Deep Multiview Clustering by Contrasting Cluster Assignments —Supplementary Material

Jie Chen[1], Hua Mao[2], Wai Lok Woo[2], Xi Peng[1*]
[1] College of Computer Science, Sichuan University, China
[2] Department of Computer and Information Sciences, Northumbria University
chenjie2010@scu.edu.cn; {hua.mao,wailok.woo}@northumbria.ac.uk;
pengx.gm@gmail.com

1. Additional Experiments

Table 1. Computation times (in seconds) of the other contrastive learning-based methods on all the datasets.

Methods	MSRC-v1	COIL-20	Handwritten	BDGP	Scene-15	MNIST-USPS	Fashion
DSIMVC	676.78	480.59	2368.74	1234.33	1770.68	4236.72	2368.74
DCP	106.69	158.73	265.39	219.8	626.42	509.07	785.25
DSMVC	261.74	886.39	802.22	865.46	1233.89	1164.54	4130.62
MFL	123.81	496.78	688.11	68.31	1430.29	511.67	939.47
CVCL	**97.55**	**149.05**	**235.69**	**41.91**	**605.12**	**449.78**	**687.03**

1.1. Investigating the Computational Costs

We compare the proposed CVCL method with the other contrastive learning-based methods in terms of their computational costs. With the enhanced learning capabilities, the importance of the computational cost may become secondary to the improved performance achievable by contrastive learning-based methods. Table 1 shows the running times of all the competing algorithms on all the datasets. It is clear that CVCL performs more efficiently than the other algorithms. This demonstrates the advantages of the proposed CVCL method in terms of computational efficiency.

1.2. Discussion

The instances of a sample from different views may sit on different underlying distributions. This means that the contrastive learning of the high-level and low-level features may not be reasonable in MFL [1]. For a given sample, the results of cluster assignments of its instances from multiple views trend to be consistent in CVCL. In contrast with MFL, CVCL ensures consistency among the cluster assignments produced from multiple views. The semantic label of each sample can be predicted using Eq. (14). Moreover, we provide a theoretical analysis for soft cluster assignment alignment. This explains why CVCL performs significantly better than MFL on some of the datasets.

2. Detailed Proofs

2.1. Proof of Theorem 1

Theorem 1 *Assume that there are N samples and K clusters. Given two views v_1 and v_2 and $l^{(v_1,v_2)}$ in Eq. (6), the following inequality holds:*

$$l^{(v_1,v_2)} \geq e^{\log(2K-1)-N/\tau}.$$

Proof *Let $\mathbf{p}_j^{(v_1)}$ and $\mathbf{p}_j^{(v_2)}$ be the jth columns of $\mathbf{P}^{(v_1)}$ and $\mathbf{P}^{(v_2)}$, respectively. The ith elements $p_{ij}^{(v_1)}$ and $p_{ij}^{(v_2)}$ in $\mathbf{p}_j^{(v_1)}$ and $\mathbf{p}_j^{(v_2)}$ represent the cluster assignment probabilities, i.e., $0 \leq p_{ij}^{(v_1)} \leq 1$ and $0 \leq p_{ij}^{(v_2)} \leq 1$, respectively, where $1 \leq i \leq N$. Thus, we have*

$$0 \leq s\left(\mathbf{p}_j^{(v_1)}, \mathbf{p}_j^{(v_2)}\right) \leq N \quad \textit{and} \quad e^{s\left(\mathbf{p}_j^{(v_1)}, \mathbf{p}_j^{(v_1)}\right)} \geq 1.$$

Suppose that

$$l = -\frac{e^{s\left(\mathbf{p}_k^{(v_1)}, \mathbf{p}_k^{(v_2)}\right)/\tau}}{\sum\limits_{j=1, j\neq k}^{K} e^{s\left(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_1)}\right)/\tau} + \sum\limits_{j=1}^{K} e^{s\left(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_2)}\right)/\tau}}$$

and we obtain

$$\begin{aligned} \log l = & \log\left(\sum_{j=1, j\neq k}^{K} e^{s\left(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_1)}\right)/\tau} + \sum_{j=1}^{K} e^{s\left(\mathbf{p}_j^{(v_1)}, \mathbf{p}_k^{(v_2)}\right)/\tau}\right) \\ & - s\left(\mathbf{p}_k^{(v_1)}, \mathbf{p}_k^{(v_2)}\right)/\tau \\ & \geq \log\left(2K-1\right) - N/\tau. \end{aligned}$$

Hence,

$$l^{(v_1,v_2)} \geq e^{\log(2K-1)-N/\tau}.$$

□

[*] Corresponding author

2.2. Proof of Theorem 2

Theorem 2 *For n_v given views of multiview data, L_c in Eq. (7) is minimized if f is strictly aligned $\forall v_1, v_2 \in \{1, 2, ...n_v\}$ and $v_1 \neq v_2$.*

Proof *According to $\mathbf{p}_i^{(v_1)} = \mathbf{p}_i^{(v_2)}$ and the result of $p_{ik}^{(v_1)}$, we obtain*

$$s\left(\mathbf{p}_i^{(v_1)}, \mathbf{p}_i^{(v_2)}\right) = \left(\mathbf{p}_i^{(v_1)}\right)^T \mathbf{p}_i^{(v_2)} = k_i$$

where k_i equals the number of samples in the ith cluster. Similarly,

$$s\left(\mathbf{p}_i^{(v_1)}, \mathbf{p}_j^{(v_1)}\right) = s\left(\mathbf{p}_i^{(v_1)}, \mathbf{p}_j^{(v_2)}\right) = 0.$$

Hence,

$$l^{(v_1,v_2)} = e^{\log(2K-1)-N/\tau}.$$

This shows that L_c in Eq. (7) is minimized. □

References

[1] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He. Multi-level feature learning for contrastive multi-view clustering. In *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 16051–16060, New Orleans, Louisiana, USA, Jun. 2022. 1