Differentiable Information Bottleneck for Deterministic Multi-view Clustering

Supplementary Material

7. Supplementary A

7.1. Derivation of Rényi's α-order entropy

In this section, we further explain the derivation of α-order version from Rényi's 2-order entropy . According to its definition, the Rényi's α-order entropy can be formulated by an expectation as follows

$$H_\alpha(\mathbf{X}) = \frac{1}{1-\alpha}\log \int_{\mathcal{X}} p^\alpha(x)dx = \frac{1}{1-\alpha}\log \mathbb{E}\left[p^{\alpha-1}(\mathbf{X})\right] \quad (16)$$

where $\alpha \in (0,1)\cup(1,\infty)$, $p(x)$ is the probability density function of the random variable $\mathbf{X}$.

Next, we can calculate the expectation with sample mean as is commonly done in density estimation [17]. Then, the Rényi's α-order entropy in Eq. 16 can be rewritten as follows

$$\begin{aligned} H_\alpha(X) &\approx \hat{H}_\alpha(\mathbf{X}) \\ &= \frac{1}{1-\alpha}\log \mathbb{E}\left[p^{\alpha-1}(\mathbf{X})\right] \\ &= \frac{1}{1-\alpha}\log \frac{1}{n}\sum_{j=1}^{n} p^{\alpha-1}(x_j) \end{aligned} \quad (17)$$

Similarity, we use the Parzen density estimator with a Gaussian kernel $g_\sigma(x,y) = \exp\left(-\frac{1}{2\sigma^2}||x-y||^2\right)$ to calculate the probability density function $p(x_j)$, which can be plugged into the Eq. 17, yields

$$\begin{aligned} \hat{H}_\alpha(\mathbf{X}) &= \frac{1}{1-\alpha}\log \frac{1}{n}\sum_{j=1}^{n} p^{\alpha-1}(x_j) \\ &= \frac{1}{1-\alpha}\log \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} g_\sigma(x_i,x_j)\right)^{\alpha-1} \end{aligned} \quad (18)$$

Now, we can construct the Gram matrix G with elements $G_{ij} = g_\sigma(x_i,x_j)$. For all $\alpha > 0$ and $a \neq 1$, Rényi's α-order entropy is a general-purpose measurement which can be calculated directly by the eigenvalues of a Gram matrix G.

7.2. Estimation of Underlying Data Distribution

The estimation of mutual information is a notorious hard problem in high-dimensional space since the complicated underlying joint distribution of two high-dimensional variables is often criticized to be hard or impossible, which leads to a gap between the information-theoretic principle and its deep learning applications.

To bridge the gap, there exists two common strategies, i.e., varitional approximation [1, 27] and neural estimation [3]. As shown in Figure 6, varitional approximation [1, 27] aims to introduce an auxiliary neural network to estimate the mean and variance of the posterior distribution so as to fit the posterior distribution (since

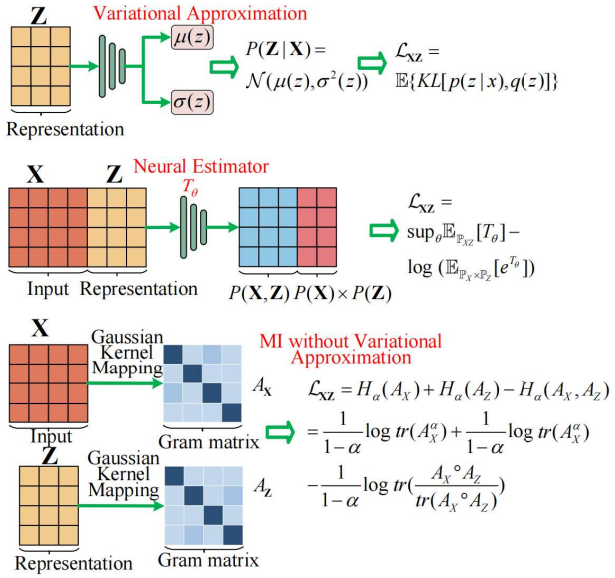

Figure 6. Mutual information measurement. Existing IB-based MVC methods utilize neural estimators to explicitly approximate the posterior distribution of representation or to fit the optimal function that can potentially estimate the data distribution. Differently, DIB is capable of fitting the mutual information by the eigenvalues of the Gram matrix $A_\mathbf{x}$ and $A_\mathbf{z}$ (where $A = \frac{G}{\mathrm{tr}(G)}$) without the need of variational approximation.

$D_{KL}\left(p\left(x,z\right)||p\left(x\right)p\left(z\right)\right) = D_{KL}\left(p\left(z|x\right)||p\left(z\right)\right)$ in an explicit distribution estimation manner. Neural estimation[3] resorts an neural estimator T_θ to fit the mutual information, which only requires to consider the expectations of original data and feature representation, and does not need to know the specific situation of joint and marginal distribution. The aforementioned methods transfer the challenging task of computing mutual information to the optimization process of neural network, which enables us to parameterize mutual information and employ it as an objective for optimization. However, the opaque function of neural networks significantly amplifies the uncertainty of the mutual information estimation process, as aforementioned methods rely heavily on these networks. In contrast, DIB fits the mutual information from the original data and feature representation directly through the Gram matrix as in **Proposition** 1. Inspired by this, we design a MI measurement without variational approximation, which has analytical gradients that allows us to parameterize the mutual information and optimize it as an objective.

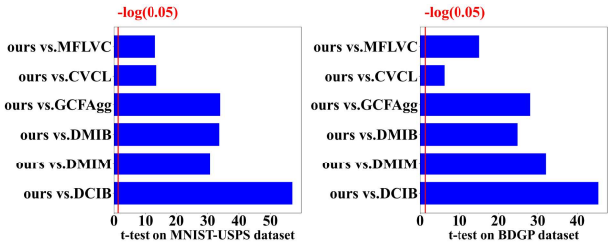

Figure 7. Significant test of DIB compared with several representative baselines on MNIST-USPS and BDGP datasets.

7.3. Formulation of Cluster Consistency

Similar to the feature consistency, we adopt contrastive learning to achieve cluster consistency. Specifically, each cluster labels $\mathbf{S}_j^v$ can form $(VK-1)$ label pairs $\{s_j^v, s_k^m\}_{k=1,\ldots,K}^{m=1,\ldots,V}$ with all labels except itself, where $\{s_j^v, s_j^m\}_{m\neq v}$ denotes $(V-1)$ positive labels pairs and $\{s_j^v, s_k^m\}_{k=1,\ldots,K}^{m=1,\ldots,V} - \{s_j^v, s_j^m\}_{m\neq v}$ denotes $V(K-1)$ negative labels pairs. And then, the cluster consistency objective $\mathcal{L}_{clu}$ between cluster labels $\{\mathbf{S}^v\}_{v=1}^V$ can be formulated as

$$\begin{aligned}\max \mathcal{L}_{clu} &= \sum_{v=1}^{V}\sum_{m\neq v} I(\mathbf{S}^v;\mathbf{S}^m) \\ &\approx \sum_{v=1}^{V}\sum_{m\neq v}\mathbb{E}\left[\log\frac{e^{d(s_j^v,s_j^m)}}{\sum_{d(s_j^v,s_k^l)\in Neg}^{K} e^{d(s_j^v,s_k^l)}}\right] \\ &+ V(V-1)\log N\end{aligned} \tag{19}$$

where $d(\cdot,\cdot)$ is the similarity measurement between two labels, and Neg represents negative label pairs.

8. Supplementary B

8.1. Significant Test

We conduct a significance test [33] to evaluate whether DIB's clustering performance is statistically better than the baseline methods. We use $-\log(p)$ values at a significance level of 0.05. A higher $-\log(p)$ value indicates a greater confidence in DIB's superiority. We select several promising baselines based on ACC for comparison. From Figure 7, DIB yields a higher $-\log(p)$ value than the baselines, signifying its statistical significance.

8.2. Comparison with variational approximation

In this study, we propose a novel mutual information measurement without variational approximation, which can fit the mutual information of high-dimensional spaces directly by eigenvalues of the normalized kernel Gram matrix. To further verify the effectiveness of the proposed MI measurement without variational approximation, we replace it in DIB with the variation approximation (with VA).

Table 3. Comparison of DIB on the proposed mutual information measurement and variational approximation on MNIST-USPS, BDGP and ESP datasets.

Dataset	Metric	with VA	DIB
MNIST-USPS	ACC	97.70	**99.86**
	NMI	97.10	**99.56**
	PUR	97.70	**99.86**
BDGP	ACC	96.92	**99.00**
	NMI	94.38	**96.65**
	PUR	96.92	**99.00**
ESP	ACC	52.46	**59.06**
	NMI	36.06	**37.77**
	PUR	55.15	**59.06**

From Table 3, we can observe DIB has higher ACC, NMI and PUR than with variational approximation on MNIST-USPS, BDGP and ESP datasets in this section, which demonstrates that fitting the mutual information directly from data can learn a better feature representation compared with variational approximation.