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 $\mathcal{D} = \{x_i\}_{i=1}^n$ mixed from two modalities.
- $\mathcal{D} \rightarrow \{x_i^{(a)}\}_{i=1}^{n_a} \cup \{x_j^{(b)}\}_{j=1}^{n_b}$, 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
Framework: Overview

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:

\[ \mathcal{L}_{\text{rec}}^A(\Theta_{AE_A}) = \| x^{(a)}_i - Dec_A(Enc_A(x^{(a)}_i)) \|_2^2, \]

\[ \mathcal{L}_{\text{rec}}^B(\Theta_{AE_B}) = \| x^{(b)}_i - Dec_B(Enc_B(x^{(b)}_i)) \|_2^2. \]
Framework: Module II

Cross-modal generators

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

\[
\mathcal{L}_{\text{cyc}}^A(\Theta_{G_{AB}}, \Theta_{G_{BA}}) = \mathbb{E}_{z_{a} \sim X_A} \left[ \| z_{a} - G_{BA}(G_{AB}(z_{a})) \|_1 \right],
\]

\[
\mathcal{L}_{\text{cyc}}^B(\Theta_{G_{AB}}, \Theta_{G_{BA}}) = \mathbb{E}_{z_{b} \sim X_B} \left[ \| z_{b} - G_{AB}(G_{BA}(z_{b})) \|_1 \right].
\]

(2)

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

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

Games between generators and discriminators:

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

(3)
Objective Function

$$\begin{align*}
\min_{\Theta_{G_{AB}}, \Theta_{G_{BA}}, \Theta_{D_A}, \Theta_{D_B}, \Theta_{A_{EA}}, \Theta_{A_{EB}}} & \quad \max_{\Theta_{G_{AB}}, \Theta_{G_{BA}}, \Theta_{D_A}, \Theta_{D_B}} \quad L^A_{\text{adv}} + L^B_{\text{adv}} + \lambda_1 (L^A_{\text{cyc}} + L^B_{\text{cyc}}) + \lambda_2 (L^A_{\text{rec}} + L^B_{\text{rec}}) \\
\end{align*}$$

(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
Yangbangyi Jiang1,2, Qianqian Xu3, Zhiyong Yang1,2, Xiaochun Cao1,2,5, Qingsning Huang1,3,4,5
1Institute of Information Engineering, CAS
2University of Chinese Academy of Sciences
3Institute of Computing Technology, CAS
4BDKM, CAS
5Peng Cheng Laboratory

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Supervision</th>
<th>Class Label</th>
<th>Modality Pairing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised Multi-modal Learning</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Unsupervised Multi-modal Learning</td>
<td>✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsupervised Mixed-modal Learning</td>
<td>✗</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Framework: Learn the cross-modal translation
- easy to obtain via cycle-consistency
- unifying: transforming all the samples into a modality specific space

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$\mathcal{D}_A$</th>
<th>$\mathcal{D}_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-modal data</td>
<td>$\mathcal{D}<em>A = {x_a^{(i)}}</em>{i=1}^m$</td>
<td>$\mathcal{D}<em>B = {x_b^{(i)}}</em>{i=1}^m$</td>
</tr>
</tbody>
</table>

Multi-modal data: $\mathcal{D}_A = \{x_a^{(i)}\}_{i=1}^m 
\mathcal{D}_B = \{x_b^{(i)}\}_{i=1}^m$

Mixed-modal data: $\mathcal{D}_A = \{x_a^{(i)}\}_{i=1}^m 
\mathcal{D}_B = \{x_b^{(i)}\}_{i=1}^m$

Goal: Learning unified representations for the modalities, then grouping the samples into 4 categories.

Results

<table>
<thead>
<tr>
<th>Table 1: Dataset statistics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Wikipedia</td>
</tr>
<tr>
<td>NUS-WIDE-10K</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Performance comparisons on Wikipedia.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm Accuracy</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>k-means</td>
</tr>
<tr>
<td>DCM</td>
</tr>
<tr>
<td>DCN</td>
</tr>
<tr>
<td>DBCEC</td>
</tr>
<tr>
<td>DBSAT</td>
</tr>
<tr>
<td>Ours</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Performance comparisons on NUS-WIDE-10K.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm Accuracy</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>k-means</td>
</tr>
<tr>
<td>DCM</td>
</tr>
<tr>
<td>DCN</td>
</tr>
<tr>
<td>DBCEC</td>
</tr>
<tr>
<td>DBSAT</td>
</tr>
<tr>
<td>Ours</td>
</tr>
</tbody>
</table>