Adversarially Robust Deep Multi-View Clustering: A Novel Attack and Defense Framework[1]

2024年7月16日
Tuesday

Haonan Huang

[1] Haonan Huang, Guoxu Zhou, Yanghang Zheng, Yuning Qiu, Andong Wang, and Qibin Zhao, “Adversarially Robust Deep Multi-View Clustering: A Novel Attack and Defense Framework,” in Forty-first International Conference on Machine Learning (ICML 2024).

Outline

01 Background

02 Attack: Adversaries to DMVC Models

03 Defense: Adversarially Robust DMVC

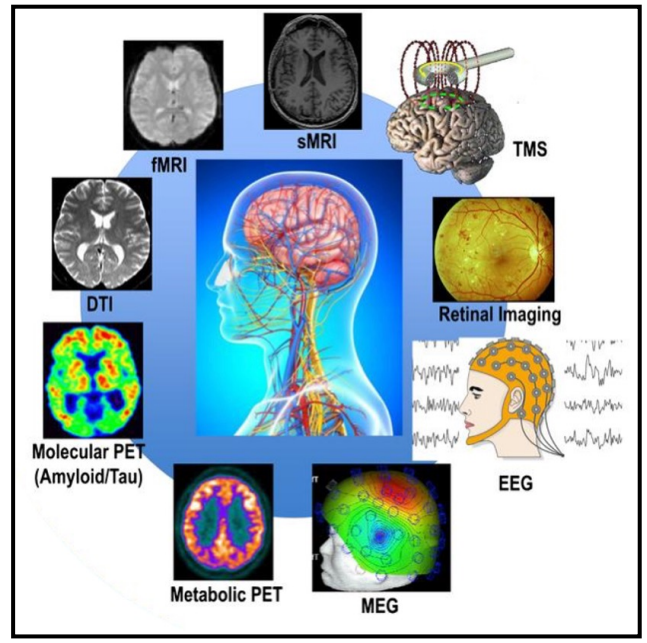
Background: Multi-view Data

Multi-view Data: Multi-view data is very common in real life, and it contains rich information. How to effectively utilize multi-view information to improve model's performance is a classic and challenging topic.

a) multilingual books

b) multi-angle images

c) Multiple sensor data

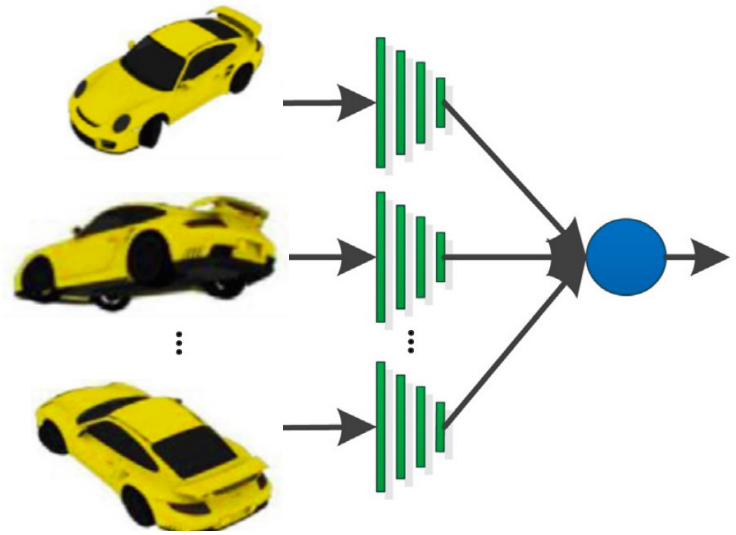


d) Multi-view Brain signals

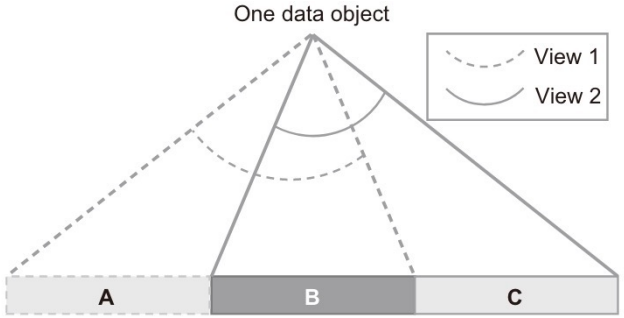
Examples of multi-view data

Background: Deep Multi-view Clustering

Deep Multi-view Clustering (DMVC) has shown to be a successful technique for enhanced and robust clustering by leveraging diverse data sources.



Deep Multi-view Learning

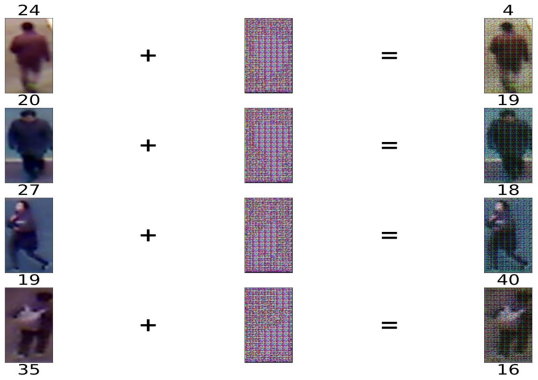


A and C denote complementarity;
B denotes the consistency

Liu, Jing, et al. "Partially shared latent factor learning with multiview data." IEEE TNNLS, 26.6 (2014): 1233-1246.
Zhou, Lihua, et al. "A Survey and an Empirical Evaluation of Multi-view Clustering Approaches." ACM Computing Surveys 56.7 (2024): 1-38.

Background: Adversarial Attack

Adversarial Attack：



Motivation1：*Is it possible to attack the unsupervised multi-view clustering model?* 🤔

Motivation2： *Is it possible to develop the unsupervised adversarial robust DMVC model?* 🤔

Our main contributions:

🌟 **Adversarial Attack Framework:** For the *first* time, an adversarial attack method for DMVC is introduced based on attacking loss functions related to the multi-view **Complementarity and Consistency**.

🌟 **Defense Mechanism:** For the *first* time, Adversarially Robust Deep Multi-View Clustering (AR-DMVC), is introduced to enhance the robustness of DMVC against adversarial attacks, especially in unsupervised setting.

🌟 **More Robust Technique:** Based on **Information-Theoretic Perspective**, AR-DMVC-AM, is proposed to mitigate attacks by minimizing the mutual information between adversarial examples and clustering assignments.

How to define the attack of multi-view clustering models?

Definition 1: The attack aims to introduce minimal perturbations to images used as input for the MVC model while staying within a defined noise threshold. This intentional perturbation is designed to cause misclustering of these samples by the model, leading to a notable decrease in performance, as quantified by various evaluation metrics.:

Number of Views

DMVC Model

Constraint of Perturbation

$$\max_{\delta^v} \sum_{v=1}^{V} \|\mathbf{z}^v - \mathcal{C}(\mathbf{x}^v + \delta^v)\|^2 \quad \text{s.t. } \|\delta^v\|^2 \leq \epsilon,$$

Original Output of DMVC

Sample

Perturbation

How to define the attack of multi-view clustering models?

Definition 2: (Attacking Multi-view Complementarity and Consistency) A successful attack method disrupts both complementarity and consistency when it induces noticeable differences between the pre-attack and post-attack states of learned view-specific representations and the consensus representation through the target model.

view-specific representation before the attack

view-specific representation post the attack

$$\mathcal{L}_{\text{a-com}} := \sum_{v=1}^{V} \mathbb{E}_{\mathbf{x}^v} \left\| \mathcal{C}^v\left(\mathbf{x}^v\right) - \mathcal{C}^v\left(\mathbf{x}^v + \mathcal{G}\left(\mathbf{x}^v\right)\right) \right\|^2,$$

$$\mathcal{L}_{\text{a-con}} := \mathbb{E}_{\{\mathbf{x}^v\}_v} \left\| \mathcal{C}\left(\{\mathbf{x}^v\}_v\right) - \mathcal{C}\left(\{\mathbf{x}^v + \mathcal{G}\left(\mathbf{x}^v\right)\}_v\right) \right\|^2.$$

common representation before the attack

common representation after the attack

How to define the attack of multi-view clustering models?

■ How to train the Generator and Discriminator:

Vanilla Minimax GAN Loss

$$\mathcal{L} := \sum_{v=1}^{V} \mathbb{E}_{\mathbf{x}^v}[\log(\mathcal{D}(\mathbf{x}^v)) + \log(1 - \mathcal{D}(\mathbf{x}^v + \mathcal{G}(\mathbf{x}^v)))].$$

Restriction on the Perturbation

$$\mathcal{L}_{\text{constraint}} := \sum_{v=1}^{V} \mathbb{E}_{\mathbf{x}^v}\left[\min\left\{\epsilon - \|\mathcal{G}(\mathbf{x}^v)\|^2, 0\right\}\right].$$

$$\max_{\mathcal{D}} \min_{\mathcal{G}} \mathcal{L} - \mu_1 \mathcal{L}_{\text{a-com}} - \mu_2 \mathcal{L}_{\text{a-con}} - \mu_3 \mathcal{L}_{\text{constraint}}$$

$$\mathcal{L}_{\text{a-com}} := \sum_{v=1}^{V} \mathbb{E}_{\mathbf{x}^v} \left\| \mathcal{C}^v(\mathbf{x}^v) - \mathcal{C}^v(\mathbf{x}^v + \mathcal{G}(\mathbf{x}^v)) \right\|^2,$$

The Attack of Complementarity

$$\mathcal{L}_{\text{a-con}} := \mathbb{E}_{\{\mathbf{x}^v\}_v} \left\| \mathcal{C}(\{\mathbf{x}^v\}_v) - \mathcal{C}(\{\mathbf{x}^v + \mathcal{G}(\mathbf{x}^v)\}_v) \right\|^2.$$

The Attack of Consistency

Remark: If we solely attack the complementarity of multiple views (i.e., optimizing only $\mathcal{L}_{\text{a-com}}$), we may fail to disrupt the final learned consensus representation, potentially yielding identical results before and after the attack. Similarly, if we exclusively target the consistency of multiple views (i.e., optimizing only $\mathcal{L}_{\text{a-con}}$), we cannot ensure that each view has been adequately attacked, potentially affecting only a subset of views. Therefore, our model is rational, as it ensures that each view is attacked while preserving a consensus representation of changes.

Adversarially Robust Deep Multi-View Clustering

- We first define the basic DMVC loss:

$$\mathcal{L}_{\text{CL-MVC}}\left(\mathbf{x}_i^u, \mathbf{x}_i^v; \theta\right) := \mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{DDC}},$$

Contrastive Loss

Clustering Module

- Adversarial Training-DMVC Loss (AR-DMVC):

Basic DMVC Loss

adversarial training is adopted

$$\mathcal{L}_{\text{AR-DMVC}} = \mathcal{L}_{\text{CL-MVC}}\left(\mathbf{x}_i^u, \mathbf{x}_i^v; \theta\right) + \lambda \mathcal{L}_{\text{CL-MVC}}\left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v; \theta\right)$$

$$\text{where } \tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v = \underset{\substack{\tilde{\mathbf{x}}_i^u \in \mathcal{B}_\epsilon[\mathbf{x}_i^v] \\ \tilde{\mathbf{x}}_i^v \in \mathcal{B}_\epsilon[\mathbf{x}_i^v]}}{\arg\max} \mathcal{L}_{\text{CL}}\left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v; \theta\right)$$

contrastive loss for the adversarial multi-view data

Adversarially Robust Deep Multi-View Clustering

Attack Mitigator: From the information-theoretic perspective, we introduce the conditional mutual information to measure the information between the adversarial input $\tilde{\mathbf{x}}$ and the corresponding predictive clustering assignment $\tilde{\mathbf{a}}$, i.e.,

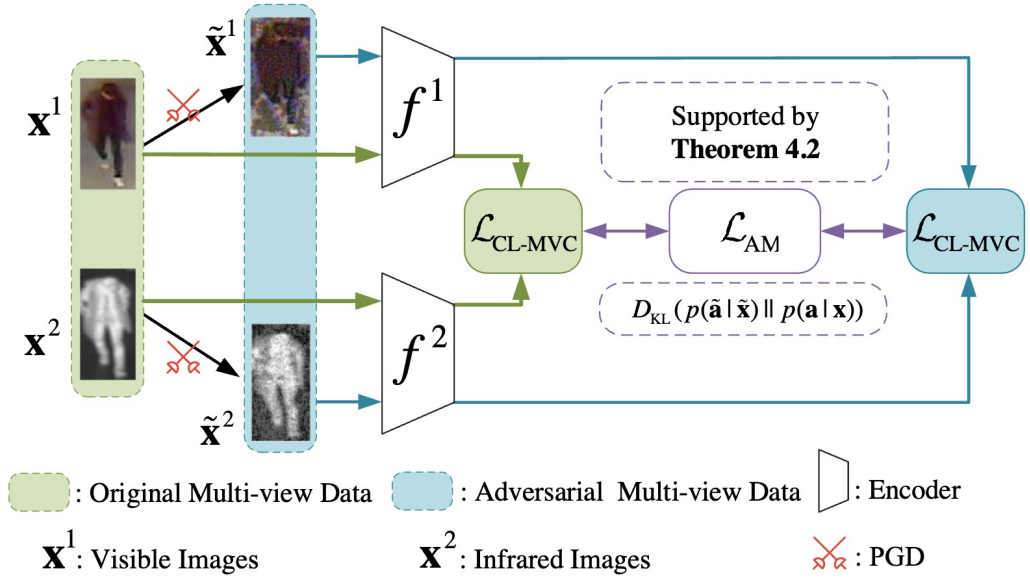
$$I\left(\tilde{\mathbf{x}}; \tilde{\mathbf{a}} \mid \mathbf{x}\right)$$

Theorem 4.2 (it can be upper-bounded in a more simplified formulation.)

$$\begin{aligned} I\left(\tilde{\mathbf{x}}; \tilde{\mathbf{a}} \mid \mathbf{x}\right) &= \mathbb{E}_{\tilde{\boldsymbol{x}}, \boldsymbol{x} \sim p(\tilde{\mathbf{x}}, \mathbf{x})} \mathbb{E}_{\boldsymbol{a} \sim p(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}})}\left[\log \frac{p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right)}{p\left(\tilde{\boldsymbol{x}} \mid \mathbf{x}\right)}\right] \\ &= \mathbb{E}_{\tilde{\boldsymbol{x}}, \boldsymbol{x} \sim p(\tilde{\mathbf{x}}, \mathbf{x})} \mathbb{E}_{\boldsymbol{a} \sim p(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}})}\left[\log \frac{p\left(\tilde{\mathbf{a}} \mid \tilde{\boldsymbol{x}}\right) p\left(\mathbf{a} \mid \boldsymbol{x}\right)}{p\left(\mathbf{a} \mid \boldsymbol{x}\right) p\left(\tilde{\mathbf{a}} \mid \boldsymbol{x}\right)}\right] \\ &= D_{\mathrm{KL}}\left(p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right) \| p\left(\mathbf{a} \mid \mathbf{x}\right)\right) - D_{\mathrm{KL}}\left(p\left(\mathbf{a} \mid \tilde{\mathbf{x}}\right) \| p\left(\mathbf{a} \mid \mathbf{x}\right)\right) \\ &\leq D_{\mathrm{KL}}\left(p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right) \| p\left(\mathbf{a} \mid \mathbf{x}\right)\right). \end{aligned}$$

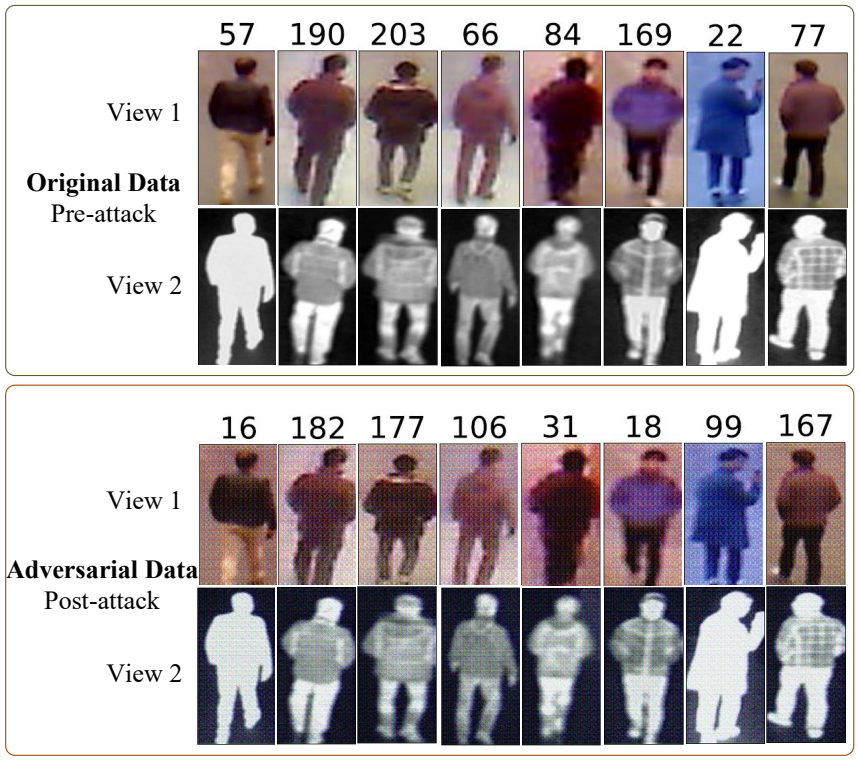
$$\begin{aligned} \mathcal{L}_{\text{AR-DMVC-AM}} = & \mathcal{L}_{\text{CL-MVC}}\left(\mathbf{x}_i^u, \mathbf{x}_i^v; \theta\right) \\ & + \lambda \mathcal{L}_{\text{CL-MVC}}\left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v; \theta\right) \\ & + \gamma D_{\mathrm{KL}}\left(p\left(\tilde{\mathbf{a}} \mid \tilde{\mathbf{x}}\right) \| p\left(\mathbf{a} \mid \mathbf{x}\right)\right) \end{aligned}$$

$$\text{where } \tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v = \underset{\substack{\tilde{\mathbf{x}}_i^u \in \mathcal{B}_\epsilon[\mathbf{x}_i^u] \\ \tilde{\mathbf{x}}_i^v \in \mathcal{B}_\epsilon[\mathbf{x}_i^v]}}{\arg\max} \mathcal{L}_{\text{CL}}\left(\tilde{\mathbf{x}}_i^u, \tilde{\mathbf{x}}_i^v; \theta\right).$$



Experimental Results

Adversarial samples generated by our attack on RegDB.



Attack Visualization Results

Table 1. Pre-attack (PRE) and post-attack (POST) performance for deep multi-view clustering models on four datasets.

Model		RegDB			NoisyFashion			NoisyMNIST			PatchedMNIST		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
EAMC (CVPR'20)	Pre	0.64	0.86	0.62	0.57	0.70	0.52	0.74	0.88	0.75	0.62	0.17	0.20
	Post	*0.33*	*0.57*	*0.23*	*0.30*	*0.20*	*0.11*	*0.25*	*0.11*	*0.06*	*0.53*	*0.13*	*0.15*
SiMVC (CVPR'21)	Pre	0.56	0.86	0.54	0.54	0.53	0.37	0.91	0.94	0.90	0.79	0.44	0.49
	Post	*0.30*	*0.61*	*0.24*	*0.30*	*0.25*	*0.13*	*0.29*	*0.20*	*0.12*	*0.49*	*0.13*	*0.12*
CoMVC (CVPR'21)	Pre	0.45	0.73	0.38	0.69	0.71	0.59	0.99	0.99	0.99	0.81	0.48	0.52
	Post	*0.25*	*0.47*	*0.14*	*0.40*	*0.35*	*0.25*	*0.31*	*0.20*	*0.13*	*0.61*	*0.20*	*0.21*
Multi-VAE (ICCV'21)	Pre	0.47	0.76	0.40	0.64	0.66	0.54	0.84	0.89	0.82	0.76	0.41	0.45
	Post	*0.43*	*0.71*	*0.33*	*0.47*	*0.43*	*0.30*	*0.46*	*0.39*	*0.28*	*0.51*	*0.19*	*0.16*
AECoDDC (CVPR'23)	Pre	0.43	0.72	0.36	0.78	0.78	0.70	0.99	0.99	0.99	0.65	0.21	0.29
	Post	*0.23*	*0.46*	*0.13*	*0.39*	*0.39*	*0.23*	*0.24*	*0.11*	*0.06*	*0.46*	*0.11*	*0.10*
InfoDDC (CVPR'23)	Pre	0.26	0.58	0.20	0.46	0.42	0.26	0.78	0.86	0.75	0.61	0.28	0.67
	Post	*0.22*	*0.50*	*0.11*	*0.25*	*0.19*	*0.10*	*0.33*	*0.22*	*0.14*	*0.53*	*0.13*	*0.15*
SEM (NeurIPS'23)	Pre	0.40	0.67	0.30	0.85	0.85	0.79	0.62	0.61	0.42	0.48	0.26	0.22
	Post	*0.33*	*0.63*	*0.21*	*0.31*	*0.30*	*0.14*	*0.21*	*0.11*	*0.07*	*0.45*	*0.16*	*0.14*
AR-DMVC (ours)	Pre	0.55	0.84	0.48	0.68	0.69	0.56	0.99	0.99	0.99	0.83	0.52	0.58
	Post	*0.42*	*0.66*	*0.31*	*0.54*	*0.48*	*0.33*	*0.90*	*0.79*	*0.80*	*0.65*	*0.34*	*0.36*
AR-DMVC-AM (ours)	Pre	0.54	0.85	0.50	0.69	0.73	0.59	0.99	0.99	0.99	0.81	0.46	0.52
	Post	***0.52***	***0.79***	***0.42***	***0.67***	***0.67***	***0.55***	***0.93***	***0.85***	***0.85***	***0.74***	***0.35***	***0.40***

Clustering Performance

Thanks for your interest.

Paper

Code