Supplementary Material

1. The Contrastive learning-based MVC method

The Contrastive learning-based MVC methods utilize the CL loss(e.g., InfoNCE loss, NT-Xent loss) to align the representation from different views. The multi-view Contrastive learning loss is as follows

$$\mathcal{L}_{Contrastive}^{MV} = \frac{1}{n * V * (V-1)} \sum_{i=1}^{n} \sum_{u=1}^{V} \sum_{v=1}^{V} \nVdash_{u \neq v} l_i^{uv} \tag{1}$$

l_i^{uv} is the CL loss between two samples from two views and is defined as follows

$$l_i^{uv} = -log \frac{e^{\frac{s_{ii}^{uv}(z_i^u, z_i^v)}{\tau}}}{\sum_{s' \in Neg(z_i^u, z_i^v)} e^{\frac{s'}{\tau}}} \tag{2}$$

and $s_{ij}^{uv}(z_i^u, z_i^v) = \frac{(z_i^u)^T z_i^v}{\|z_i^u\| \cdot \|z_i^v\|}$ denotes the cosine similarity between two samples from two views. τ denotes the temperature hyperparameter and $Neg(z_i^u, z_i^v)$ denotes the set of similarities of all negative sample pairs.

2. The Mutual Information Maximization-based MVC method

The Mutual Information Maximization-based Multi-View Clustering (MVC) method illustrates that the integration of maximizing mutual information between views and minimizing conditional entropy can be effectively incorporated into a multi-view learning framework. This methodology facilitates the extraction of task-relevant information while discarding task-irrelevant data. We generalize the loss function of this approach to an arbitrary number of views:

$$\begin{aligned} \mathcal{L}_{MI}^{MV} = & \frac{2}{V(V-1)} \sum_{u=1}^{V-1} \sum_{v=u+1}^{V} -(I(\mathbf{Z}^{(u)}, \mathbf{Z}^{(v)}) \\ & + \alpha(H(\mathbf{Z}^{(u)}) + H(\mathbf{Z}^{(v)}))) \end{aligned} \tag{3}$$

where the computation for each pair of views is given by:

$$\begin{aligned} & I(\mathbf{Z}^{(u)}, \mathbf{Z}^{(v)}) + \alpha(H(\mathbf{Z}^{(u)}) + H(\mathbf{Z}^{(v)})) \\ & = -\sum_{a=1}^{D} \sum_{b=1}^{D} \mathbf{P}_{ab}^{(uv)} log \frac{\mathbf{P}_{ab}^{(uv)}}{(\mathbf{P}_a^{(u)})^{(\alpha+1)} * (\mathbf{P}_b^{(v)})^{(\alpha+1)}} \end{aligned} \tag{4}$$

Where I denotes mutual information, and H represents information entropy, with the parameter α used for entropy regularization. From an information-theoretic perspective, the entropy $H(\mathbf{Z}^i)$ reflects the amount of information contained in the i-th view representation $\mathbf{Z}^i$. Maximizing $H(\mathbf{Z}^u)$ and $H(\mathbf{Z}^v)$ prevents the trivial solution of assigning all samples to a single cluster. Eq.(4) is derived from the formula for calculating mutual information, which represents the feature dimension of the view-specific representations processed through the MLP. To express $I(\mathbf{Z}^{(u)}, \mathbf{Z}^{(v)})$, we define the joint probability distribution matrix $\mathbf{P}$. This matrix $\mathbf{P}$ represents the distribution of the two view representations, $\mathbf{Z}^{(u)}$ and $\mathbf{Z}^{(v)}$, as discrete cluster assignments across D 'classes'. We first compute $\tilde{\mathbf{P}}^{(uv)} = \frac{1}{n} \sum_{i=1}^{n} z_i^{(u)} (z_i^{(v)})^T$, and then symmetrize it as $\tilde{\mathbf{P}}^{(uv)} = \frac{1}{2}(\tilde{\mathbf{P}}^{(uv)} + (\tilde{\mathbf{P}}^{(uv)})^T)$ for estimation. The marginal distributions $\mathbf{P}^{(u)}$ and $\mathbf{P}^{(v)}$ are obtained by summing over the rows and columns of $\mathbf{P}^{(uv)}$, respectively.

3. The information bottleneck-based MVC method

The information bottleneck-based MVC method [**?**] extends the information bottleneck approach to the unsupervised learning setting. Under the condition of mutual redundancy among various views, the multi-view information bottleneck loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{IB}^{MV} = & \frac{2}{V(V-1)} \sum_{u=1}^{V-1} \sum_{v=u+1}^{V} (-I_{\theta\psi}(\mathbf{Z}^{(u)}; \mathbf{Z}^{(v)}) \\ & + \beta D_{SKL}(p_\theta(\mathbf{Z}^{(u)}|\mathbf{V}^{(u)}) || p_\psi(\mathbf{Z}^{(v)}|\mathbf{V}^{(v)})) \end{aligned} \tag{5}$$

Given two views, $\mathbf{V}^{(u)}$ and $\mathbf{V}^{(v)}$, which are mutually redundant with respect to a label y, we aim to define an objective function for the representation $\mathbf{Z}^{(u)}$ of $\mathbf{V}^{(u)}$. This objective function should discard as much information as possible without losing any label information. The representation $\mathbf{Z}^{(u)}$ should be sufficient with respect to $\mathbf{V}^{(v)}$ to ensure the sufficiency of y, while enhancing the robustness of the representation by discarding irrelevant information. Consequently, we can combine these two requirements using a relaxed Lagrangian objective to obtain the minimal sufficient representation $\mathbf{Z}^{(u)}$ for $\mathbf{V}^{(u)}$:

$$\mathcal{L}_1(\theta; \lambda_1) = I_\theta(\mathbf{Z}^{(u)}; \mathbf{V}^{(u)} | \mathbf{V}^{(v)}) - \lambda_1 I_\theta(\mathbf{V}^{(v)}; \mathbf{V}^{(u)}) \tag{6}$$

Symmetrically, we define the loss $\mathcal{L}_2$ to obtain the minimal sufficient representation $\mathbf{Z}^{(v)}$ for $\mathbf{V}^{(v)}$:

$$\mathcal{L}_2(\psi; \lambda_2) = I_\psi(\mathbf{Z}^{(v)}; \mathbf{V}^{(v)} | \mathbf{V}^{(u)}) - \lambda_2 I_\psi(\mathbf{V}^{(u)}; \mathbf{V}^{(v)}) \tag{7}$$

Based on Eq.(6) and (7), we can obtain the average of the loss expressions for the minimal sufficient representationsă$\mathbf{Z}^{(u)}$ăandă $\mathbf{Z}^{(v)}$ of the two views, namely:

$$\begin{aligned}\mathcal{L}_{\frac{1+2}{2}}(\theta,\psi;\lambda_1,\lambda_2) = {} & \frac{I_\theta(\mathbf{Z}^{(u)};\mathbf{V}^{(u)}|\mathbf{V}^{(v)}) + I_\psi(\mathbf{Z}^{(v)};\mathbf{V}^{(v)}|\mathbf{V}^{(u)})}{2} \\ & - \frac{\lambda_1 I_\theta(\mathbf{V}^{(u)};\mathbf{V}^{(v)}|\mathbf{Z}^{(u)}) + \lambda_2 I_\psi(\mathbf{V}^{(u)};\mathbf{V}^{(v)}|\mathbf{Z}^{(u)})}{2}\end{aligned} \tag{8}$$

Considering $\mathbf{Z}^{(u)}$ăandă $\mathbf{Z}^{(v)}$ on the same domain $\mathbb{Z}$, $I_\theta(\mathbf{Z}^{(u)};\mathbf{V}^{(u)}|\mathbf{V}^{(v)})$ is upper bounded by $D_{KL}(p_\psi(\mathbf{Z}^{(u)}|\mathbf{V}^{(u)}))||(p_\theta(\mathbf{Z}^{(v)}|\mathbf{V}^{(v)}))$, analogously,$I_\psi(\mathbf{Z}^{(v)};\mathbf{V}^{(v)}|\mathbf{V}^{(u)})$ is upper bounded by $D_{KL}(p_\psi\mathbf{Z}^{(v)}|\mathbf{V}^{(v)}))||(p_\theta\mathbf{Z}^{(u)}|\mathbf{V}^{(u)})$.According to the chain role of mutual information,$I_\theta(\mathbf{Z}^{(u)};\mathbf{V}^{(v)})$ is lower bounded by $I_{\theta\psi}(\mathbf{Z}^{(u)};\mathbf{Z}^{(v)})$.Therefore, the loss function in Eq.(8) can be upper-bounded with:

$$\begin{aligned}\mathcal{L}_{\frac{1+2}{2}}(\theta,\psi;\lambda_1,\lambda_2) \leq {} & D_{SKL}(p_\psi(\mathbf{Z}^{(u)}|\mathbf{V}^{(u)}))||(p_\theta(\mathbf{Z}^{(v)}|\mathbf{V}^{(v)})) \\ & - \frac{\lambda_1+\lambda_2}{2} I_{\theta\psi}(\mathbf{Z}^{(u)};\mathbf{Z}^{(v)})\end{aligned} \tag{9}$$

Multiplying both terms with $\beta : \frac{\lambda_1+\lambda_2}{2}$and re-parametrizing the objective, we obtain Eq.(5).

4. The proof of the non-transitivity of multi-view mutual redundancy

In our preceding discussion, we examined three categories of view-pair self-supervised learning techniques employed in Multi-View Clustering (MVC). These methods are predicated on the fundamental assumption of redundancy among multiple views, which we refer to as the multi-view mutual redundancy assumption. Nevertheless, from a theoretical standpoint, this assumption is not universally valid. In practical applications, the degree of redundancy can vary significantly across different datasets and scenarios, which may impact the effectiveness of these methods. It is of paramount importance to recognize that the assumption of multi-view mutual redundancy is not universally applicable. In instances where this assumption is invalid, it may be necessary to consider alternative methods or approaches.

Below, we present a straightforward proof method demonstrating the non-transitivity of multi-view mutual redundancy in multi-view settings.

Given three views $\mathbf{V}^1$, $\mathbf{V}^2$,$\mathbf{V}^3$ and a clustering task $\mathbf{Y}$,for which satisfy the following conditions:

- $\mathbf{V}^1$, $\mathbf{V}^2$ are mutually redundant views for clustering task $\mathbf{Y}$
- $\mathbf{V}^2$, $\mathbf{V}^3$ are mutually redundant views for clustering task $\mathbf{Y}$

If $\mathbf{V}^1$ and $\mathbf{V}^3$ are not mutually redundant views for $\mathbf{Y}$, we can define $\mathbf{Y}$ as the exclusive or operator applied to $\mathbf{V}^1$ and $\mathbf{V}^3$ $(\mathbf{Y} := \mathbf{V}^1 XOR\, \mathbf{V}^3)$.The XOR operation yields a result of 1 when the two input values are different, and 0 when they are the same. This indicates that Y is determined by both $\mathbf{V}^1$ and $\mathbf{V}^3$. Therefore, if Y and one of these variables are known, the other variable can be accurately deduced. This reflects that the information provided by $\mathbf{V}^1$ and $\mathbf{V}^3$ is complementary rather than redundant.

By employing the chaining principle of mutual information and the definition of conditional mutual information, the following can be derived:

$$\begin{aligned} I(\mathbf{V}^1;\mathbf{Y}|\mathbf{V}^2) &= H(\mathbf{V}^1|\mathbf{V}^2) - H(\mathbf{V}^1|\mathbf{V}^2\mathbf{Y}) \\ &= H(\mathbf{V}^1) - H(\mathbf{V}^1) = 0 \end{aligned} \tag{10}$$

$$\begin{aligned} I(\mathbf{V}^2;\mathbf{Y}|\mathbf{V}^3) &= H(\mathbf{V}^2|\mathbf{V}^3) - H(\mathbf{V}^2|\mathbf{V}^3\mathbf{Y}) \\ &= H(\mathbf{V}^2) - H(\mathbf{V}^2) = 0 \end{aligned} \tag{11}$$

$$\begin{aligned} I(\mathbf{V}^1;\mathbf{Y}|\mathbf{V}^3) &= H(\mathbf{V}^1|\mathbf{V}^3) - H(\mathbf{V}^1|\mathbf{V}^3\mathbf{Y}) \\ &= H(\mathbf{V}^1) = 1 \end{aligned} \tag{12}$$

$$\begin{aligned} I(\mathbf{V}^3;\mathbf{Y}|\mathbf{V}^1) &= H(\mathbf{V}^3|\mathbf{V}^1) - H(\mathbf{V}^3|\mathbf{V}^1\mathbf{Y}) \\ &= H(\mathbf{V}^3) = 1 \end{aligned} \tag{13}$$

Under this definition, the conditional mutual information is non-zero, demonstrating the complementary dependency of the two variables on Y. This clearly indicates that they are not mutually redundant.

5. Dataset Information

Caltech101-20[1] consists of 2,386 images of 20 subjects with the views of Gabor, wavelet moments (WM), CENTRIST, HOG, GIST and LBP features.**Handwritten (HW)**[2] dataset is a specialized image dataset designed for handwritten character recognition and analysis. **UCI digit**[?] is a handwritten digit dataset containing 2,000 samples of 3 views. **ALOI**[?] is a dataset focusing on object recognition and visual classification in the field of computer vision. **BBC**[3] dataset is a well-known dataset used for text classification and natural language processing research, particularly in news categorization and topic identification tasks. The **FashionMNIST**(FM)[?] dataset consists of 70,000 samples with 4 views.

[1] http://www.vision.caltech.edu/ImageDatasets/Caltech101/

[2] https://archive.ics.uci.edu/ml/datasets/Multiple+Features

[3] http://mlg.ucd.ie/datasets/segment.html