Appendix of
”Rethinking Minimal Sufficient Representation in Contrastive Learning”

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

\section*{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

\begin{assumption}
Random variable \( z_1 \) is conditionally independent from any other variable in the system once random variable \( v_1 \) is observed, i.e., \( I(z_1, s|v_1) = 0, \forall s \).
\end{assumption}

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.

\begin{lemma}
Let \( z_1^{\text{suf}} \) and \( z_1^{\text{min}} \) are the sufficient representation and the minimal sufficient representation of view \( v_1 \) for \( v_2 \) in contrastive learning respectively, we have

\begin{align}
I(z_1^{\text{min}}, v_2, T) &= I(z_1^{\text{suf}}, v_2, T) = I(v_1, v_2, T) \quad (1) \\
I(z_1^{\text{min}}, T|v_2) &= 0 \quad (2)
\end{align}

Proof. 1) From the Definition 1 in the main text and the

Assumption 1, we have

\[
I(v_1, v_2, T) - I(z_1^{\text{suf}}, v_2, T) = [I(v_1, v_2) - I(v_1, v_2|T)] - I(z_1^{\text{suf}}, v_2) - I(z_1^{\text{suf}}, v_2|T) = I(z_1^{\text{suf}}, v_2|T) - I(v_1, v_2|T) = [H(v_2|T) - H(v_2|z_1^{\text{suf}}, T)] - [H(v_2|T) - H(v_2|v_1, T)] = H(v_2|v_1, T) - H(v_2|z_1^{\text{suf}}, T) = [I(z_1^{\text{suf}}, v_2|v_1, T) + H(v_2|v_1, z_1^{\text{suf}}, T)] - I(v_1, v_2|v_1, T) = I(z_1^{\text{suf}}, v_2|v_1, T) - I(v_1, v_2|z_1^{\text{suf}}, T) = I(z_1^{\text{suf}}, v_2|v_1, T) = 0
\]

Therefore, we have

\[
I(z_1^{\text{suf}}, v_2, T) = I(v_1, v_2, T)
\]

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

\[
I(z_1^{\text{min}}, v_2, T) = I(v_1, v_2, T)
\]

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

\[
I(z_1^{\text{min}}, v_1|v_2) = 0 \quad I(z_1^{\text{min}}, T|v_1) = 0
\]

Applying these two equations, we have

\[
I(z_1^{\text{min}}, T|v_2) = I(z_1^{\text{min}}, T|v_1, v_2) + I(z_1^{\text{min}}, T|v_1, v_2) = I(z_1^{\text{min}}, T, v_1|v_2) = I(z_1^{\text{min}}, v_1|v_2) - I(z_1^{\text{min}}, v_1|T, v_2) = 0
\]

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

\[\tag*{\blacksquare}\]

\footnote{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

$$H(T|z_1) = H(T) - I(T, z_1)$$

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

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

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

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

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

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

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

$$I(v_1, T) \geq I(z_1^{suf}, T)$$

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

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

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 $\varepsilon$ is uniform, Laplace and Gaussian distribution, the minimum expected squared prediction error $R_e$ becomes $\frac{1}{12} \exp[2H(T|z_1)]$, $\frac{1}{2\pi\sigma^2} \exp[2H(T|z_1)]$ and $\frac{1}{2\pi\sigma^2} \exp[2H(T|z_1)]$ respectively. Therefore, we unify them as

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

where $\alpha$ 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 $z_1^{\text{suf}}$, we have

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

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

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

\[ \square \]

**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]

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

2) MINE [1]

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

3) InfoNCE [10]

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

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 $\hat{I}_{\text{NWJ}}$ and $\hat{I}_{\text{MINE}}$, 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 $\hat{I}(z, v)$ 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.

**C.2. Reconstructed samples**

In order to show the reconstruction effect of our Implementation I, we provide the original images and the reconstructed images for comparison. 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].
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

$$L_{MIB} = D_{SKL}(p(z_1|v_1)||p(z_2|v_2))$$

$$= \frac{1}{2}KL(p(z_1|v_1)||p(z_2|v_2))$$

$$= KL(p(z_2|v_2)||p(z_1|v_1))$$

(9)

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

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

(10)

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

(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 $(v_1, v_2)$, $\text{diag}(e)$ 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

$$L_{MIB} = \frac{1}{2} \sum_{i=1}^{d} \left[ \frac{\sigma_1^4}{\sigma_2^4} + \frac{\sigma_2^4}{\sigma_1^4} + \frac{(\mu_1(v_2) - \mu_2(v_1))^2}{\sigma_2^4} + \frac{(\mu_1(v_1) - \mu_2(v_2))^2}{\sigma_1^4} - 2 \right]$$

(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 $\sigma_1^4/\sigma_2^4 + \sigma_2^4/\sigma_1^4$ takes the minimum value 2, so the regularization term becomes

$$L_{MIB} = \frac{1}{2} \sum_{i=1}^{d} \frac{(\mu_1^2 - \mu_2^2)^2}{\sigma_i^4}$$

(13)

In practice, minimizing $L_{MIB}$ makes the variance $\sigma_1^2$ and $\sigma_2^2$ very large, and the sampled representations change drastically and have very poor performance in downstream tasks. If the upper bound of the variance $\sigma_1^2$ and $\sigma_2^2$ 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

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

(14)

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

(15)

where $I$ is the identity matrix, $\sigma^2$ 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

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

(16)

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

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

(17)

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

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

(18)

References


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


