_1^{suf}, T, \mathbf{v}_2) \\ &\geq I(\mathbf{z}_1^{suf}, T) - I(\mathbf{z}_1^{suf}, T, \mathbf{v}_2) \\ &\geq I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) \geq 0 \end{aligned}$$

Combining these three equations, we can get the Theorem 1. □

The proof of Theorem 2.

Proof. According to Feder & Merhav (1994), the relationship between the Bayes error rate P_e and the conditional entropy $H(T|\mathbf{z}_1)$ is

$$-\ln(1 - P_e) \leq H(T|\mathbf{z}_1)$$

which is equivalent to

$$P_e \leq 1 - \exp[-H(T|\mathbf{z}_1)]$$

Applying the Lemma 2, for arbitrary learned representation $\mathbf{z}_1$, its Bayes error rate P_e satisfies

$$P_e \leq 1 - \exp[-(H(T) - I(\mathbf{z}_1, T|\mathbf{v}_2) - I(\mathbf{z}_1, \mathbf{v}_2, T))]$$

for the sufficient representation $\mathbf{z}_1^{suf}$, its Bayes error rate P_e^{suf} satisfies

$$P_e^{suf} \leq 1 - \exp[-(H(T) - I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) - I(\mathbf{v}_1, \mathbf{v}_2, T))]$$

for the minimal sufficient representation $\mathbf{z}_1^{min}$, its Bayes error rate P_e^{min} satisfies

$$P_e^{min} \leq 1 - \exp[-(H(T) - I(\mathbf{v}_1, \mathbf{v}_2, T))]$$

Note that $0 \leq P_e \leq 1 - 1/|T|$, so we use the threshold function $\Gamma(x) = \min\{\max\{x, 0\}, 1 - 1/|T|\}$ to prevent overflow. □

The proof of Theorem 3.

Proof. According to Frénay et al. (2013), when the conditional distribution $p(\varepsilon|\mathbf{z}_1)$ of estimation error ε is uniform, Laplace and Gaussian distribution, the minimum expected squared prediction error R_e becomes $\frac{1}{12}\exp[2H(T|\mathbf{z}_1)]$, $\frac{1}{2e^2}\exp[2H(T|\mathbf{z}_1)]$ and $\frac{1}{2\pi e}\exp[2H(T|\mathbf{z}_1)]$ respectively. Therefore, we unify them as

$$R_e = \alpha \cdot \exp[2H(T|\mathbf{z}_1)]$$

where α is a constant coefficient which depends on the conditional distribution $p(\varepsilon|\mathbf{z}_1)$. Applying the Lemma 2, for arbitrary learned representation $\mathbf{z}_1$, we have

$$R_e = \alpha \cdot \exp[2 \cdot (H(T) - I(\mathbf{z}_1, T|\mathbf{v}_2) - I(\mathbf{z}_1, \mathbf{v}_2, T))]$$

for the sufficient representation $\mathbf{z}_1^{suf}$, we have

$$R_e^{suf} = \alpha \cdot \exp[2 \cdot (H(T) - I(\mathbf{z}_1^{suf}, T|\mathbf{v}_2) - I(\mathbf{v}_1, \mathbf{v}_2, T))]$$

for the minimal sufficient representation $\mathbf{z}_1^{min}$, we have

$$R_e^{min} = \alpha \cdot \exp[2 \cdot (H(T) - I(\mathbf{v}_1, \mathbf{v}_2, T))]$$

□

B MORE EXPERIMENTS

In this section, we provide more experiments to support our work.

Table 4: Instance segmentation results on Cityscapes validation set, averaged over 5 random seeds.

Model	AP^{mk}	AP^{mk}_{50}	person	rider	car	truck	bus	train	mcycle	bicycle
Non Pre-training	27.8	53.7	29.7	23.3	50.6	19.5	45.9	21.6	14.8	17.0
SimCLR	28.5	55.5	**28.9**	22.5	**49.9**	22.0	47.5	22.9	16.6	17.6
SimCLR+RC (ours)	**29.4**	**57.2**	28.2	**22.8**	49.4	**24.5**	**49.4**	**25.4**	**17.9**	**17.7**
BYOL	28.3	55.1	**29.0**	**23.0**	**49.7**	25.5	44.6	22.5	15.7	16.7
BYOL+LBE (ours)	**29.2**	**55.5**	28.6	22.6	49.3	**26.2**	**47.8**	**25.8**	**16.8**	16.7

B.1 INSTANCE SEGMENTATION ON CITYSCAPES

Experimental setup. We train the models on TinyImageNet (Le & Yang, 2015), and fine-tune on Cityscapes (Cordts et al., 2016) for instance segmentation tasks. Cityscapes dataset has fine annotations for 2975 train, 500 val and 1525 test images. The instance segmentation task involves 8 object categories with different numbers of fine training instances:

Category	person	rider	car	truck	bus	train	mcycle	bicycle
Instance Number	17.9k	1.8k	26.9k	0.5k	0.4k	0.2k	0.7k	3.7k

For fine-tuning, we use Mask R-CNN (He et al., 2017) models with the ResNet-FPN-50 (Lin et al., 2017) backbone and fine-tune all layers end-to-end. We apply the default schedule from Detectron2 (Wu et al., 2019), except for using 36k iterations. For contrastive learning models, We train the models for 400 epochs with batch size 512 using LARS optimizer (You et al., 2017) and a cosine decay learning rate schedule. We set $\lambda_1 = \lambda_2 = 0.1$ for SimCLR and $\lambda_1 = \lambda_2 = 0.001$ for BYOL. The performance is measured by mask AP (averaged precision over IoU thresholds) and AP_{50} (mask AP at an IoU of 0.5).

Results. Since the test labels are not public, we report the results on Cityscapes validation set in Table 4 and the best result in each block is in bold. The results without pre-training are also provided for clear comparison. We report the best result between RC and LBE. On average, pre-training on TinyImageNet improves the performance of instance segmentation, and increasing $I(\mathbf{z}, \mathbf{v})$ can achieve further improvements. A main challenge of Cityscapes is training models in the low-data regime, especially for truck, bus, train and mcycle categories. Pre-training effectively alleviates this problem and our method has better performance. On the other hand, pre-training damages the performance on the categories (e.g., person, car and rider) which have sufficient training instances, and the reason may be that there is no similar categories to them in TinyImageNet and pre-training introduces harmful inductive bias.

Table 5: Downstream classification accuracy (%) on the source dataset (CIFAR10 or STL-10) and six transfer datasets.

Model	CIFAR10	DTD	MNIST	FaMNIST	CUBirds	VGGFlower	TrafficSigns
BarTwins	86.85	28.56	95.39	86.19	7.49	35.91	88.50
BarTwins+RC (ours)	**86.91**	28.97	96.60	86.72	7.90	38.94	90.92
BarTwins+LBE (ours)	86.38	**29.54**	**96.72**	**86.88**	**8.47**	**41.44**	**92.76**
Model	STL-10	DTD	MNIST	FaMNIST	CUBirds	VGGFlower	TrafficSigns
BarTwins	80.59	36.86	94.27	86.63	7.47	44.89	73.73
BarTwins+RC (ours)	**82.21**	36.97	94.45	86.71	7.89	46.31	78.94
BarTwins+LBE (ours)	81.13	**37.32**	**96.33**	**87.13**	**8.08**	**49.82**	**82.08**

B.2 More results in classification tasks

In the main text, we verify the effectiveness of increasing $I(\mathbf{z}, \mathbf{v})$ on two classic contrastive learning models: SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020a). SimCLR perfectly matches the contrastive learning framework, maximizing the lower bound estimate of the mutual information $I(\mathbf{z}_1, \mathbf{z}_2)$. BYOL avoids the dependence on the large amount of negative samples, and adopts the asymmetric structure and prediction loss. They both satisfy the characteristic that the views provide supervision information to each other, so they all tend to learn the minimal sufficient representation. We further verify the effect of increasing $I(\mathbf{z}, \mathbf{v})$ on Barlow Twins (Zbontar et al., 2021) which also satisfies this characteristic but uses a very different loss. It makes the cross-correlation matrix between the representations of different views as close to the identity matrix as possible. We use the same settings as in Experiment 4.1 and set $\lambda_1 = \lambda_2 = 1$. For STL-10, we use the *unlabeled* split for contrastive learning and the *train* and *test* split for linear evaluation.

The results are shown in Table 5 and the best result in each block is in bold. Increasing $I(\mathbf{z}, \mathbf{v})$ can improve the accuracy in downstream classification tasks of the learned representations in Barlow Twins, which indicates that our analysis results are applicable to various contrastive losses. In fact, our analysis mainly relies on the characteristic that the views provide supervision information to each other and all supervision information for one view comes from the other view.

B.3 Reconstructed samples

In order to show the effect of our reconstruction module more clearly, we provide the reconstructed images after training. As an example, we use SimCLR contrastive loss and take CIFAR10 as the training dataset. All experimental settings are the same as in Section 4.1. The classification accuracy is shown in Table 1, and the original and reconstructed images are shown in Figure 5. As we can see, the reconstructed images retain the shape or outline information in the original images, so as

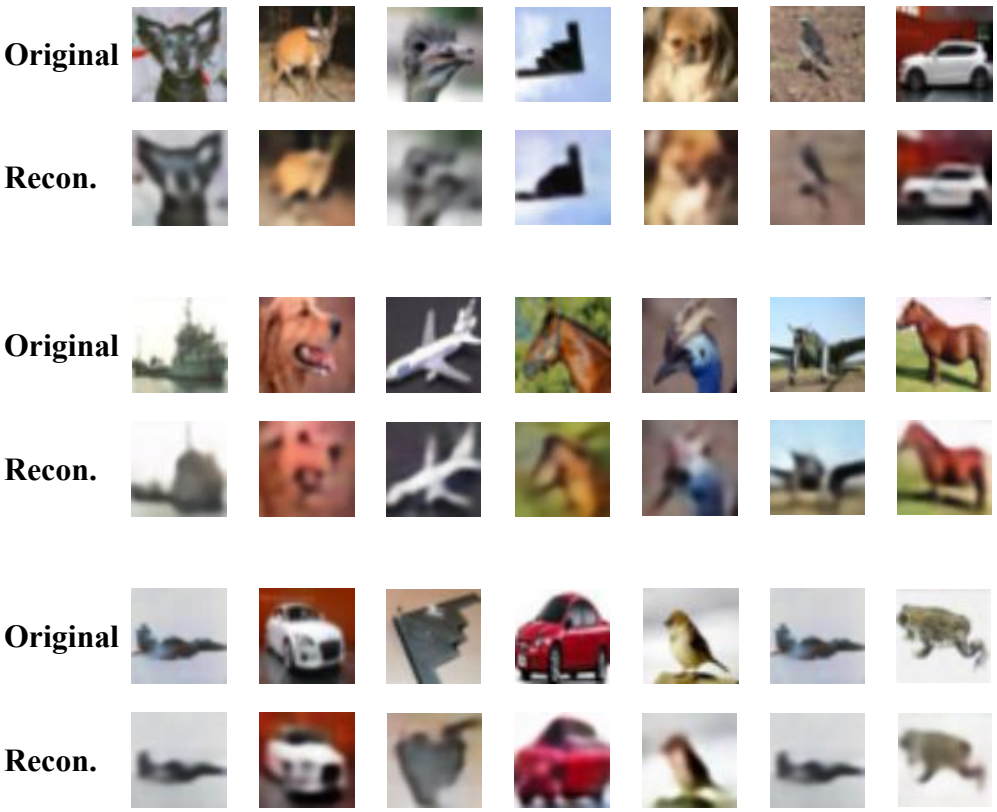


Figure 5: Demonstration of the effect of our reconstruction module. We provide the original images and the reconstructed images for comparison. We use SimCLR contrastive loss and take CIFAR10 as the training dataset.

the obtained representations. Since we use the mean square error loss to optimize the reconstruction module, the reconstructed images are blurry and this phenomenon is also observed in vanilla variational auto-encoder (Kingma & Welling, 2014).

C Derivation of L_{MIB} and L_{IP}

Federici et al. (2020) and Tsai et al. (2021) propose to eliminate the non-shared information between views to get the minimal sufficient representation. To this end, they propose their respective regularization term. Here we derive the specific form used in the Section 4.2.

In Federici et al. (2020), the regularization term is

$$\begin{aligned} L_{MIB} &= D_{SKL}(p(\mathbf{z}_1|\mathbf{v}_1)\|p(\mathbf{z}_2|\mathbf{v}_2)) \\ &= \frac{1}{2}\left(D_{\mathrm{KL}}(p(\mathbf{z}_1|\mathbf{v}_1)\|p(\mathbf{z}_2|\mathbf{v}_2)) + D_{\mathrm{KL}}(p(\mathbf{z}_2|\mathbf{v}_2)\|p(\mathbf{z}_1|\mathbf{v}_1))\right) \end{aligned}$$

According to the description in their paper and the official code [1], they model the two stochastic encoders $p(\mathbf{z}_1|\mathbf{v}_1)$ and $p(\mathbf{z}_2|\mathbf{v}_2)$ as

$$\begin{aligned} p(\mathbf{z}_1|\mathbf{v}_1) &= \mathcal{N}(\mathbf{z}_1; \boldsymbol{\mu}_1, \mathrm{diag}(\sigma_1^2)) \\ p(\mathbf{z}_2|\mathbf{v}_2) &= \mathcal{N}(\mathbf{z}_2; \boldsymbol{\mu}_2, \mathrm{diag}(\sigma_2^2)) \end{aligned}$$

where $\boldsymbol{\mu}_1(\mathbf{v}_1)$,$\sigma_1^2(\mathbf{v}_1)$,$\boldsymbol{\mu}_2(\mathbf{v}_2)$ and $\sigma_2^2(\mathbf{v}_2)$ are all functions of the input ($\mathbf{v}_1$ or $\mathbf{v}_2$), $\mathrm{diag}(\boldsymbol{e})$ creates a matrix in which the diagonal elements consist of vector $\boldsymbol{e}$ and all off-diagonal elements are zeros.

[1] https://github.com/mfederici/Multi-View-Information-Bottleneck

The regularization term has the closed form

$$L_{MIB} = \frac{1}{4}\sum_{i=1}^{d}\left[\frac{\sigma_1^{i2}}{\sigma_2^{i2}} + \frac{\sigma_2^{i2}}{\sigma_1^{i2}} + \frac{(\boldsymbol{\mu}_1^i - \boldsymbol{\mu}_2^i)^2}{\sigma_2^{i2}} + \frac{(\boldsymbol{\mu}_2^i - \boldsymbol{\mu}_1^i)^2}{\sigma_1^{i2}} - 2\right]$$

where d is the dimension of $\mathbf{z}_1$ and $\mathbf{z}_2$. We want to minimize L_{MIB}, and when $\sigma_1^2 = \sigma_2^2$, the term $\sigma_1^{i2}/\sigma_2^{i2} + \sigma_2^{i2}/\sigma_1^{i2}$ takes the minimum value 2, so the regularization term becomes

$$L_{MIB} = \frac{1}{2}\sum_{i=1}^{d}\frac{(\boldsymbol{\mu}_1^i - \boldsymbol{\mu}_2^i)^2}{\sigma_1^{i2}} \tag{22}$$

In practice, minimizing L_{MIB} makes the variance σ_1^2 and σ_2^2 very large, and the sampled representations change drastically and have very poor performance in downstream tasks. If the upper bound of the variance σ_1^2 and σ_2^2 is fixed, such as using the sigmoid activation function to limit it to $(0, 1)$, they will converge to the maximum value as the training progresses. Therefore, we might as well fix the variance and model the two stochastic encoders $p(\mathbf{z}_1|\mathbf{v}_1)$ and $p(\mathbf{z}_2|\mathbf{v}_2)$ as

$$p(\mathbf{z}_1|\mathbf{v}_1) = \mathcal{N}(\mathbf{z}_1; f_1(\mathbf{v}_1), \sigma^2\boldsymbol{I})$$
$$p(\mathbf{z}_2|\mathbf{v}_2) = \mathcal{N}(\mathbf{z}_2; f_2(\mathbf{v}_2), \sigma^2\boldsymbol{I})$$

where $\boldsymbol{I}$ is the identity matrix, σ^2 is the given variance, $f_i, i = 1, 2$ are deterministic encoders. This also guarantees a fair comparison with our second implementation. According to the Equation (22), the regularization term is equivalent to

$$L_{MIB} = \|f_1(\mathbf{v}_1) - f_2(\mathbf{v}_2)\|_2^2$$

We calculate the expectation of the regularization term on the data distribution $p(\mathbf{v}_1, \mathbf{v}_2)$ and get

$$L_{MIB} = \mathbb{E}_{p(\mathbf{v}_1,\mathbf{v}_2)}[\|f_1(\mathbf{v}_1) - f_2(\mathbf{v}_2)\|_2^2]$$

In Tsai et al. (2021), the author define the inverse predictive loss

$$L_{IP} = \mathbb{E}_{p(\mathbf{v}_1,\mathbf{v}_2)}[\|\mathbf{z}_1 - \mathbf{z}_2\|_2^2] = \mathbb{E}_{p(\mathbf{v}_1,\mathbf{v}_2)}[\|f_1(\mathbf{v}_1) - f_2(\mathbf{v}_2)\|_2^2]$$

D CHOICE OF MUTUAL INFORMATION LOWER BOUND ESTIMATE

In our second implementation, we need to use a mutual information lower bound estimate to calculate $I(\mathbf{z}, \mathbf{v})$ where $\mathbf{v}$ is the original input (e.g., images) and $\mathbf{z}$ is the representation (feature vector). We consider three candidate estimates:

1) The bound of Nguyen, Wainwright and Jordan (Nguyen et al., 2010)

$$\hat{I}_{NWJ}(\mathbf{z}, \mathbf{v}) = \mathbb{E}_{p(\mathbf{z},\mathbf{v})}[h(\mathbf{z}, \mathbf{v})] - \mathbb{E}_{p(\mathbf{z})p(\mathbf{v})}[e^{h(\mathbf{z},\mathbf{v})-1}] \tag{23}$$

2) MINE (Belghazi et al., 2018)

$$\hat{I}_{MINE}(\mathbf{z}, \mathbf{v}) = \mathbb{E}_{p(\mathbf{z},\mathbf{v})}[h(\mathbf{z}, \mathbf{v})] - \ln(\mathbb{E}_{p(\mathbf{z})p(\mathbf{v})}[e^{h(\mathbf{z},\mathbf{v})}]) \tag{24}$$

3) InfoNCE (Poole et al., 2019)

$$\hat{I}_{NCE}(\mathbf{z}, \mathbf{v}) = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\ln\frac{p(\mathbf{z}^k|\mathbf{v}^k)}{\frac{1}{N}\sum_{l=1}^{N}p(\mathbf{z}^l|\mathbf{v}^k)}\right] \tag{25}$$

where $(\mathbf{z}^k, \mathbf{v}^k), k = 1, \cdots, N$ are N copies of $(\mathbf{z}, \mathbf{v})$ and the expectation is over $\Pi_k p(\mathbf{z}^k, \mathbf{v}^k)$. As we can see, when we calculate the bound $\hat{I}_{NWJ}$ and $\hat{I}_{MINE}$, we need to calculate the critic $h(\mathbf{z}, \mathbf{v})$ between the representation $\mathbf{z}$ and original input $\mathbf{v}$. If we use a neural network to model the critic $h(\mathbf{z}, \mathbf{v})$, we have to take the original input and the representation together as the input of the neural network. Since the distribution of the original input $\mathbf{v}$ and the representation $\mathbf{z}$ is quite different, it is very difficult. Therefore, we use the InfoNCE lower bound estimate.