DM2C: Deep Mixed-Modal Clustering

Yangbangyan Jiang, Qianqian Xu, **Zhiyong Yang**,
Xiaochun Cao, Qingming Huang

Institute of Information Engineering, CAS
University of Chinese Academy of Sciences
Institute of Computing Technology, CAS
Key Lab. of BDKM, CAS
Peng Cheng Lab.

Why multiple modalities?

Ubiquitous multi-modal data

- The related information among multiple modalities helps us to understand the data.

Supervised Learning under Multiple Modalities

- Supervision comes from class labels and modality pairing.
  - Modality pairing: a sample in modality A and another sample in modality B represent the same instance.
- Manual annotations: expensive and laborious. When involving multiple modalities, the labeling is even more complicated than that for single modal data.
- We turn to unsupervised learning under multiple modalities since it works without data labels.

Mixed-modal Setting: Fully-unsupervised Learning

- Traditional unsupervised multi-modal learning still requires extra pairing information among modalities for feature alignment.
  - *E.g.*, partial modality pairing, 'must/cannot link' constraints, co-occurrence frequency...
- Mixed-modal data: each instance is represented in only one modality.

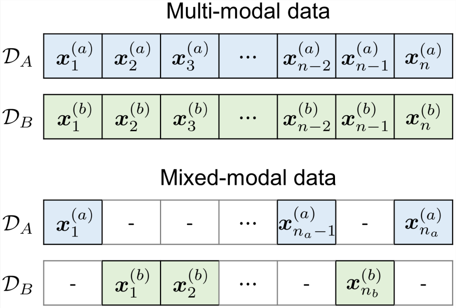
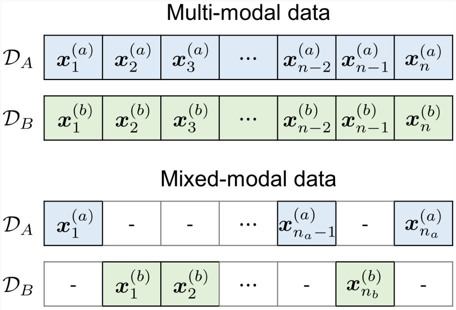


Figure 1: Examples of multi-modal and mixed-modal data with two modalities.

Mixed-modal Clustering: The Goal



- Dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{n}$ mixed from two modalities.
- $\mathcal{D} \rightarrow \{\mathbf{x}_i^{(a)}\}_{i=1}^{n_a} \cup \{\mathbf{x}_j^{(b)}\}_{j=1}^{n_b}$, where $n = n_a + n_b$.
- Mixed-modal clustering aims at learning unified representations for the modalities and then grouping the samples into k categories.

How to Learn Unified Representations?

Choice 1: learn a joint semantic space for all the modalities

- hard to find the correlation among all the modalities when pairing information is not available

Choice 2: learn the translation across the modalities

- easy to obtain the cross-modal mappings under the guidance of *cycle-consistency*
- modality unifying: transforming all the samples into a specific modality space

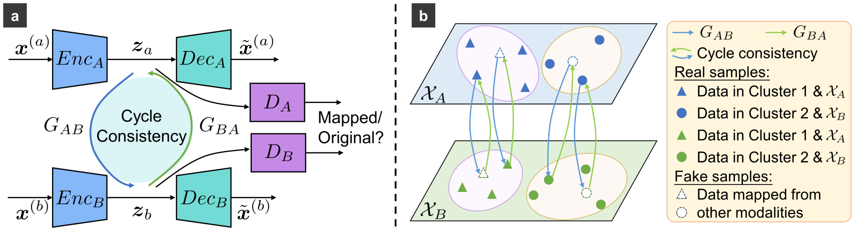
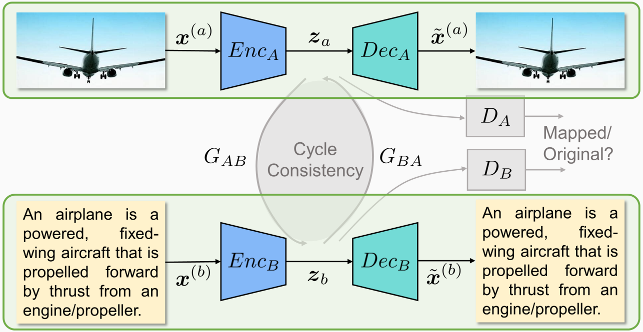


Figure 2: Overview of the proposed method.

Modules

- **Modality-specific auto-encoders**: to learn latent representations for each modality.
- **Cross-modal generators**: to learn mappings across modalities with unpaired data.
- **Discriminators**: to distinguish whether a sample is mapped from other modality spaces.

Framework: Module I

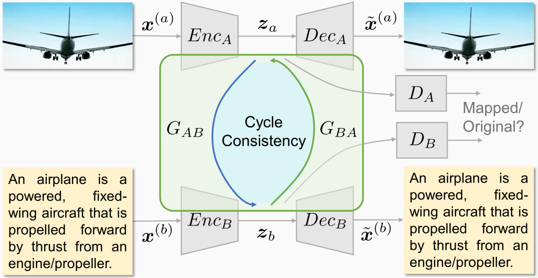


Modality-specific auto-encoders

Latent representations for each modality are learned by single-modal data reconstruction:

$$\begin{aligned} \mathcal{L}^{A}_{\text{rec}}(\Theta_{AE_A}) &= \|x_i^{(a)} - Dec_A(Enc_A(x_i^{(a)}))\|_2^2, \\ \mathcal{L}^{B}_{\text{rec}}(\Theta_{AE_B}) &= \|x_i^{(b)} - Dec_B(Enc_B(x_i^{(b)}))\|_2^2. \end{aligned} \tag{1}$$

Framework: Module II



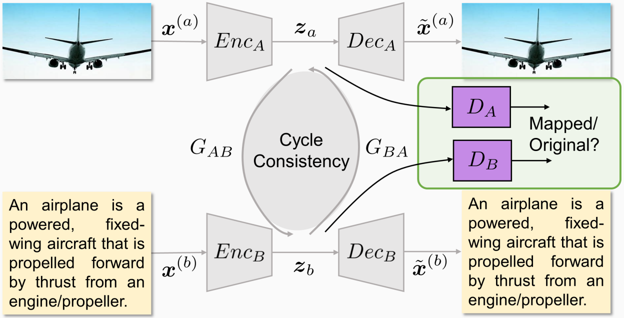
Cross-modal generators

Mappings across modalities are constrained by *cycle-consistency*:

$$\begin{aligned} \mathcal{L}^{A}_{cyc}(\Theta_{G_{AB}}, \Theta_{G_{BA}}) &= \mathbb{E}_{z_a \sim \mathcal{X}_A}\left[\|z_a - G_{BA}(G_{AB}(z_a))\|_1\right], \\ \mathcal{L}^{B}_{cyc}(\Theta_{G_{AB}}, \Theta_{G_{BA}}) &= \mathbb{E}_{z_b \sim \mathcal{X}_B}\left[\|z_b - G_{AB}(G_{BA}(z_b))\|_1\right]. \end{aligned} \tag{2}$$

Generators: produce fake samples that are transformed from other modalities rather than originally lying in a specific modality space.

Framework: Module III

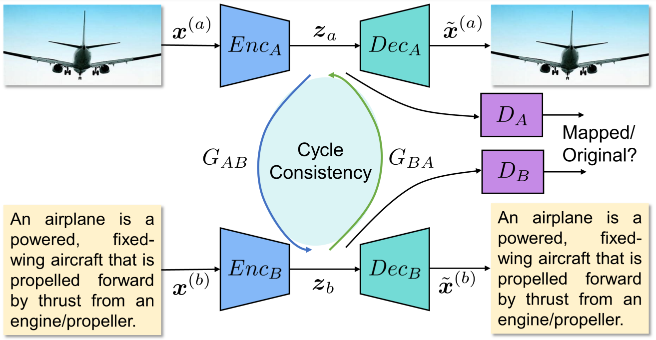


Discriminators

Discriminators: distinguish whether a sample is mapped from other modality spaces.

Games between generators and discriminators:

$$\mathcal{L}^{\mathrm{A}}_{\mathrm{adv}}(\boldsymbol{\Theta}_{G_{BA}}, \boldsymbol{\Theta}_{D_A}) = \mathbb{E}_{\boldsymbol{z}_a \sim \mathcal{X}_A}[D_A(\boldsymbol{z}_a)] - \mathbb{E}_{\boldsymbol{z}_b \sim \mathcal{X}_B}[D_A(G_{BA}(\boldsymbol{z}_b))],$$
$$\mathcal{L}^{\mathrm{B}}_{\mathrm{adv}}(\boldsymbol{\Theta}_{G_{AB}}, \boldsymbol{\Theta}_{D_B}) = \mathbb{E}_{\boldsymbol{z}_b \sim \mathcal{X}_B}[D_B(\boldsymbol{z}_b)] - \mathbb{E}_{\boldsymbol{z}_a \sim \mathcal{X}_A}[D_B(G_{AB}(\boldsymbol{z}_a))]. \quad (3)$$



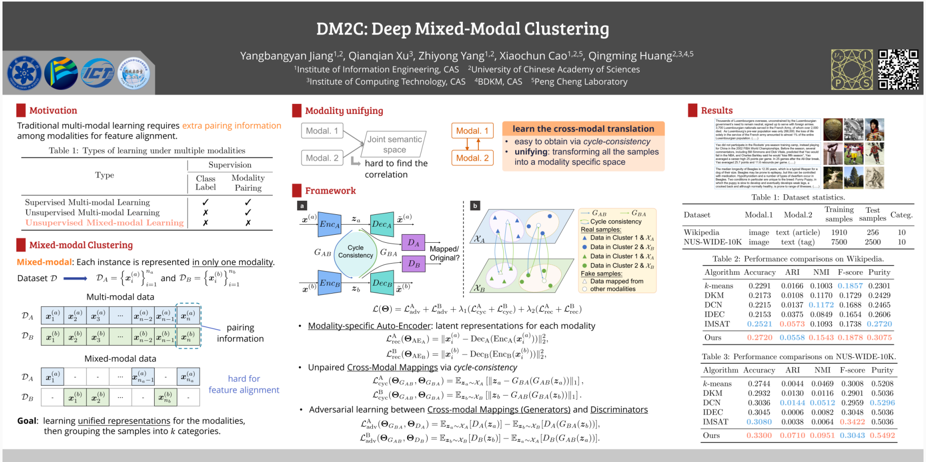
Objective Function

$$\min_{\substack{\Theta_{G_{AB}},\Theta_{G_{BA}}\\ \Theta_{AE_A},\Theta_{AE_B}}} \max_{\Theta_{D_A},\Theta_{D_B}} \mathcal{L}_{\text{adv}}^{\text{A}} + \mathcal{L}_{\text{adv}}^{\text{B}} + \lambda_1(\mathcal{L}_{\text{cyc}}^{\text{A}} + \mathcal{L}_{\text{cyc}}^{\text{B}}) + \lambda_2(\mathcal{L}_{\text{rec}}^{\text{A}} + \mathcal{L}_{\text{rec}}^{\text{B}}) \tag{4}$$

Thank You for Your Attention!

See you at the poster session!

Wed Dec 11th 10:45AM – 12:45PM @ East Exhibition Hall B+C #63



DM2C: Deep Mixed-Modal Clustering

Yangbangyan Jiang[1,2], Qianqian Xu[3], Zhiyong Yang[1,2], Xiaochun Cao[1,2,5], Qingming Huang[2,3,4,5]

[1]Institute of Information Engineering, CAS [2]University of Chinese Academy of Sciences

[3]Institute of Computing Technology, CAS [4]BDKM, CAS [5]Peng Cheng Laboratory

Motivation

Traditional multi-modal learning requires extra pairing information among modalities for feature alignment.

Table 1: Types of learning under multiple modalities

Type	Supervision: Class Label	Supervision: Modality Pairing
Supervised Multi-modal Learning	✓	✓
Unsupervised Multi-modal Learning	✗	✓
Unsupervised Mixed-modal Learning	✗	✗

Mixed-modal Clustering

Mixed-modal: Each instance is represented in only one modality.

Dataset $\mathcal{D}$ ⟶ $\mathcal{D}_A = \{x_i^{(a)}\}_{i=1}^{n_a}$ and $\mathcal{D}_B = \{x_i^{(b)}\}_{i=1}^{n_b}$

Multi-modal data

pairing information

Mixed-modal data

hard for feature alignment

Goal: learning unified representations for the modalities, then grouping the samples into k categories.

Modality unifying

Modal. 1, Modal. 2, Joint semantic space, hard to find the correlation

learn the cross-modal translation

- easy to obtain via *cycle-consistency*
- **unifying**: transforming all the samples into a modality specific space

Framework

Cycle Consistency, Mapped/Original?, G_{AB}, G_{BA}, Real samples: Data in Cluster 1 & $\mathcal{X}_A$, Data in Cluster 2 & $\mathcal{X}_B$, Data in Cluster 1 & $\mathcal{X}_A$, Data in Cluster 2 & $\mathcal{X}_B$; Fake samples: Data mapped from other modalities

$$\mathcal{L}(\Theta) = \mathcal{L}_{adv}^{A} + \mathcal{L}_{adv}^{B} + \lambda_1(\mathcal{L}_{cyc}^{A} + \mathcal{L}_{cyc}^{B}) + \lambda_2(\mathcal{L}_{rec}^{A} + \mathcal{L}_{rec}^{B})$$

- Modality-specific Auto-Encoder: latent representations for each modality

$$\mathcal{L}_{rec}^{A}(\Theta_{AE_A}) = \|x_i^{(a)} - \mathrm{Dec}_A(\mathrm{Enc}_A(x_i^{(a)}))\|_2^2,$$
$$\mathcal{L}_{rec}^{B}(\Theta_{AE_B}) = \|x_i^{(b)} - \mathrm{Dec}_B(\mathrm{Enc}_B(x_i^{(b)}))\|_2^2,$$

- Unpaired Cross-Modal Mappings via *cycle-consistency*

$$\mathcal{L}_{cyc}^{A}(\Theta_{G_{AB}}, \Theta_{G_{BA}}) = \mathbb{E}_{z_a\sim\mathcal{X}_A}[\|z_a - G_{BA}(G_{AB}(z_a))\|_1],$$
$$\mathcal{L}_{cyc}^{B}(\Theta_{G_{AB}}, \Theta_{G_{BA}}) = \mathbb{E}_{z_b\sim\mathcal{X}_B}[\|z_b - G_{AB}(G_{BA}(z_b))\|_1].$$

- Adversarial learning between Cross-modal Mappings (Generators) and Discriminators

$$\mathcal{L}_{adv}^{A}(\Theta_{G_{BA}}, \Theta_{D_A}) = \mathbb{E}_{z_a\sim\mathcal{X}_A}[D_A(z_a)] - \mathbb{E}_{z_b\sim\mathcal{X}_B}[D_A(G_{BA}(z_b))],$$
$$\mathcal{L}_{adv}^{B}(\Theta_{G_{AB}}, \Theta_{D_B}) = \mathbb{E}_{z_b\sim\mathcal{X}_B}[D_B(z_b)] - \mathbb{E}_{z_a\sim\mathcal{X}_A}[D_B(G_{AB}(z_a))].$$

Results

Table 1: Dataset statistics.

Dataset	Modal.1	Modal.2	Training samples	Test samples	Categ.
Wikipedia	image	text (article)	1910	256	10
NUS-WIDE-10K	image	text (tag)	7500	2500	10

Table 2: Performance comparisons on Wikipedia.

Algorithm	Accuracy	ARI	NMI	F-score	Purity
k-means	0.2291	0.0166	0.1003	0.1857	0.2301
DKM	0.2173	0.0108	0.1170	0.1729	0.2429
DCN	0.2215	0.0137	0.1172	0.1688	0.2465
IDEC	0.2153	0.0375	0.0849	0.1654	0.2606
IMSAT	0.2521	0.0573	0.1093	0.1738	0.2720
Ours	0.2720	0.0558	0.1543	0.1878	0.3075

Table 3: Performance comparisons on NUS-WIDE-10K.

Algorithm	Accuracy	ARI	NMI	F-score	Purity
k-means	0.2744	0.0044	0.0469	0.3008	0.5208
DKM	0.2932	0.0130	0.0116	0.2901	0.5036
DCN	0.3036	0.0144	0.0512	0.2959	0.5296
IDEC	0.3045	0.0006	0.0082	0.3048	0.5036
IMSAT	0.3080	0.0038	0.0064	0.3422	0.5036
Ours	0.3300	0.0710	0.0951	0.3043	0.5492