This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Rethinking Minimal Sufficient Representation in Contrastive Learning

Haoqing Wang*[1] Xun Guo[2] Zhi-Hong Deng[1] Yan Lu[2]
[1]Peking University [2]Microsoft Research Asia
{wanghaoqing, zhdeng}@pku.edu.cn {xunguo, yanlu}@microsoft.com

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 broad downstream tasks. Since all supervision information for one view comes from the other view, contrastive learning approximately obtains 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 cannot be guaranteed that all task-relevant information is shared between views. Therefore, we assume the non-shared task-relevant information cannot 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 cannot 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. Our code is available at https://github.com/Haoqing-Wang/InfoCL.

1. Introduction

Recently, contrastive learning [6–8, 18, 51] between different views of the data achieves outstanding success in the field of self-supervised representation learning. The learned representations are useful for broad downstream tasks in practice, such as classification, detection and segmentation [20]. 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* [43]. 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 [15], the non-shared information is often ignored, so that the minimal sufficient representation is approximately obtained.

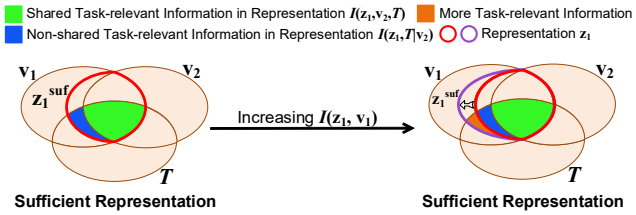


Figure 1. Demonstration of our motivation using information diagrams. Based on the (approximately minimal) sufficient representation learned in contrastive learning, increasing $I(z_1, v_1)$ approximately introduces more non-shared task-relevant information.

Tian *et al*. [40] find that the optimal views for contrastive learning depend on the downstream tasks when the minimal sufficient representation is obtained. In other words, the optimal views for task T_1 may not be suitable for task T_2. The reason may be that some information relevant to T_2 is not shared between these views. In this work, we formalize this conjecture and assume that the non-shared task-relevant information cannot be ignored. Based on this assumption, we theoretically prove that the minimal sufficient representation contains less task-relevant information than other sufficient representations and has a 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 task, 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 representation in contrastive learning is not sufficient for the downstream tasks. This reveals that the contrastive learning

*The work was done when the author was with MSRA as an intern.

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 cannot utilize any downstream task information in the training stage, 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. This motivation is demonstrated in Fig. 1 using information diagrams. We consider two implementations to increase the mutual information. The first one reconstructs the input to make the representations containing the key information about the input [26,44]. The second one relies on the high-dimensional mutual information estimate [5,34].

Overall, we summarize our contributions as follows.

- To the best of our knowledge, this is the first work to theoretically reveal that contrastive learning has the risk of over-fitting to the shared information between views. We provide comprehensive analysis based on the internal mechanism of contrastive learning that the views provide supervision information to each other.
- To alleviate this problem, when the downstream task information is not available, we propose to increase the mutual information between the representation and input to approximately introduce more task-relevant information, as shown in Fig. 1.
- We verify the effectiveness of our method for SimCLR [7], BYOL [18] and Barlow Twins [51] in classification, detection and segmentation tasks. We also provide extensive analytical experiments to further understand our hypotheses, theoretical analysis and model.

2. Related works

Contrastive learning. Contrastive learning between different views of the data is a successful self-supervised representation learning framework. The views are constructed by exploiting the structure of the unlabeled data, such as local patches and the whole image [23], different augmentations of the same image [2, 7, 20, 48], or video and text pairs [31,38]. Recently, Tian *et al*. [40] find that the optimal views for contrastive learning are task-dependent under the assumption of minimal sufficient representation. In other words, even if the given views are optimal for some downstream tasks, they may not be suitable for other tasks. In this work, we theoretically analyze this discovery and find that the contrastive learning models 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. Some recent works [15,43] propose to learn the minimal sufficient representation. They assume that almost all the information relevant to downstream tasks is shared between views, which is an overly idealistic assumption and conflicts with the discovery in [40].

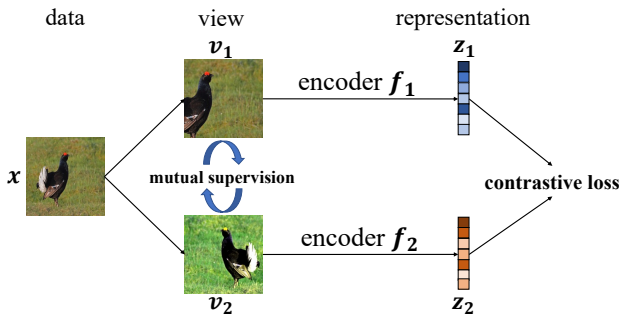


Figure 2. Internal mechanism of contrastive learning: the views provide supervision information to each other.

Information bottleneck theory. Based on the information bottleneck theory [36, 41, 42], a model extracts all task-relevant information in the first phase of learning (drift phase) to ensure sufficiency, and then compresses the task-irrelevant information in the second phase (diffusion phase). Our analysis shows that the learned representation in contrastive learning is not sufficient for the downstream tasks and can be seen as in the drift phase. We need to introduce more task-relevant information to achieve sufficiency.

3. 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-relevant information to the representations. Note that although the analysis is about the information content in the representations, its specific form is also very important. Therefore, our theoretical analysis is actually based on the premise that the information content in the representations is represented in the most appropriate form.

3.1. Contrastive learning

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

$$\max_{f_1, f_2} I(z_1, z_2) \tag{1}$$

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, v_1 and v_2 are usually two views of the data x. When 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 ($f_1 = f_2$).

In contrastive learning, the variable v_2 provides supervision information for v_1 and plays the similar role as the label y in the supervised learning, and vice versa [15]. This internal mechanism is illustrated in Fig. 2. Similar to the information bottleneck theory [1, 42] in the supervised learning, we can define the sufficient representation and minimal sufficient representation of v_1 (or v_2) for v_2 (or v_1) in contrastive learning [40, 43].

Definition 1. *(Sufficient Representation in Contrastive Learning) The representation z_1^{suf} of v_1 is sufficient for v_2 if and only if $I(z_1^{suf}, v_2) = I(v_1, v_2)$.*

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

Definition 2. *(Minimal Sufficient Representation in Contrastive Learning) The sufficient representation z_1^{min} of v_1 is minimal if and only if $I(z_1^{min}, v_1) \leq I(z_1^{suf}, v_1)$, $\forall z_1^{suf}$ 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.*, $I(z_1^{min}, v_1|v_2) = 0$.

Applying Data Processing Inequality [11] to the Markov chain $v_1 \to v_2 \to z_2$ and $z_2 \to v_1 \to z_1$, we have

$$I(v_1, v_2) \geq I(v_1, z_2) \geq I(z_1, z_2) \tag{2}$$

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

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.* [40] find that the optimal views for contrastive learning are task-dependent under the assumption of minimal sufficient representation. This discovery is intuitive since various downstream tasks need different information that is unknown during training.

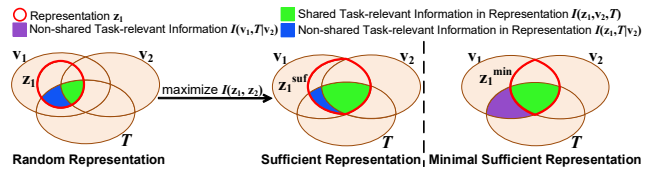


Figure 3. Information diagrams of different representations in contrastive learning. We consider the situation where the non-shared task-relevant information $I(v_1, T|v_2)$ cannot be ignored. Contrastive learning makes the representations extracting the shared information between views to obtain the sufficient representation which is approximately minimal. The minimal sufficient representation contains less task-relevant information from 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.*, $I(v_1, T|v_2)$ cannot be ignored. Eliminating all non-shared information has the risk of damaging the performance of the representations in the downstream tasks.

3.2. Analysis on 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 than other sufficient representations in the downstream tasks. All proofs for the below theorems are provided in Appendix.

Considering the symmetry between v_1 and v_2, without loss of generality, 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 [11, 14]. 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 z_1^{min} contains less task-relevant information from input v_1 than other sufficient representation z_1^{suf}, and $I(z_1^{min}, T)$ has a gap of $I(v_1, T|v_2)$ with the upper bound $I(v_1, T)$. Formally, we have*

$$\begin{aligned} &I(v_1, T) = I(z_1^{min}, T) + I(v_1, T|v_2) \\ \geq &I(z_1^{suf}, T) = I(z_1^{min}, T) + I(z_1^{suf}, T|v_2) \\ \geq &I(z_1^{min}, T) \end{aligned} \tag{3}$$

Theorem 1 indicates that z_1^{suf} can have better performance in task T than z_1^{min} because it contains more task-relevant information. When non-shared task-relevant information $I(v_1, T|v_2)$ 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 [16] 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 $P_e = 1 - \mathbb{E}_{p(z_1)}[\max_{t \in T} p(\widehat{T} = t|z_1)]$ 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 z_1, its Bayes error rate $P_e = \Gamma(\bar{P}_e)$ with*

$$\bar{P}_e \leq 1 - \exp[-(H(T) - I(z_1, T|v_2) - I(z_1, v_2, T))] \quad (4)$$

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

$$\bar{P}_e^{suf} \leq 1 - \exp[-(H(T) - I(z_1^{suf}, T|v_2) - I(v_1, v_2, T))] \quad (5)$$

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

$$\bar{P}_e^{min} \leq 1 - \exp[-(H(T) - I(v_1, v_2, T))] \quad (6)$$

Since $I(z_1^{suf}, T|v_2) \geq 0$, Theorem 2 indicates for classification task T, the upper bound of P_e^{min} is larger than P_e^{suf}. In other words, 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 $I(v_1, v_2, T)$ are not related to the representations, increasing $I(z_1^{suf}, T|v_2)$ can reduce the Bayes error rate in classification task T. When $I(z_1^{suf}, T|v_2) = I(v_1, T|v_2)$, z_1^{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 $R_e = \min_{\widetilde{T}} \mathbb{E}[(T - \widetilde{T}(z_1))^2] = \mathbb{E}[\varepsilon^2]$ with $\varepsilon(T, z_1) = T - \mathbb{E}[T|z_1]$.

Theorem 3. *(Minimum Expected Squared Prediction Error of Representations) For arbitrary learned representation z_1, when the conditional distribution $p(\varepsilon|z_1)$ is uniform, Laplacian or Gaussian distribution, the minimum expected squared prediction error R_e satisfies*

$$R_e = \alpha \cdot \exp[2 \cdot (H(T) - I(z_1, T|v_2) - I(z_1, v_2, T))] \quad (7)$$

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

$$R_e^{suf} = \alpha \cdot \exp[2 \cdot (H(T) - I(z_1^{suf}, T|v_2) - I(v_1, v_2, T))] \quad (8)$$

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

$$R_e^{min} = \alpha \cdot \exp[2 \cdot (H(T) - I(v_1, v_2, T))] \quad (9)$$

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

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 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 $I(z_1^{suf}, T|v_2)$ 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 $I(v_1, T|v_2)$ with the optimal representation, as shown in Theorem 1.

3.3. 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 non-shared 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 $I(z_1, T|v_2)$. However, we cannot utilize any downstream task information during training, so it is impossible to increase $I(z_1, T|v_2)$ 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

$$\max_{f_1, f_2} I(z_1, z_2) + \sum_{i=1}^{2} \lambda_i I(z_i, v_i) \quad (10)$$

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 $I(z, v) = H(v) - H(v|z)$ and $H(v)$ is not related with z, we can equivalently decrease the conditional entropy $H(v|z) = -\mathbb{E}_{p(z,v)}[\ln p(v|z)]$. Concretely, we use the representation z to reconstruct the original input v, as done in auto-encoder models [44]. Decreasing the entropy of reconstruction encourages the representation z to contain more information about the original input v. However, the conditional distribution $p(v|z)$ is intractable in practice, so we use $q(v|z)$ as an approximation and get $\mathbb{E}_{p(z,v)}[\ln q(v|z)]$, which is the lower bound of $\mathbb{E}_{p(z,v)}[\ln p(v|z)]$. We can increase $\mathbb{E}_{p(z,v)}[\ln q(v|z)]$ as an alternative objective. According to the type of input v (*e.g.*, images, text or audio), $q(v|z)$ 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 $q(v|z)$ is the Gaussian distribution $\mathcal{N}(v; \mu(z), \sigma^2 I)$ with given variance σ^2 and deterministic mean function $\mu(\cdot)$ which is usually parameterized by neural networks, we have

$$\mathbb{E}_{p(z,v)}[\ln q(v|z)] \propto -\mathbb{E}_{p(z,v)}[\|v - \mu(z)\|_2^2] + c \quad (11)$$

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

$$\max_{f_1, f_2, \mu} I(z_1, z_2) - \sum_{i=1}^{2} \lambda_i \mathbb{E}_{p(z_i, v_i)}[\|v_i - \mu_i(z_i)\|_2^2] \quad (12)$$

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 [4], the bound of Nguyen, Wainwright and Jordan [32], MINE [5] and InfoNCE [34]. We choose the InfoNCE lower bound and the detailed discussion is provided in Appendix. Concretely, the InfoNCE lower bound is

$$\hat{I}_{NCE}(z, v) = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} \ln \frac{p(z^k|v^k)}{\frac{1}{N}\sum_{l=1}^{N} p(z^l|v^k)}\right] \quad (13)$$

where $(z^k, v^k), k = 1, \cdots, N$ are N copies of (z, v) and the expectation is over $\Pi_k p(z^k, v^k)$. 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 $p(z|v)$ 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 [26] during training. For example, when we define $p(z|v)$ as the Gaussian distribution $\mathcal{N}(z; f(v), \sigma^2 I)$ 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

$$\tilde{I}_{NCE}(z, v) = \mathbb{E}\left[-\frac{1}{N}\sum_{k=1}^{N} \ln \sum_{l=1}^{N} \exp(-\rho\|z^l - f(v^k)\|_2^2)\right] \quad (14)$$

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 $I(z, v) = H(z) - H(z|v) = H(z) - \frac{d}{2}(\ln 2\pi + \ln \sigma^2 + 1)$ with d being representation dimension. It also be denoted as uniformity property [47]. The final optimization objective is

$$\max_{f_1, f_2} I(z_1, z_2) + \sum_{i=1}^{2} \lambda_i \tilde{I}_{NCE}(z_i, v_i) \quad (15)$$

Since the objective term Eq. (14) is calculated at the representation-level, when we use the convolutional neural networks (*e.g.*, ResNet [22]) 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 [42]. 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.

4. Experiments

In this section, we first verify the effectiveness of increasing $I(z, v)$ on various datasets, and then provide some analytical experiments. We choose three classic contrastive learning models as our baselines: SimCLR [7], BYOL [18] and Barlow Twins [51]. We denote our first implementation Eq. (12) as "RC" for "ReConstruction" and the second implementation Eq. (15) 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 [7]. For "LBE", we set $\sigma = 0.1$ and $\rho = 0.05$.

4.1. Effectiveness of increasing $I(z, v)$

We consider different types of the downstream task, including classification, detection and segmentation tasks. The results of Barlow Twins are provided in Appendix.

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
Model	ImageNet	DTD	CIFAR10	CIFAR100	CUBirds	VGGFlower	TrafficSigns
SimCLR	61.01	70.16	82.30	59.86	36.49	93.52	95.27
SimCLR+RC (ours)	**61.60**	**71.22**	**83.30**	**63.56**	37.42	**94.53**	**96.47**
SimCLR+LBE (ours)	61.37	70.95	83.20	61.99	**37.78**	94.34	95.99

Table 1. Linear evaluation accuracy (%) on the source dataset (CIFAR10, STL-10 or ImageNet) and other transfer datasets.

Pretraining. We train the models on CIFAR10 [27], STL-10 [10] and ImageNet [12]. For CIFAR10 and STL-10, we use the ResNet18 [22] backbone and the models are trained for 200 epochs with batch size 256 using Adam optimizer with learning rate 3e-4. For ImageNet, we use the ResNet50 [22] backbone and the models are trained for 200 epochs with batch size 1024 using LARS optimizer [50] and a cosine decay learning rate schedule.

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 [27], DTD [9], MNIST [28], FashionMNIST [49], CUBirds [45], VGG Flower [33] and Traffic Signs [25]. 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 [7, 24, 39, 46]. 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 over-fitting 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.

Object detection and instance segmentation. We conduct object detection on VOC07+12 [13] using Faster R-CNN [35], and detection and instance segmentation on COCO [29] using Mask R-CNN [21], following the setup in [20]. All methods use the R50-C4 [21] backbone that is initialized using the ResNet50 pre-trained on ImageNet. The results are show in Table 2. 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.

Model	AP	AP_{50}	AP_{75}
random initialization	33.8	60.2	33.1
SimCLR	45.5	76.2	47.5
SimCLR+RC (ours)	48.1 (+2.6)	78.0 (+1.8)	50.9 (+3.4)
SimCLR+LBE (ours)	47.4 (+1.9)	77.8 (+1.6)	49.7 (+2.2)

(a) Object detection on VOC07+12

Model	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
random initialization	35.6	54.6	38.2
SimCLR	38.6	58.5	41.8
SimCLR+RC (ours)	39.3 (+0.7)	59.0 (+0.5)	42.6 (+0.8)
SimCLR+LBE (ours)	39.0 (+0.4)	58.7 (+0.2)	42.3 (+0.5)

(b) Object detection on COCO

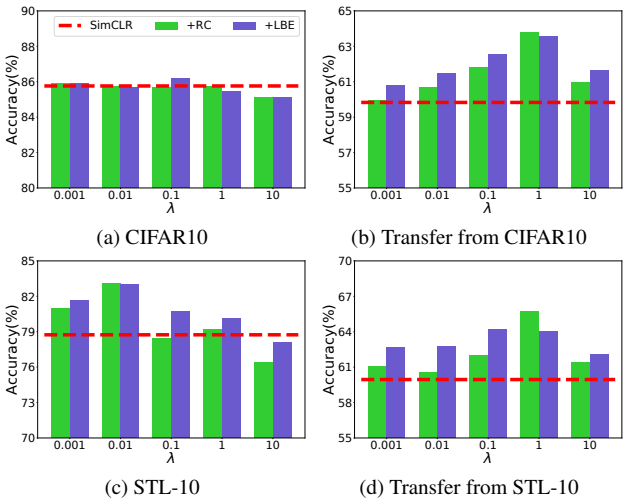
Model	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random initialization	31.4	51.5	33.5
SimCLR	33.9	55.2	36.0
SimCLR+RC (ours)	34.5 (+0.6)	55.7 (+0.5)	36.6 (+0.6)
SimCLR+LBE (ours)	34.2 (+0.3)	55.4 (+0.2)	36.4 (+0.4)

(c) Instance segmentation on COCO

Table 2. Object detection and instance segmentation on VOC07+12 and COCO. The models on COCO are fine-tuned using the default 2× schedule. In magenta are the gaps of at least +0.5 point to the baseline, SimCLR.

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**

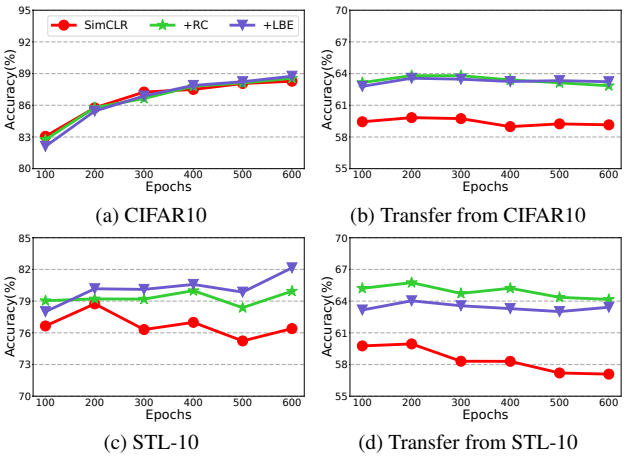
Table 3. Linear evaluation accuracy (%) on CIFAR10 and the transfer datasets. † represents adding Gaussian noise to the representations.



(a) CIFAR10 (b) Transfer from CIFAR10

(c) STL-10 (d) Transfer from STL-10

Figure 4. Linear evaluation accuracy on the source dataset (CIFAR10 or STL-10) and the averaged accuracy on all transfer datasets with varying hyper-parameter λ.



(a) CIFAR10 (b) Transfer from CIFAR10

(c) STL-10 (d) Transfer from STL-10

Figure 5. Linear evaluation accuracy on the source dataset (CIFAR10 or STL-10) and the averaged accuracy on all transfer datasets with varying epochs.

4.2. Analytical experiments

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

Eliminating non-shared information. Some recent works [15, 43] propose to eliminate the non-shared information between views in the representation to get the minimal sufficient representation. To this end, Federici *et al.* [15] minimize the regularization term

$$L_{MIB} = \frac{1}{2}\left[KL(p(z_1|v_1)\|p(z_2|v_2)) + KL(p(z_2|v_2)\|p(z_1|v_1))\right] \quad (16)$$

where $KL(\cdot||\cdot)$ represents the Kullback-Leibler divergence. When $p(z_1|v_1)$ and $p(z_2|v_2)$ are modeled as $\mathcal{N}(z_i; f_i(v_i), \sigma^2 I), i = 1, 2$ with given variance σ^2, it can be rewritten as $L_{MIB} = \mathbb{E}_{p(v_1,v_2)}\left[\|f_1(v_1) - f_2(v_2)\|_2^2\right]$. Identically, Tsai *et al.* [43] minimize the inverse predictive loss $L_{IP} = \mathbb{E}_{p(v_1,v_2)}\left[\|f_1(v_1) - f_2(v_2)\|_2^2\right]$. The detailed derivation is provided in Appendix. 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 in Table 3 and the best result in each block is in bold. 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 [30, 37]. Therefore, we assume that the hyper-parameters λ_1 and λ_2 control the amount of increasing $I(z_1, v_1)$ and $I(z_2, v_2)$ respectively. Larger λ_1 is expected to increase $I(z_1, v_1)$ 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.

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**

Table 4. Linear evaluation accuracy (%) on the source dataset (CIFAR10 or CIFAR100) and the transfer datasets.

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 $I(z, v)$ 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 Fig. 5. 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 non-shared information. For the classification tasks on the transfer datasets, the shared information between views is not sufficient. As shown in Fig. 5 (b) and (d), 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 but not on STL-10. As shown in Fig. 5 (a) and (c), the accuracy on CIFAR10 increases with more epochs and increasing $I(z, v)$ cannot make a difference. But the accuracy on STL-10 decreases with more epochs, and increasing $I(z, v)$ 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 tasks on the *train* and *test* split.

Increasing $I(z, x)$ in supervised learning. According to the information bottleneck theory [42], 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 cross-entropy 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 Table 4 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 over-fitting on the label information. This discovery helps to obtain more general representations in the field of supervised pre-training and we left it for the future work.

5. Limitations

Our work has the following limitations. 1) Based on our experimental observation, the assumption that non-shared task-relevant information cannot be ignored usually well holds for the cross-domain transfer tasks, but may not be satisfied for the tasks on the training dataset. 2) Increasing $I(z, v)$ can also introduce noise (task-irrelevant) information which may increase the data demand in the downstream tasks, so one may need to adjust the coefficients λ_1 and λ_2 to achieve effective trade-off for the different downstream tasks. 3) Due to limited computing resources, we cannot reproduce the best results of SimCLR on ImageNet which need the batch size of 4096 and more training epochs.

6. Conclusions

In this work, we explore the relationship between the learned representations and downstream tasks in contrastive learning. Although some 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 approximately obtains the minimal sufficient representation, which means it may over-fit to the shared information between views. To this end, we propose to increase the mutual information between the representation and input to approximately introduce more non-shared task-relevant information when the downstream tasks are unknown. For the future work, we can consider combining the reconstruction models [3, 19] and contrastive learning for convolutional neural networks or vision transformers, since reconstruction can learn more sufficient information and contrast can make the representations more discriminative.

References

[1] 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

[2] 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

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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[14] Meir Feder and Neri Merhav. Relations between entropy and error probability. *IEEE Transactions on Information theory*, 40(1):259–266, 1994. 3

[15] 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. 1, 2, 3, 7

[16] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013. 4

[17] 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*, 2020. 5

[18] 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*, 2020. 1, 2, 5

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 8

[20] 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*, pages 9729–9738, 2020. 1, 2, 6

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[22] 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*, pages 770–778, 2016. 5, 6

[23] 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. 2

[24] Chih-Hui Ho and Nuno Vasconcelos. Contrastive learning with adversarial examples. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020. 6

[25] 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)*, pages 1–8. Ieee, 2013. 6

[26] 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. 2, 5

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 6

[28] 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. 6

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[30] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020. 7

[31] 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*, pages 9879–9889, 2020. 2

[32] 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. 5

[33] 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*, pages 722–729. IEEE, 2008. 6

[34] 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

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 6

[36] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. 2

[37] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *8th International Conference on Learning Representations, 2020*, 2020. 7

[38] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 2

[39] Alex Tamkin, Mike Wu, and Noah D. Goodman. Viewmaker networks: Learning views for unsupervised representation learning. In *9th International Conference on Learning Representations, 2021*, 2021. 6

[40] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020. 1, 2, 3

[41] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 2

[42] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015. 2, 3, 5, 8

[43] 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. 1, 2, 3, 7

[44] 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. 2, 5

[45] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[46] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2021. 6

[47] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 5

[48] 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*, pages 3733–3742, 2018. 2

[49] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 6

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

[51] 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. 1, 2, 5