Revising Mutual Information by Rethinking Minimal Sufficient Representation in Contrastive Learning

Project Report submitted by

P.VignaTej Reddy (420217)

Ch.Prashanth kumar (420121)

M.Viswa Chandra(420210)

Under the supervision of

Mr.Y.Gireesh

DEPARTMENT OF
COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
ANDHRA PRADESH

MAY 2023

CERTIFICATE

This is to certify that the project titled **Revising Mutual Information by Rethinking Minimal Sufficient Representation in Contrastive Learning** is a bonafide record of the work done by

P.VignaTej Reddy (420217)

Ch.Prashanth kumar (420121)

M.Vishwa Chandhra(420210)

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **COMPUTER SCIENCE** of the **NATIONAL INSTITUTE OF TECHNOLOGY, Tadepalligudeum**, during the year 2020-2024.

Mr.Y.Gireesh	**Dr.K.Hima Bindu**
Project Guide	Head of the Department

Project Presentation held on _____

ABSTRACT

In the discipline of self-supervised representation learning, contrastive learning between various views of the data achieves amazing success, and the learnt representations are helpful in a variety of downstream tasks such as classification, detection and segementation etc . Contrastive learning roughly achieves the minimal sufficient representation that incorporates the shared information and excludes the non-shared information between views because all supervisory information for one view originates from the other view. Given the variety of the downstream activities, it cannot be ensured that all information pertinent to the task is transmitted between views.As a result, we make the assumption that non-shared task-relevant information cannot be disregarded and theoretically demonstrate that the minimal sufficient representation in contrastive learning is insufficient for the downstream tasks, which results in performance loss.From our study, We propose a solution for this problem .The possibility of the contrastive learning models over-fitting to the shared information between perspectives is revealed, which is a new issue. As a workaround for this issue, as we are unable to use any downstream task information during training, we suggest increasing the mutual information between the representation and input as a regularisation to roughly provide more task-relevant information. Our experiments support the validity of our analysis and the efficiency of our approach. It considerably raises the efficiency of a number of conventional contrastive learning models in downstream tasks.

ACKNOWLEDGEMENT

We would like to thank the following people for their support and guidance without whom the completion of this project in fruition would not be possible.

Mr.Y.Gireesh, our project guide, for helping us and guiding us in the course of this project .

Dr.K.Hima Bindu, the Head of the Department, Department of COMPUTER SCIENCE.

Our internal reviewers, **Dr.K.Hima Bindu** , **Mrs K.Sindhu** for their insight and advice provided during the review sessions.

We would also like to thank our individual parents and friends for their constant support.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

In the field of self-supervised representation learning, contrastive learning between several interpretations of the data obtains excellent results. In reality, the learnt representations are helpful for a variety of downstream tasks such segmentation, detection, and classification. A sufficient representation in contrastive learning is one that includes all shared information between perspectives, whereas a minimal sufficient representation only includes the shared information and leaves out the non-shared information. Through contrastive learning, a sufficient representation is obtained by maximising the mutual information between the representations of various views. Additionally, since all supervision data for one view is derived from the other view , non-shared data is frequently disregarded, resulting in a roughly accurate representation. After the minimal suitable representation is acquired, the best perspectives for contrastive learning rely on the subsequent tasks. Therefore, the information required for task T1 is different from Task T2. Moreover we don't know task at the time of training. In this study, we formalise this hypothesis and make an assumption that the information relevant to non-shared tasks cannot be ignored. On the basis of this assumption, we theoretically demonstrate that performance reduction happens because of the minimal sufficient representations' non-ignorable gap with the optimal representation, which contains less task-relevant information than other sufficient representations. In this report, for understanding, we take two kinds of tasks, named classification and regression and we will demonstrate that minimal sufficient representations lowest fessible error is higher than other sufficient representations.We take use of SimCLR model to show that learned rep-

resentations in contrastive learning are insufficient for some downstream tasks where task relavent information is not exchanged between views.The heavily trained contrastive learning models have backlash of running at the risk of becoming overly adapted to the information that is transmitted across views. In order to mitigate this, we must add extra information to the representations that is unshared task relevent information.

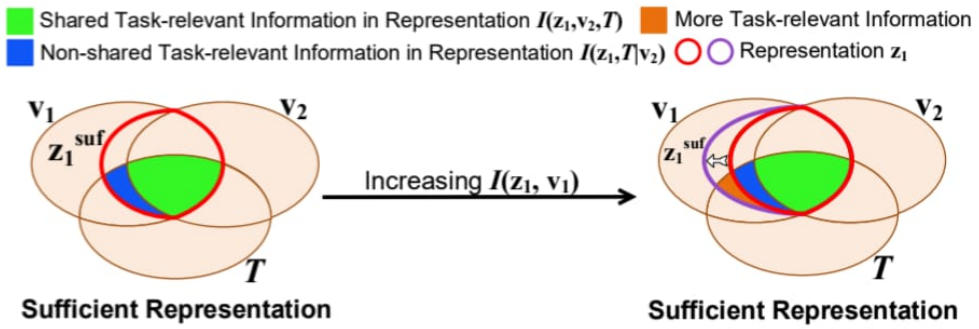


Figure 1: Reqiured diagram for our study. Increasing I(z1, v1) approximately introduces additional non-shared task-relevant information based on the (approximately minimal) adequate representation learned in contrastive learning.

It is not possible to accomplish this directly at the training stage because we don't know information about the downstream tasks. As alternative solution, we provide an objective that roughly introduces more task-relevant information by increasing the mutual information between the representation and input. Information visualisations are used in Fig. 1 to illustrate this motivation. To improve the mutual information, we take into consideration two implementations. The first creates representations of the input that contain the essential details about the input through the reconstruction of the input. The second one is based on a high-dimensional estimation of mutual information, it makes use of lower bound estimates of mutual information. Overall, we say the following: Models run the risk of becoming overly adapted to the information that is transmitted across views. To mitigate this, we must add more information to the representation that is relevant to separate tasks. This study conceptually demonstrates that contrastive learning runs the risk of being overly adapted to the mutual information across views. We offer a thorough analysis based on the internal contrastive learning

process in which the views exchange supervision data. When the downstream task information is unavailable, we suggest increasing the mutual information between the representation and input to roughly introduce more task-relevant information, as shown in Fig. 1. This will help to alleviate the problem.We test the efficiency of our approach for classification, detection, and segmentation tasks using SimCLR, BYOL, and Barlow Twins.We provide some analytical tests to help other you understand our model, hypothesis and assumptions.

Chapter 2

Review of Literature

Contrastive learning is an effective and sucessfull approach for self-supervised representation learning. It works as comparative learning between various interpretations(views) of the data. The views are created by utilising the unlabeled data's structure, such as local patches and the entire image, various augmentations (argumentation means changing the form of the data to an relevant form,i.e for an image, we do cropping, rotation, color changes etc) of the same image or video and text pairs. The objective is to maximize the similarity between the positive pair and minimize the similarity between the positive pair and a set of negative pairs. Contrastive learning's main application is that it can learn high-quality representations from unlabeled data, which can be used to improve the performance of downstream tasks that require labeled data. In real world, labeled data is often more scarce and expensive to obtain than unlabeled data, as it requires human annotation or expert knowledge to generate accurate labels. many of ML models till date relied on labelled data. To address the challenge of limited labeled data and having advantage of vast unlabelled data, researchers have developed various techniques for self-supervised and unsupervised learning techniques and one them is contrastive learning. Here we have used SimCLR and BYOL, which are two popular self-supervised learning techniques in the field of computer vision. BYOL is based on the idea of a ”predictor” network and a ”target” network. The predictor network takes an input image and produces a low-dimensional latent vector, which is used to predict the latent vector produced by the target network on the same image. The target network is a copy of the predictor network that is updated using a slow-moving exponential

moving average of the weights of the predictor network. BYOL uses contrastive loss function to learn to predict the latent vectors produced by the target network. BYOL has advantage of simplicity with high performance and efficiency, But it is sensitive to hyperparameters and there is trade off between accuracy and efficiency in BYOL. SimCLR is a self-supervised learning technique for training deep neural networks to learn high-quality representations from unlabeled data. The contrastive loss function used in SimCLR is based on the InfoNCE (Normalized Mutual Information Estimation) loss, which encourages the model to learn to maximize the mutual information between the positive pairs while minimizing the mutual information between the positive and negative pairs. SimCLR has advantage of high performance and robustness to hyper-parameters, but it requires large dataset and hefty computational resources. Assuming minimal sufficient representation, researchers recently discovered that the best views for contrastive learning vary based on the task. According to recent research, under the premise of minimal sufficient representation, the best views for contrastive learning depend on the task at hand. In other words, views optimal for a downstream task may be irrelevent for other downstream task. Theoretically analysing this discovery, we find that the contrastive learning models may over-fit to the shared information between views. To address this issue, we therefore increase the mutual information between representation and input. Minimum sufficient representation is learned by some recent works. They assume that practically all information relevant to tasks is shared between views, which is an overly idealistic assumption and conflicts with our discovery.We simplify recently proposed contrastive SimCLR algorithm with resnets(i.e. resnet18) without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of this model. In other words, even if the provided views are ideal for some jobs that come after them, they might not be appropriate for all tasks. Theoretically analysing this finding, we find that the contrastive learning models may over-fit to the shared information between views. To address this issue, we therefore suggest increasing the mutual information between representation and input . Our

study demonstrates that the learnt representation in contrastive learning is insufficient for the downstream tasks. To be sufficient, we must include extra information that is pertinent to the task.

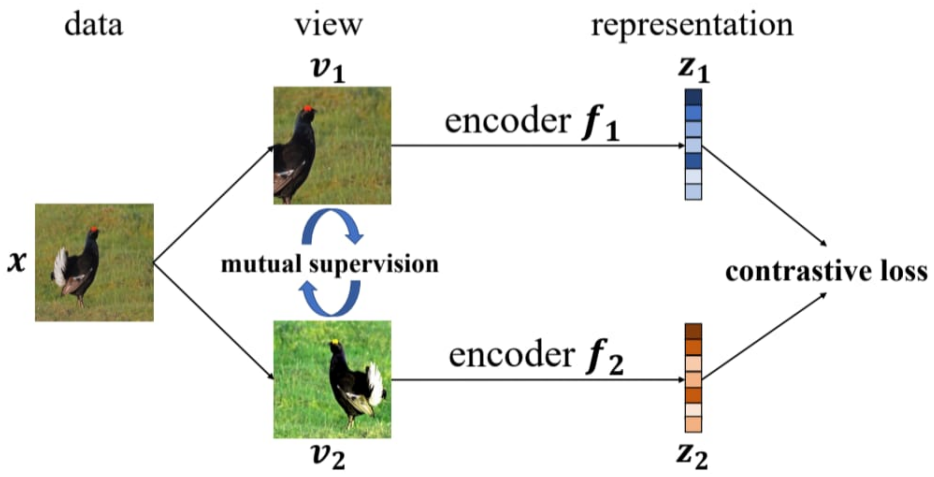


Figure 2. Internal mechanism of contrastive learning: the views provide supervision information to each other.

In other words, even if the provided views are ideal for some jobs that come after them, they might not be appropriate for all tasks. Theoretically analysing this finding, we find that the contrastive learning models may over-fit to the shared information between views. To address this issue, we therefore suggest increasing the mutual information between representation and input .

In order to ensure sufficiency, a model based on the information bottleneck theory captures all task-relevant information during the first learning phase (the ”drift phase”), and then compresses it during the second learning phase (the ”diffusion phase”). Our study demonstrates that the learnt representation in contrastive learning is in the drift phase and is insufficient for the downstream tasks. To be sufficient, we must include extra information that is pertinent to the task.

Chapter 3

Theoretical analysis and model

In this part, we first discuss the foundations of contrastive learning, and then we will conceptually examine the drawbacks of contrastive learning's minimal sufficient representation, and at last, we will provide an approach for adding additional task-relevant information to the representations. As a result, the foundation of our theoretical analysis is the assumption that the information contained in the representations is given in the most appropriate manner.

3.1 CONTRASTIVE LEARNING

Contrastive learning, an unsupervised learning model, is a general framework, which maximizes the mutual information between the representations of two random variables v_1 and v_2 with the joint distribution $p(v_1, v_2)$

$$\max_{f_1, f_2} I(z_1, z_2)$$

where $z_i = f_i(v_i), i = 1, 2$ are also random variables and $f_i, i = 1, 2$ are encoding functions. In practice, two views of the data x are usually representated as v_1 and v_2. Where v_1 and v_2 have the same marginal distributions $(p(v_1) = p(v_2))$, the function f_1 and f_2 can be the same or different $(f_1 = f_2)$.

3.1.1 Definition 1

(Sufficient Representation in Contrastive Learning) The representation z_1^{suf} of v_1 is sufficient for v_2 if and only if $I\left(z_1^{\text{suf}}, v_2\right) = I(v_1, v_2)$.

The sufficient representation z_1^{suf} of v_1 keeps all the information about z_2 in v_1 In other words, z_1^{suf} contains all the shared information between v_1 and v_2, i.e., $I\left(v_1, v_2 \mid z_1^{suf}\right) = 0$. Symmetrically, the sufficient representation z_2^{suf} of v_2 for v_1 satisfies $I\left(v_1, z_2^{\text{suf}}\right) = I\left(v_1, v_2\right)$

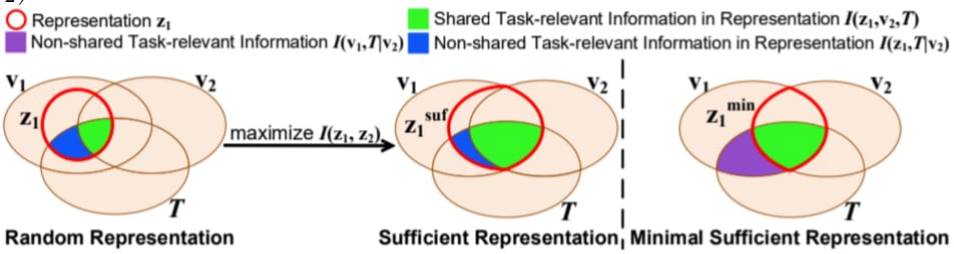


Figure 3. Information diagrams of different representations in contrastive learning. We consider the situation where the non-shared task-relevant information $I\left(v_1, T \mid v_2\right)$ cannot be ignored. Contrastive learning makes the representations extracting the shared information hetween views to obtain the sufficient representation which is approximately minimal. The minimal sufficient representation contains less task-retevant information-frem the input than other sufficient representations.

3.1.2 Definition 2

(Minimal Sufficientnt. Representation in Contrastive Learning) The sufficient representation z_1^{min} of v_1 is minimal if and only if $I\left(z_1^{\text{min}}, v_1\right) \in I\left(z_1^{suf}, \overline{v_1}\right), \forall z_1^{suf}$ that is sufficient

Among all sufficient representations, the minimal sufficient representation z_1^{min} contains the least information about v_1. Further, it is usually assumed that z_1^{min} only contains the shared information between views and eliminates other non-shared information, i.e., $I\left(z_1^{min}, v_1 \mid v_2\right) = 0$.

$$I\left(v_1, v_2\right) \geq I\left(v_1, z_2\right) \geq I\left(z_1, z_2\right)$$

i.e., $I\left(v_1, v_2\right)$ is the upper bound of $I\left(z_1, z_2\right)$. Considering that $I\left(v_1, v_2\right)$ remains unchanged during the optimization process, contrastive learning optimizes the functions f_1 and f_2 so that $I\left(z_1, z_2\right)$ approximates $I\left(v_1, v_2\right)$. When these functions have enough capacity and are well learned based on sufficient data, we can assume $I\left(z_1, z_2\right) =$

$I(v_1, v_2)$, which means the learned representations in contrastive learning are sufficient. They are also approximately minimal since all supervision information comes from the other view. Therefore, the shared information controls the properties of the representations We introduce a random variable T to represent the information needed for a downstream task, which can be a classification, regression, or clustering work. The learnt representations in contrastive learning are often employed in various downstream tasks. According to reasearch, under the supposition of minimal sufficient representation, the best views for contrastive learning depend on the task. Since different downstream tasks require different information that is unknown during training, this discovery makes sense.

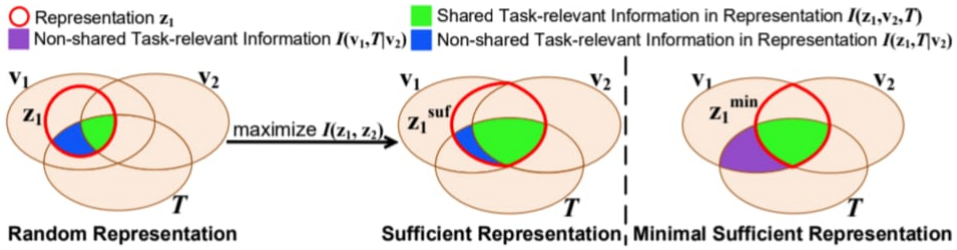


Figure 3. Information diagrams of different representations in contrastive learning. We consider the situation where the non-shared task-relevant information $I(v_1, T \mid v_2)$ cannot be ignored. Contrastive learning makes the representations extracting the shared information hetween views to obtain the sufficient representation which is approximately minimal. The minimal sufficient representation contains less task-retevant information-frem the input than other sufficient representations.

It is difficult for the given views to share all the information required by 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., $I(v_1, T \mid v_2)$ cannot be ignored. Eliminating all non-shared information has the risk of damaging the performance of the representations in the downstream tasks

3.2 Examination of minimal sufficient representation

Because it completely eliminates the non-shared information between views, which may be crucial for some downstream tasks, the minimal sufficient representation is not a good choice for downstream tasks. We formalise this issue and provide theoretical support for the claim that, in contrastive learning, the minimal sufficient representation should do less well in the subsequent tasks than other sufficient representations. Without loss of generality, considering the symmetry between v_1 and v_2, we take v_2 as the supervision signal for v_1 and take v_1 as the input of a task. It is generally believed that the more task-relevant information contained in the representations, the better performance can be obtained. Therefore, we examine the task-relevant information contained in the representations.

3.2.1 Task-Relevant Information in Representations

In contrastive learning, for a downstream task T, the minimal sufficient representation z_1^{min} contains less task-relevant information from input v_1 than other sufficient representation z_1^{suf}, and $I\left(z_1^{\text{min}}, T\right)$ has a gap of $I\left(v_1, T \mid v_2\right)$ with the upper bound $I\left(v_1, T\right)$. Formally, we have

$$
\begin{aligned}
& I\left(v_1, T\right)=I\left(z_1^{\text{min}}, T\right)+I\left(v_1, T \mid v_2\right) \\
\geq & I\left(z_1^{\text{suf}}, T\right)=I\left(z_1^{\text{min}}, T\right)+I\left(z_1^{\text{suf}}, T \mid v_2\right) \\
\geq & I\left(z_1^{\text{min}}, T\right)
\end{aligned}
$$

Theorem 1 indicates that z_1^{suf} can have better performance in task T than z_1^{min} because it contains more taskrelevant information. When non-shared task-relevant information $I\left(v_1, T \mid v_2\right)$ is significant, z_1^{min} has poor performance because it loses a lot of useful information. See Fig. 3 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 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 z_1 and $\widehat{T}$ be the prediction for T based on z_1, we have $P_e = 1-\mathbb{E}_{p(z_1)}\left[\max_{t\in T} p\left(\widehat{T}=t \mid z_1\right)\right]$ 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.

3.2.2 Bayes error for classification Task

For arbitrary learned representation z_1, its Bayes error rate $P_e = \Gamma\left(\bar{P}_e\right)$ with

$$\bar{P}_e \leq 1 - \exp\left[-\left(H(T) - I\left(z_1, T \mid v_2\right) - I\left(z_1, v_2, T\right)\right)\right]$$

Specifically, for sufficient representation z_1^{suf}, its Bayes error rate $P_e^{\text{suf}} = \Gamma\left(\bar{P}_e^{\text{suf}}\right)$ with

$$\bar{P}_e^{\text{suf}} \leq 1 - \exp\left[-\left(H(T) - I\left(z_1^{\text{suf}}, T \mid v_2\right) - I\left(v_1, v_2, T\right)\right)\right]$$

for minimal sufficient representation $z_1^{\min}$, its Bayes error rate $P_e^{min} = \Gamma\left(\bar{P}_e^{min}\right)$ with

$$\bar{P}_e^{\min} \leq 1 - \exp\left[-\left(H(T) - I\left(v_1, v_2, T\right)\right)\right]$$

Since $I\left(z_1^{\text{suf}}, T \mid v_2\right) \geq 0$, Theorem 2 indicates for classification task T, the upper bound of $P_e^{\min}$ is larger than P_e^{suf}. In other words, $z_1^{\min}$ is expected to obtain a higher classification error rate in the task T than z_1^{suf}. According to the Eq. (5), considering that $H(T)$ and $I\left(v_1, v_2, T\right)$ are not related to the representations, increasing $I\left(z_1^{\text{suf}}, T \mid v_2\right)$ can reduce the Bayes error rate in classification task T. When $I\left(z_1^{\text{suf}}, T \mid v_2\right) = I\left(v_1, T \mid v_2\right), z_1^{\text{suf}}$ contains all the useful information for task T in 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 z_1, we consider the smallest achievable expected squared prediction error $R_e = \min_{\widetilde{T}} \mathbb{E}\left[\left(T - \widetilde{T}\left(z_1\right)\right)^2\right] = \mathbb{E}\left[\varepsilon^2\right]$ with $\varepsilon\left(T, z_1\right) = T - \mathbb{E}\left[T \mid z_1\right]$

3.2.3 Minimum Expected Squared Prediction Error for regression Task

For arbitrary learned representation z_1, when the conditional distribution $p(\varepsilon \mid z_1)$ is uniform, Laplacian or Gaussian distribution, the minimum expected squared prediction error R_e satisfies

$$R_e = \alpha \cdot \exp\left[2 \cdot \left(H(T) - I\left(z_1, T \mid v_2\right) - I\left(z_1, v_2, T\right)\right)\right]$$

Specifically, for sufficient representation z_1^{suf}, its minimum expected squared prediction error R_e^{suf} satisfies

$$R_e^{\text{suf}} = \alpha \cdot \exp\left[2 \cdot \left(H(T) - I\left(z_1^{\text{suf}}, T \mid v_2\right) - I\left(v_1, v_2, T\right)\right)\right]$$

for minimal sufficient representation z_1^{min}, its minimum expected squared prediction error R_e^{min} satisfies

$$R_e^{\min} = \alpha \cdot \exp\left[2 \cdot \left(H(T) - I\left(v_1, v_2, T\right)\right)\right]$$

where the constant coefficient α depends on the conditional distribution $p(\varepsilon \mid z_1)$.

The assumption about the 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 $\overline{\text{come}}$ from simple distributions (e.g., Gaussian distribution) in statistical learning theory. Similar to the classification tasks, Theorem 3 indicates that for regression tasks, z_1^{suf} can achieve lower expected squared prediction error than z_1^{min} and increasing $I\left(z_1^{\text{suf}}, T \mid v_2\right)$ can improve the performance.

Theorem 2 and Theorem 3 analyze the disadvantages of the minimal sufficient representation $z_1^{\min}$ in classification tasks and regression tasks respectively. The essential reason is that $z_1^{\min}$ has less task-relevant information than z_1^{suf} and has a non-ignorable gap $I\left(v_1, T \mid v_2\right)$ with the optimal representation, as shown in Theorem 1.

3.2.4 More 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 nonshared task-relevant information. Moreover, contrastive learning approximately learns the minimal sufficient representation, 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 from v_1, i.e., increasing $I(z_1, T \mid v_2)$. However, we cannot utilize any downstream task information during training, so it is impossible to increase $I(z_1, T \mid v_2)$ directly. We consider increasing $I(z_1, v_1)$ as an alternative because the increased information from v_1 in z_1 may be relevant to some downstream tasks, and this motivation is demonstrated in Fig. 1. In addition, increasing $I(z_1, v_1)$ also helps to extract the shared information between views at the beginning of the optimization process. Concretely, considering the symmetry between v_1 and v_2, our optimization objective is

$$\max_{f_1, f_2} I(z_1, z_2) + \sum_{i=1}^{2} \lambda_i I(z_i, v_i)$$

which consists of the original optimization objective Eq. (1) in contrastive learning and the objective terms we proposed. The coefficients λ_1 and λ_2 are used to control the amount of increasing $I(z_1, v_1)$ and $I(z_2, v_2)$ respectively. For optimizing $I(z_1, z_2)$, we adopt the commonly used implementations in contrastive learning models [7, 17, 51]. For optimizing $I(z_i, v_i), i = 1, 2$, we consider two implementations. Implementation I Since $I(z, v) = H(v) - H(v \mid z)$ and $H(v)$ is not related with z, we can equivalently decrease the conditional entropy $H(v \mid z) = -\mathbb{E}_{p(z,v)}[\ln p(v \mid z)]$. Concretely, we use the representation z to reconstruct the original input v, as done in auto-encoder models. Decreasing the entropy of reconstruction encourages the representation z to contain more information about the original input v. However, the conditional distribution $p(v \mid z)$ is intractable in practice, so we use $q(v \mid z)$ as an approximation and get $\mathbb{E}_{p(z,v)}[\ln q(v \mid z)]$, which is the lower bound of $\mathbb{E}_{p(z,v)}[\ln p(v \mid z)]$. We can increase $\mathbb{E}_{p(z,v)}[\ln q(v \mid z)]$ as an alternative objective. According to the type of input v (e.g.,

images, text or audio), $q(v \mid z)$ 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 z. For example, when $q(v \mid z)$ is the Gaussian distribution $\mathcal{N}\left(v; \mu(z), \sigma^2 I\right)$ with given variance σ^2 and deterministic mean function $\mu(\cdot)$ which is usually parameterized by neural networks, we have

$$\mathbb{E}_{p(z,v)}[\ln q(v \mid z)] \propto -\mathbb{E}_{p(z,v)}\left[\|v - \mu(z)\|_2^2\right] + c$$

where c is a constant to representation z. The final optimization objective is

$$\max_{f_1, f_2, \mu} I\left(z_1, z_2\right) - \sum_{i=1}^{2} \lambda_i \mathbb{E}_{p(z_i, v_i)}\left[\|v_i - \mu_i\left(z_i\right)\|_2^2\right]$$

Implementation II Although the above implementation is effective and preferred in practice, it needs to reconstruct the input, which is challenging for complex input and introduces more 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, the bound of Nguyen, Wainwright and Jordan, MINE and InfoNCE . We choose the InfoNCE lower bound and the detailed discussion is provided in Appendix. Concretely, the InfoNCE lower bound is

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

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

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

where ρ is a scale factor. In fact, it pushes the representations away from each other to increase $H(z)$, which can increase mutual information $I(z,v)$ since $I(z,v) = H(z) - H(z \mid v) = H(z) - \frac{d}{2}\left(\ln 2\pi + \ln \sigma^2 + 1\right)$ with d being representation dimension. It also be denoted as uniformity property. The final optimization objective is

$$\max_{f_1,f_2} I\left(z_1, z_2\right) + \sum_{i=1}^{2} \lambda_i \tilde{I}_{NCE}\left(z_i, v_i\right)$$

Since the objective term Eq. (14) is calculated at the representation-level, when we use the convolutional neural networks (e.g., ResNet) to parameterize f, it can be applied to the output activation of multiple internal blocks.

Discussion. It is worth noting that increasing $I(z,v)$ does not conflict with the information bottleneck theory. 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 but not to compress it. On the other hand, we cannot introduce too much information from the input v either, which may contain harmful noise. Here we use the coefficients λ_1 and λ_2 to control this.

Chapter 4

Experiments

In this section, we first verify the effectiveness of increasing $I(z, v)$ on CIFAR10 dataset, and then provide some analytical experiments. We choose a classic contrastive learning models as our baseline: SimCLR. We denote our first implementation as "RC" for "Re-Construction" and the second implementation as "LBE" for "Lower Bound Estimate". For all experiments, we use random cropping, flip and random color distortion as the data augmentation, as suggested by. For "LBE", we set $\sigma = 0.1$ and $\rho = 0.05$.

4.0.1 Effectiveness of increasing $I(z, v)$

We consider various types of the downstream task, including classification, regression, detection and segmentation tasks. The results of SimCLR are provided in this report.

Pretraining

We train the models on CIFAR10. For CIFAR10, we use the ResNet18 backbone and the models are trained for 30 epochs with batch size 256 using Adam optimizer with learning rate 3e-4.

Linear evaluation

We follow the linear evaluation protocol where a linear classifier is trained on top of the frozen backbone. The linear evaluation is conducted on the source dataset and several transfer datasets: CIFAR100, CIFAR10. The linear classifier is trained for 100 epochs using SGD optimizer. Table 1 shows the results on CIFAR10, STL-10 and ImageNet, and the best result in each block is in bold. Our implemented results of the

baselines are consistent with. Increasing $I(z, v)$ 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 $I(z, v)$ achieves significant improvement. In other words, increasing $I(z, v)$ can prevent the models from overfitting to the shared information between views. What's more, it is effective for various contrastive learning models, which means our analysis results are widely applicable in contrastive learning. In fact, they all satisfy the internal mechanism.

ACCURACIES	
MODEL	CIFAR-10
SimCLR	76.91
SimCLR + RC (ours)	77.31
SimCLR + LBE (ours)	76.74

We conduct dual object detection on CIFAR10 using faster R-CNN, and detection and instance segmentation on COCO using Mask R-CNN , following given setup. All methods use the R50-C4 backbone that is initialized using the ResNet50 pre-trained on ImageNet. Increasing $I(z, v)$ significantly improves the precision in object detection and instance segmentation tasks. These dense prediction tasks require some local semantic information from the input. Increasing $I(z, v)$ can make the representation z contain more information from the input v which may not be shared between views, thereby obtaining better precision.

4.0.2 Analytical experiments

We provide some analytical experiments to further understand our hypotheses, theoretical analysis and models.

Eliminating non-shared information

Some recent works propose to eliminate the non-shared information between views in the representation to get the minimal sufficient representation. To this end, let us, minimize the regularization term

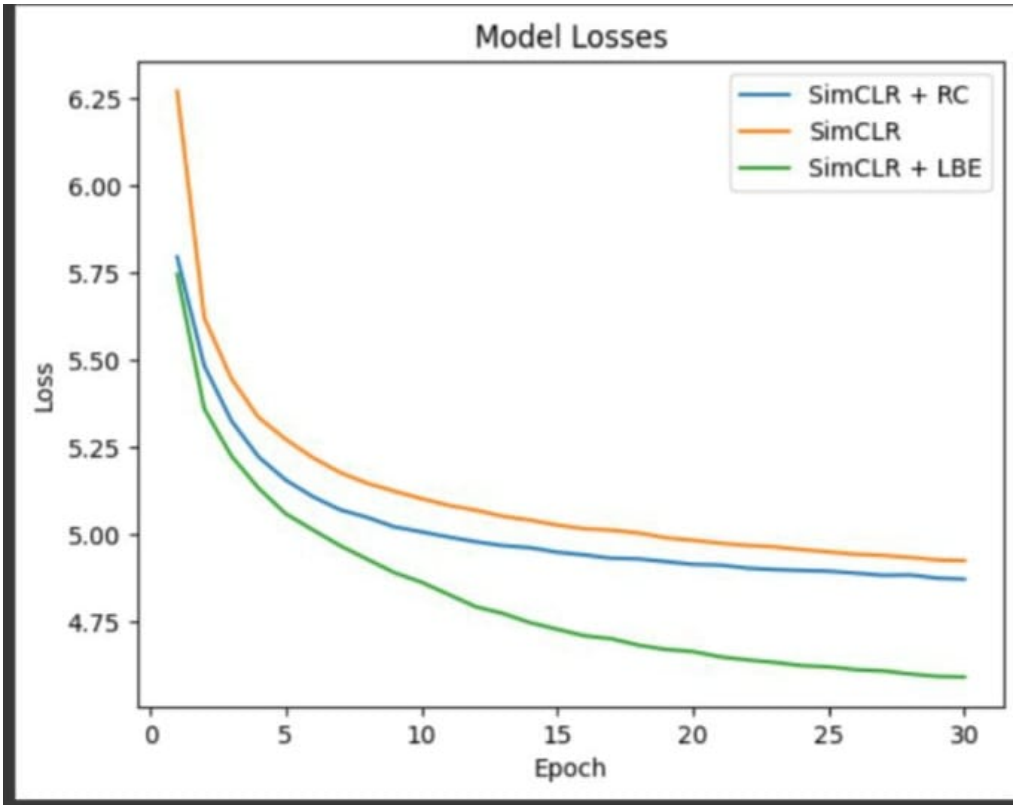
$$L_{MIB} = \frac{1}{2}\left[KL\left(p\left(z_1 \mid v_1\right) \| p\left(z_2 \mid v_2\right)\right) + KL\left(p\left(z_2 \mid v_2\right) \| p\left(z_1 \mid v_1\right)\right)\right]$$

where $KL(\cdot\|\cdot)$ represents the Kullback-Leibler divergence. When $p\left(z_1 \mid v_1\right)$ and $p\left(z_2 \mid v_2\right)$ are modeled as $\mathcal{N}\left(z_i; f_i\left(v_i\right), \sigma^2 I\right), i = 1, 2$ with given variance σ^2, it can be rewritten as $L_{MIB} = \mathbb{E}_{p(v_1,v_2)}\left[\|f_1\left(v_1\right) - f_2\left(v_2\right)\|_2^2\right]$. Identically, let us minimize the inverse predictive loss $L_{IP} = \mathbb{E}_{p(v_1,v_2)}\left[\|f_1\left(v_1\right) - f_2\left(v_2\right)\|_2^2\right]$. The detailed derivation will be provided. We evaluate these two regularization terms in the linear evaluation tasks and choose their coefficient with best accuracy on the source dataset. The results are shown and the best result in each block is highlighted. Although these two regularization terms have the same form, L_{MIB} 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 cannot change the accuracy in downstream classification tasks much. This means that the sufficient representation learned in contrastive learning is approximately minimal and we don't need to further remove the non-shared information.

Changing the amount of increasing $I(z, v)$. Quantifying the mutual information between the high-dimensional variables is very difficult, and often leads to inaccurate calculation in practice . Therefore, we assume that the hyper-parameters λ_1 and λ_2 control the amount of increasing $I\left(z_1, v_1\right)$ and $I\left(z_2, v_2\right)$ respectively. Larger λ_1 is expected to increase $I\left(z_1, v_1\right)$ more, so as λ_2. 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 Fig. 4. We report the accuracy on the source dataset (CIFAR10 or STL-10) and the averaged accuracy on all transfer datasets. As we can see, increasing $I(z, v)$ 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 $I(z, v)$ may introduce noise beside useful information.

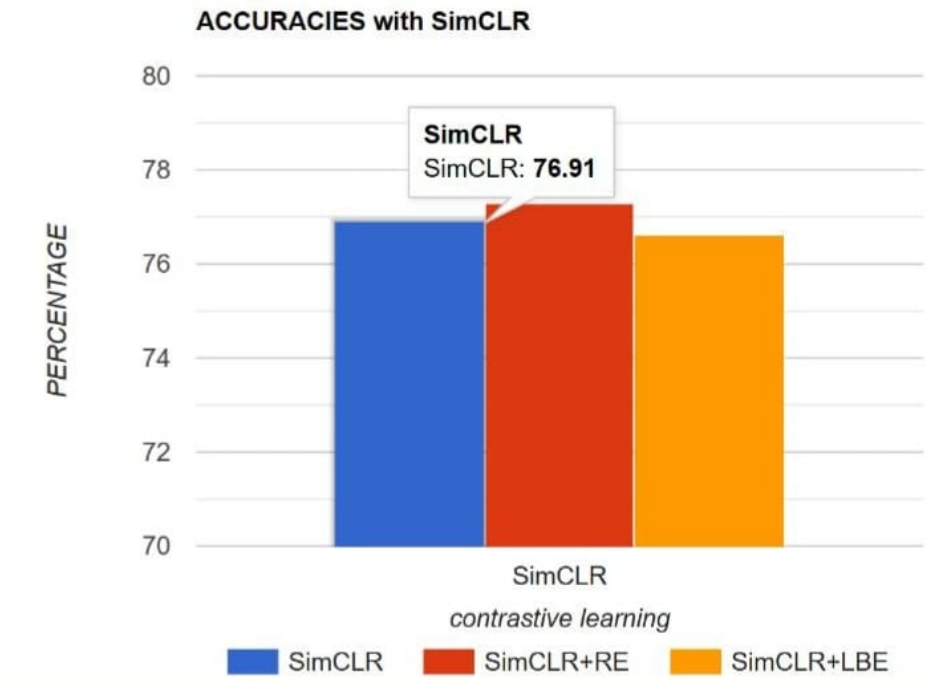


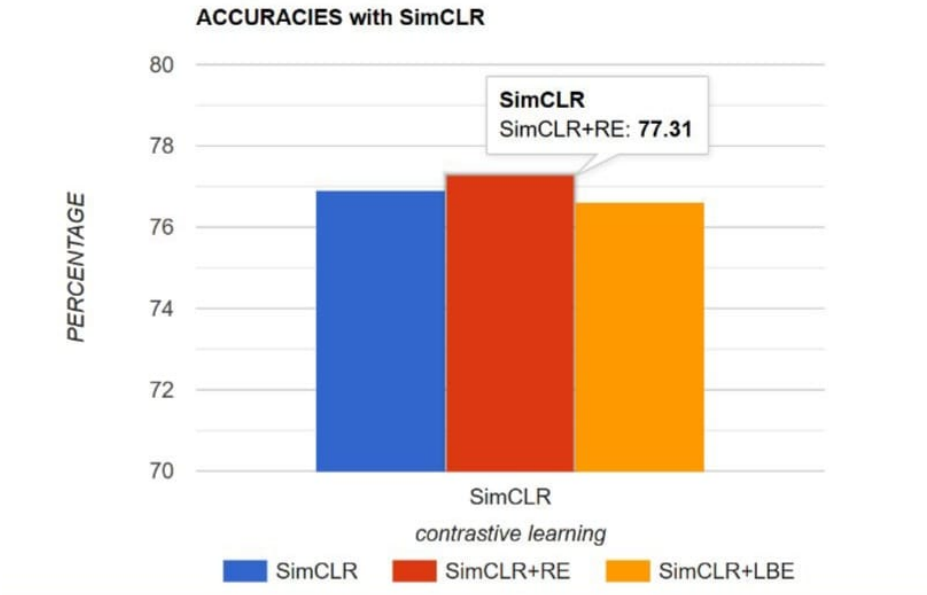
Training with more epochs. In the above experiments, we train all models for 30 epochs. Here we further show the behavior of the contrastive learning models and increasing $I(z, v)$ when training with more epochs. We choose SimCLR as the baseline and train all models for 100, 200, 300, 400, 500 and 600 epochs. The results are shown in the refrence[1]. With more training epochs, the learned representations in contrastive learning are more approximate to the minimal sufficient representation which mainly contain the shared information between views and ignore the nonshared information. For the classification tasks on the transfer datasets, the shared information between views is not sufficient. As shown in above Figures, the accuracy on the transfer datasets decreases with more epochs and the learned representations over-fit to the shared information between views. Increasing $I(z, v)$ 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. As shown in Figures below, the accuracy on CIFAR10 increases with more epochs and increasing $I(z, v)$

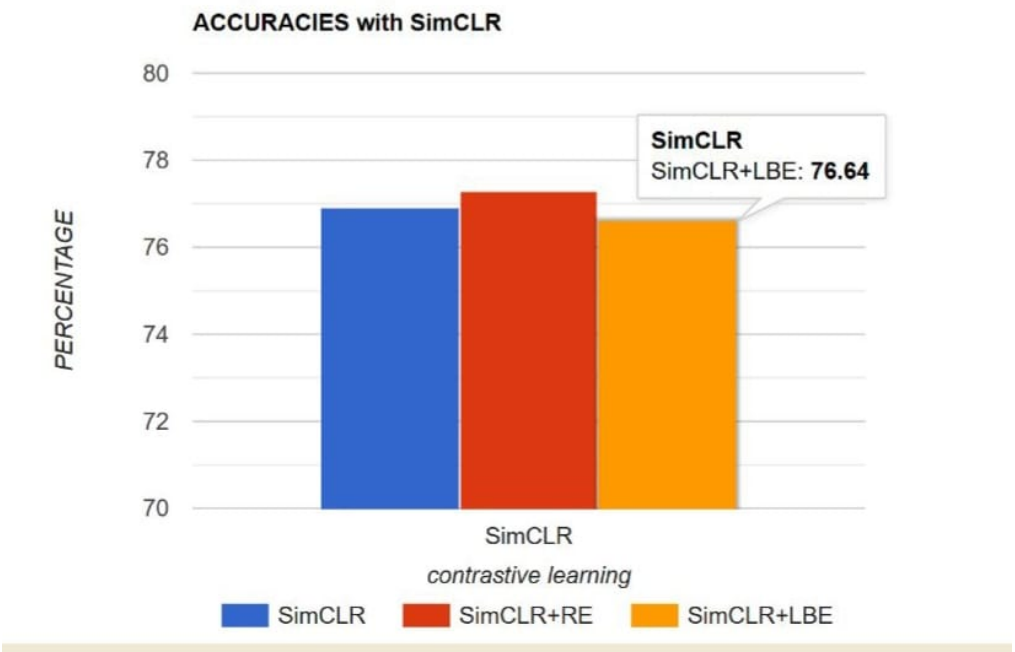
cannot make a difference. In fact, we use the unlabeled split for contrastive training on CIFAR-10, so it is intuitive that the shared information between views is not sufficient for the classification tasks on the train and test split.

Increasing $I(z, x)$ in supervised learning. According to the information bottleneck theory, a model extracts the approximate minimal sufficient statistics of the input x with respect to the label y in supervised learning. In other words, the representation 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 some downstream tasks, so we evaluate the effect of increasing $I(z, x)$ in supervised learning. We train the ResNet18 backbone using the crossentropy classification loss on CIFAR10 and CIFAR100, and choose $\lambda_1 = \lambda_2 = \lambda$ from $\{0.001, 0.01, 0.1, 1, 10\}$. The linear evaluation results are shown in BarGraphs and the best result in each block is in bold. As we can see, increasing $I(z, x)$ improves the performance on the transfer datasets and achieves comparable results on the source dataset, which means it can effectively alleviate the overfitting on the label information. This discovery helps to obtain more general representations in the field of supervised pre-training.

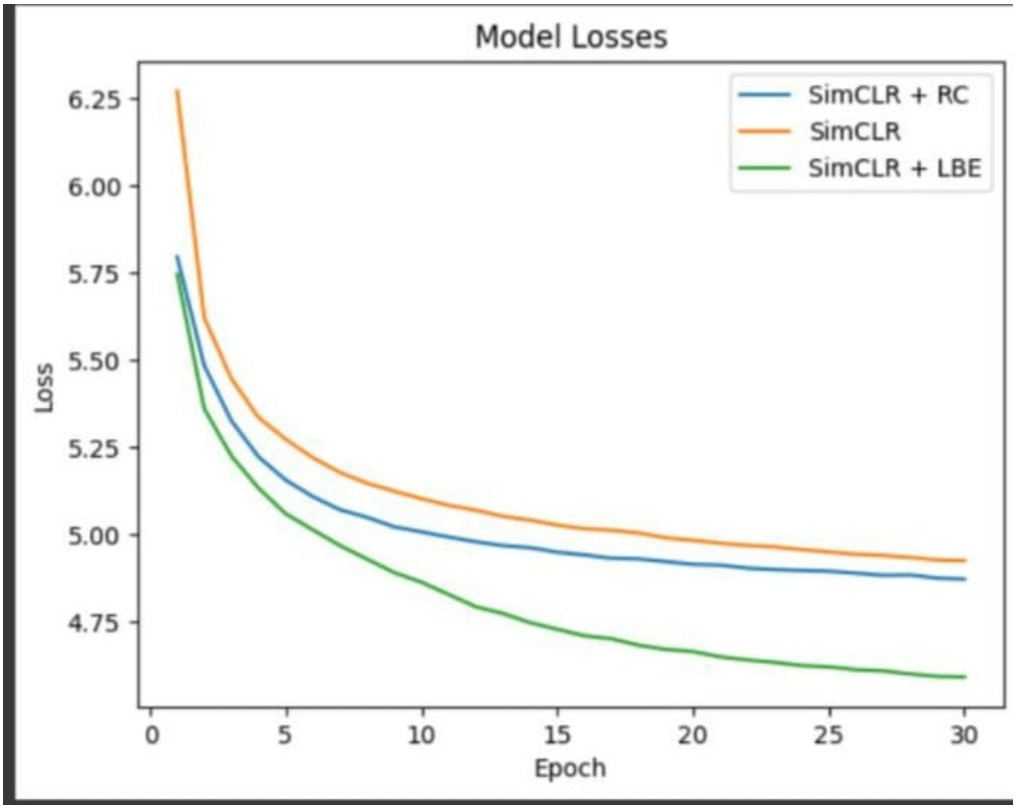
4.1 Observed Accuracies







4.2 Epoch vs Loss(oberseved)



4.3 Limitations

These were few limitations of our work.

1) Based on observations, our main assumption that non-ignorable non-shared task-relevant information usually holds for the cross-domain transfer tasks, but they may not be satisfied for the tasks for the training dataset.

2) Increasing $I(z, v)$ can also introduce noise (task-irrelevant outliers) information which could increase the inaccuracy in the downstream tasks, so we need to adjust the coefficients λ_1 and λ_2 for good output of different downstream tasks.

3) Because of limited computing resources, we couldn't do analysis on big datasets, and we ran only 30 epocs and we reduce dimentions to 32x32.

Chapter 5

Conclusion

In this work, we investigate the interaction between contrastive learning's downstream tasks and the learnt representations. We theoretically and empirically demonstrate that the minimal adequate representation is insufficient for downstream tasks since it loses non-shared task-relevant information, despite the fact that several works propose to learn the minimal sufficient representation. According to our research, contrastive learning roughly achieves the minimal representation that is necessary, which means it can over-fit to the shared information between views. To do this, when the downstream tasks are unclear, we suggest increasing the mutual information between the representation and input to roughly introduce more non-shared task-relevant information.According to our research, contrastive learning roughly achieves the minimal representation that is necessary, which means it can over-fit to the shared information between views. To do this, when the downstream tasks are unclear, we suggest increasing the mutual information between the representation and input to roughly introduce more non-shared task-relevant information. Since reconstruction can learn more adequate information and contrast can make the representations more discriminative, we can think about merging the reconstruction models and contrastive learning for convolutional neural networks or vision transformers in future work.

Chapter 6

Reference

These are all my referenced papers

[1] Rethinking Minimal Sufficient Representation in Contrastive Learning Haoqing Wang, Xun Guo, Zhi-Hong Deng, Yan Lu.

Link for paper: https://arxiv.org/pdf/2002.05709.pdf

[2] A Simple Framework for Contrastive Learning of Visual Representations Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton

link for the paper: https://arxiv.org/abs/2002.05709

[3] 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, pages 5171–5180. PMLR, 2019. 2, 5

[4] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. The Journal of Machine Learning Research, 19(1):1947–1980, 2018. 3

[5] 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. 2

[6] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021. 8

[7] 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, pages 201–208, 2003. 5

[8] 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, pages 531–540. PMLR, 2018. 2, 5

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Advances in Neural Information Processing Systems 33, 2020, 2020. 1 [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. 1, 2, 5, 6

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In IEEE Conference on Computer Vision and Pattern Recognition, 2021. Computer Vision Foundation / IEEE, 2021. 1

[12] 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, pages 3606–3613, 2014. 6

[13] 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, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 6