Appendix of
”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

A. Proofs of theorems

In this section, we provide the proofs of the theorems in the main text. Since the random variable $z_1 = f_1(v_1)$ is the representation of random variable v_1 where f_1 is an encoding function, we have

Assumption 1. *Random variable z_1 is conditionally independent from any other variable s in the system once random variable v_1 is observed, i.e., $I(z_1, s|v_1) = 0, \forall s$.*

This assumption is also adopted in [5]. When f_1 is a deterministic function, this assumption strictly holds. And when f_1 is a random function, the information in z_1 consists of the information from v_1 and the information introduced by the randomness of function f_1 which can be considered irrelevant to other variables in the system, so this assumption still holds. Next, we first present two lemmas for subsequent proofs.

Lemma 1. *Let z_1^{suf} and z_1^{min} are the sufficient representation and the minimal sufficient representation of view v_1 for v_2 in contrative learning respectively, we have*

$$I(z_1^{min}, v_2, T) = I(z_1^{suf}, v_2, T) = I(v_1, v_2, T) \tag{1}$$

$$I(z_1^{min}, T|v_2) = 0 \tag{2}$$

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

$$\begin{aligned}
& I(v_1, v_2, T) - I(z_1^{suf}, v_2, T) \\
&= [I(v_1, v_2) - I(v_1, v_2|T)] - [I(z_1^{suf}, v_2) - I(z_1^{suf}, v_2|T)] \\
&= I(z_1^{suf}, v_2|T) - I(v_1, v_2|T) \\
&= [H(v_2|T) - H(v_2|z_1^{suf}, T)] - [H(v_2|T) - H(v_2|v_1, T)] \\
&= H(v_2|v_1, T) - H(v_2|z_1^{suf}, T) \\
&= [I(z_1^{suf}, v_2|v_1, T) + H(v_2|v_1, z_1^{suf}, T)] \\
&\quad -[I(v_1, v_2|z_1^{suf}, T) + H(v_2|v_1, z_1^{suf}, T)] \\
&= I(z_1^{suf}, v_2|v_1, T) - I(v_1, v_2|z_1^{suf}, T) \\
&= I(z_1^{suf}, v_2|v_1, T) = 0
\end{aligned}$$

Therefore, we have

$$I(z_1^{suf}, v_2, T) = I(v_1, v_2, T)$$

The above proof process only uses the sufficiency of z_1^{suf} for v_2, so we have

$$I(z_1^{min}, v_2, T) = I(v_1, v_2, T)$$

2) From the Definition 2 in the main text and the Assumption 1, we have

$$I(z_1^{min}, v_1|v_2) = 0 \qquad I(z_1^{min}, T|v_1) = 0$$

Applying these two equations, we have

$$\begin{aligned}
I(z_1^{min}, T|v_2) &= I(z_1^{min}, T|v_1, v_2) + I(z_1^{min}, T, v_1|v_2) \\
&= I(z_1^{min}, T, v_1|v_2) \\
&= I(z_1^{min}, v_1|v_2) - I(z_1^{min}, v_1|T, v_2) = 0
\end{aligned}$$

□

We consider the conditional entropy of the task variable T given the representation z_1.

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

Lemma 2. *For arbitrary learned representation z_1, the conditional entropy $H(T|z_1)$ of the task variable T given z_1 satisfies*

$$H(T|z_1) = H(T) - I(z_1, T|v_2) - I(z_1, v_2, T) \quad (3)$$

Specifically, for the sufficient representation z_1^{suf}, the conditional entropy $H(T|z_1^{suf})$ satisfies

$$H(T|z_1^{suf}) = H(T) - I(z_1^{suf}, T|v_2) - I(v_1, v_2, T) \quad (4)$$

for the minimal sufficient representation z_1^{min}, the conditional entropy $H(T|z_1^{min})$ satisfies

$$H(T|z_1^{min}) = H(T) - I(v_1, v_2, T) \quad (5)$$

Proof. We have

$$\begin{aligned} H(T|z_1) &= H(T) - I(T, z_1) \\ &= H(T) - [I(T, z_1, v_2) + I(T, z_1|v_2)] \\ &= H(T) - I(z_1, T|v_2) - I(z_1, v_2, T) \end{aligned}$$

Applying the Eq. (1), the conditional entropy $H(T|z_1^{suf})$ satisfies

$$\begin{aligned} H(T|z_1^{suf}) &= H(T) - I(z_1^{suf}, T|v_2) - I(z_1^{suf}, v_2, T) \\ &= H(T) - I(z_1^{suf}, T|v_2) - I(v_1, v_2, T) \end{aligned}$$

Further, applying the Eq. (2), the conditional entropy $H(T|z_1^{min})$ satisfies

$$\begin{aligned} H(T|z_1^{min}) &= H(T) - I(z_1^{min}, T|v_2) - I(v_1, v_2, T) \\ &= H(T) - I(v_1, v_2, T) \end{aligned}$$

□

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

The proof of Theorem 1.

Proof. We decompose the Theorem 1 into three equations and prove them in turn.
1) $I(v_1, T) = I(z_1^{min}, T) + I(v_1, T|v_2)$.

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

2) $I(z_1^{suf}, T) = I(z_1^{min}, T) + I(z_1^{suf}, T|v_2)$.

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

3) $I(v_1, T|v_2) \geq I(z_1^{suf}, T|v_2) \geq 0$.

Applying the Data Processing Inequality [3] to the Markov chain $T \rightarrow v_1 \rightarrow z_1^{suf}$, we have $I(v_1, T) \geq I(z_1^{suf}, T)$, so

$$\begin{aligned} I(v_1, T|v_2) &= I(v_1, T) - I(v_1, T, v_2) \\ &= I(v_1, T) - I(z_1^{suf}, T, v_2) \\ &\geq I(z_1^{suf}, T) - I(z_1^{suf}, T, v_2) \\ &\geq I(z_1^{suf}, T|v_2) \geq 0 \end{aligned}$$

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

The proof of Theorem 2.

Proof. According to [4], the relationship between the Bayes error rate P_e and the conditional entropy $H(T|z_1)$ is

$$-\ln(1 - P_e) \leq H(T|z_1)$$

which is equivalent to

$$P_e \leq 1 - \exp[-H(T|z_1)]$$

Applying the Lemma 2, for arbitrary learned representation z_1, its Bayes error rate P_e satisfies

$$P_e \leq 1 - \exp[-(H(T) - I(z_1, T|v_2) - I(z_1, v_2, T))]$$

for the sufficient representation z_1^{suf}, its Bayes error rate P_e^{suf} satisfies

$$P_e^{suf} \leq 1 - \exp[-(H(T) - I(z_1^{suf}, T|v_2) - I(v_1, v_2, T))]$$

for the minimal sufficient representation z_1^{min}, its Bayes error rate P_e^{min} satisfies

$$P_e^{min} \leq 1 - \exp[-(H(T) - I(v_1, v_2, T))]$$

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

The proof of Theorem 3.

Proof. According to [6], when the conditional distribution $p(\varepsilon|z_1)$ of estimation error ε is uniform, Laplace and Gaussian distribution, the minimum expected squared prediction error R_e becomes $\frac{1}{12}\exp[2H(T|z_1)]$, $\frac{1}{2e^2}\exp[2H(T|z_1)]$ and $\frac{1}{2\pi e}\exp[2H(T|z_1)]$ respectively. Therefore, we unify them as

$$R_e = \alpha \cdot \exp[2H(T|z_1)]$$

where α is a constant coefficient which depends on the conditional distribution $p(\varepsilon|z_1)$. Applying the Lemma 2, for arbitrary learned representation z_1, we have

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

for the sufficient representation z_1^{suf}, we have

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

for the minimal sufficient representation z_1^{min}, we have

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

□

B. Choice of mutual information lower bound estimate

In our Implementation II, we need to use a mutual information lower bound estimate to calculate $I(z, v)$ where v is the original input (e.g., images) and z is the representation (feature vectors). We consider three candidate estimates:
1) The bound of Nguyen, Wainwright and Jordan [9]

$$\hat{I}_{NWJ}(z, v) = \mathbb{E}_{p(z,v)}[h(z, v)] - \mathbb{E}_{p(z)p(v)}[e^{h(z,v)-1}] \quad (6)$$

2) MINE [1]

$$\hat{I}_{MINE}(z, v) = \mathbb{E}_{p(z,v)}[h(z, v)] - \ln(\mathbb{E}_{p(z)p(v)}[e^{h(z,v)}]) \quad (7)$$

3) InfoNCE [10]

$$\hat{I}_{NCE}(z, v) = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} \ln \frac{p(z^k|v^k)}{\frac{1}{N}\sum_{l=1}^{N} p(z^l|v^k)}\right] \quad (8)$$

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

C. More experiments

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

C.1. Results on Barlow Twins

In the main text, we provide the results on two classic contrastive learning models: SimCLR [2] and BYOL [7]. SimCLR perfectly matches the contrastive learning framework, maximizing the lower bound estimate of the mutual information $I(z_1, z_2)$. BYOL avoids the dependence on the large amount of negative samples, and adopts prediction loss and the asymmetric structure. We further verify the effectiveness of increasing $I(z, v)$ on Barlow Twins [12] which makes the cross-correlation matrix between the representations of different views as close to the identity matrix as possible. Although the loss functions of these contrastive learning models are very different, they all satisfy the internal mechanism that the views provide supervision information to each other, so they all approximately learn the minimal sufficient representation. We use the same pre-training schedule and linear evaluation protocol as in the main text and set $\lambda_1 = \lambda_2 = 1$. For STL-10, we use the *unlabeled* split for contrastive learning and the *train* and *test* split for linear evaluation.

The results are shown in Table 1 and the best result in each block is in bold. Increasing $I(z, v)$ can improve the accuracy of the learned representations in Barlow Twins in downstream classification tasks, which indicates that our analysis results are applicable to various contrastive losses.

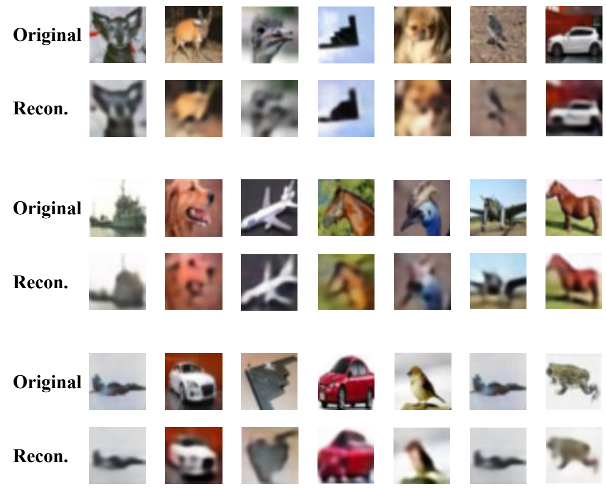

Figure 1. Demonstration of the reconstruction effect of our Implementation I. We provide the original input images and the reconstructed images for comparison. We use SimCLR contrastive loss and take CIFAR10 as the training dataset.

C.2. Reconstructed samples

In order to show the reconstruction effect of our Implementation I, we provide the reconstructed images after training. As an example, we use SimCLR contrastive loss and take CIFAR10 as the training dataset. The original input images and the reconstructed images are shown in Fig. 1. As we can see, the reconstructed images retain the shape and outline information in the original images, so as the obtained representations. Since we use the mean square error loss to optimize the reconstruction module, the reconstructed images are blurry and this phenomenon is also observed in vanilla variational auto-encoder [8].

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

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

D. Derivation of L_{MIB} and L_{IP}

Federici *et al*. [5] and Tsai *et al*. [11] propose to eliminate the non-shared information between views in the representation to get the minimal sufficient representation. To this end, they propose their respective regularization terms. Here we derive the specific forms used in the main text.

In [5], the regularization term is

$$\begin{aligned} L_{MIB} &= D_{SKL}(p(z_1|v_1)||p(z_2|v_2)) \\ &= \frac{1}{2}[KL(p(z_1|v_1)||p(z_2|v_2)) \\ &+ KL(p(z_2|v_2)||p(z_1|v_1))] \end{aligned} \tag{9}$$

According to the description in their paper and the official code [1], they model the two stochastic encoders $p(z_1|v_1)$ and $p(z_2|v_2)$ as

$$p(z_1|v_1) = \mathcal{N}(z_1; \mu_1, \mathrm{diag}(\sigma_1^2)) \tag{10}$$

$$p(z_2|v_2) = \mathcal{N}(z_2; \mu_2, \mathrm{diag}(\sigma_2^2)) \tag{11}$$

where $\mu_1(v_1)$,$\sigma_1^2(v_1)$,$\mu_2(v_2)$ and $\sigma_2^2(v_2)$ are all functions of the input (v_1 or v_2), $\mathrm{diag}(e)$ creates a matrix in which the diagonal elements consist of vector e and all off-diagonal elements are zeros. The regularization term has the analytical expression

$$L_{MIB} = \frac{1}{4}\sum_{i=1}^{d}\left[\frac{\sigma_1^{i2}}{\sigma_2^{i2}} + \frac{\sigma_2^{i2}}{\sigma_1^{i2}} + \frac{(\mu_1^i-\mu_2^i)^2}{\sigma_2^{i2}} + \frac{(\mu_2^i-\mu_1^i)^2}{\sigma_1^{i2}} - 2\right] \tag{12}$$

where d is the dimension of z_1 and z_2. We want to minimize L_{MIB}, and when $\sigma_1^2 = \sigma_2^2$, the term $\sigma_1^{i2}/\sigma_2^{i2} + \sigma_2^{i2}/\sigma_1^{i2}$ takes the minimum value 2, so the regularization term becomes

$$L_{MIB} = \frac{1}{2}\sum_{i=1}^{d}\frac{(\mu_1^i - \mu_2^i)^2}{\sigma_1^{i2}} \tag{13}$$

In practice, minimizing L_{MIB} makes the variance σ_1^2 and σ_2^2 very large, and the sampled representations change drastically and have very poor performance in downstream tasks. If the upper bound of the variance σ_1^2 and σ_2^2 is fixed, such as using the sigmoid activation function to limit it to $(0, 1)$, they will converge to the maximum value as the training progresses. Therefore, we might as well fix the variance and model the two stochastic encoders $p(z_1|v_1)$ and $p(z_2|v_2)$ as

$$p(z_1|v_1) = \mathcal{N}(z_1; f_1(v_1), \sigma^2 I) \tag{14}$$

$$p(z_2|v_2) = \mathcal{N}(z_2; f_2(v_2), \sigma^2 I) \tag{15}$$

where I is the identity matrix, σ^2 is the given variance, $f_i, i = 1, 2$ are deterministic encoders. This also guarantees a fair comparison with our Implementation II. According to the Eq. (13), the regularization term is equivalent to

$$L_{MIB} = \|f_1(v_1) - f_2(v_2)\|_2^2 \tag{16}$$

We calculate the expectation of the regularization term on the data distribution $p(v_1, v_2)$ and get

$$L_{MIB} = \mathbb{E}_{p(v_1,v_2)}[\|f_1(v_1) - f_2(v_2)\|_2^2] \tag{17}$$

In [11], the authors define the inverse predictive loss

$$L_{IP} = \mathbb{E}_{p(v_1,v_2)}[\|z_1 - z_2\|_2^2] = \mathbb{E}_{p(v_1,v_2)}[\|f_1(v_1) - f_2(v_2)\|_2^2] \tag{18}$$

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

References

[1] 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. 3

[2] 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. 3

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

[4] Meir Feder and Neri Merhav. Relations between entropy and error probability. *IEEE Transactions on Information theory*, 40(1):259–266, 1994. 2

[5] 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, 4

[6] Benoît Frénay, Gauthier Doquire, and Michel Verleysen. Is mutual information adequate for feature selection in regression? *Neural Networks*, 48:1–7, 2013. 2

[7] 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. 3

[8] 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. 3

[9] 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. 3

[10] 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. 3

[11] 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. 4

[12] 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. 3