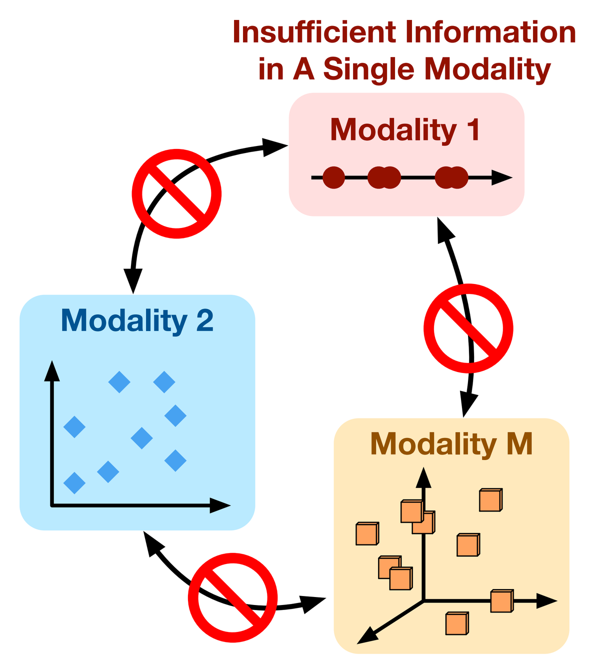
The 39th Annual AAAI Conference on Artificial Intelligence

An Optimal Transport-based Latent Mixer for Robust Multi-modal Learning

Fengjiao Gong, Angxiao Yue, Hongteng Xu

Renmin University of China, Beijing, China

Motivation

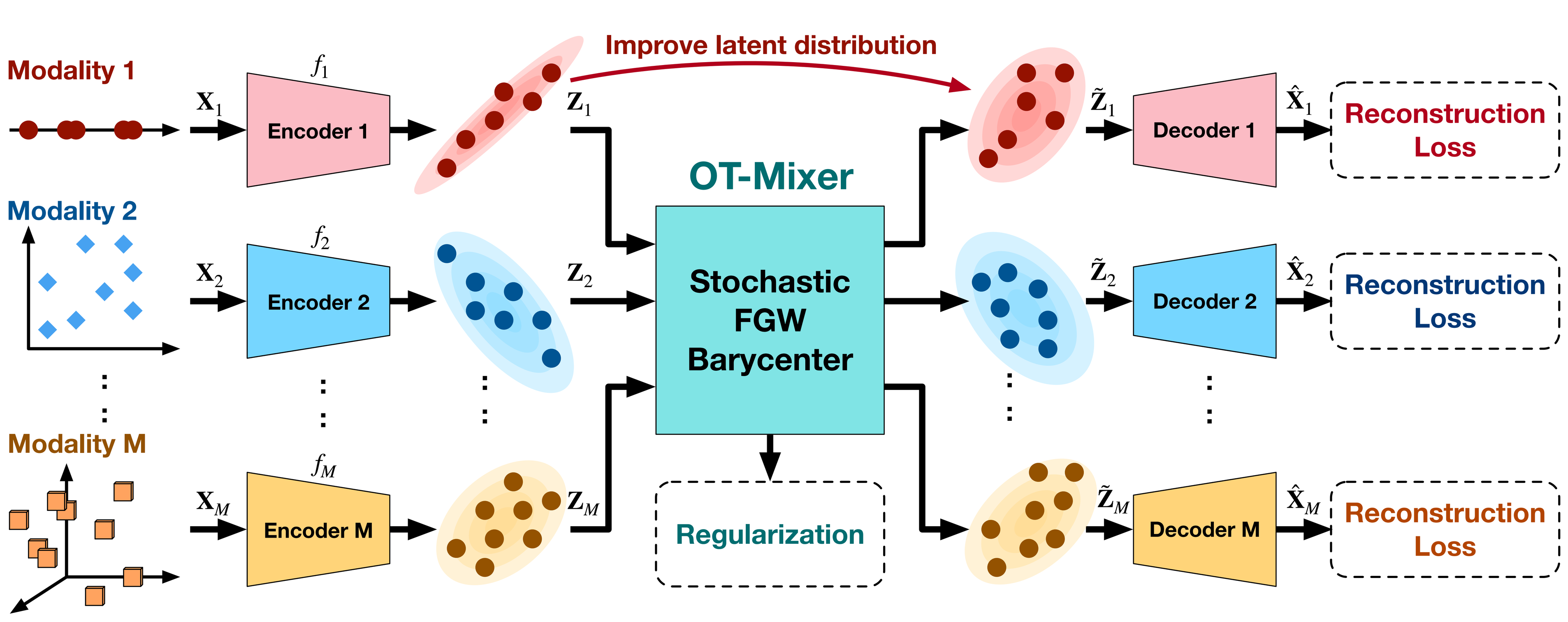


Practical scenario with unaligned and distributed multi-modal data.

- Real-world multi-modal data are often scattered to different local agents, and each agent can only access the data in a single modality.
- Due to privacy protection and data security, sharing data directly across different agents is forbidden in many applications.
- What is worse, for some agents, the data associated with its modality may be insufficient for downstream tasks because the number of the data can be limited and the features can be not informative enough for representation learning.

Proposed Method

Suppose that we have a set of multi-modal data, denoted as $\mathcal{D}=\{X_m\}_{m=1}^{M}$, where M is the number of modalities. The data of the m-th modality, i.e., $X_m=\{x_{m,j}\}_{j=1}^{N_m}\in\mathbb{R}^{N_m\times D_m}$, contains N_m D_m-dimensional samples.



The scheme of OTM-based multi-modal learning.

- Learning the WAE models

$$\min_{\{f_m,g_m\}_{m=1}^{M}}\sum_{m=1}^{M}\|X_m-\hat{X}_m\|_F^2+\lambda R(\{Z_m\}_{m=1}^{M}),\tag{1}$$

- Reconstruction loss of each modality

$$\mathcal{L}_{mix}=\sum_{m=1}^{M}\Big(\|X_m-g_m(Z_m)\|_F^2+\|Z_m-\tilde{Z}_m\|_F^2\Big),\tag{2}$$

- Regularization on the latent representations

$$\underbrace{\min_{T\in\Pi(\mu_B,\mu_C)}\sum_{i,j,k,l}|K(i,k)-I_C(j,l)|^2t_{ij}t_{kl}}_{GW(\tilde{Z}_B,I_C)},\quad \textit{clustering}$$
$$\mathcal{L}_{supervise}(h_m(\tilde{Z}_m),y_m)+\mathcal{L}_{supervise}(h_m(Z_m),y_m),\quad \textit{supervised tasks}\tag{3}$$

An Optimal Transport-based Latent Mixer

- FGW Distance $\mathrm{FGW}(Z,Z_m;\alpha)$

$$\min_{T_m\in\Pi(\mu,\mu_m)}\sum_{i,j,k,l}\alpha\underbrace{d_{\mathcal{Z}}^2(z_i,z_{m,j})t_{m,ij}}_{\text{Wasserstein term}}+(1-\alpha)\underbrace{|A(i,k)-A_m(j,l)|^2t_{m,ij}t_{m,kl}}_{\text{Gromov-Wasserstein term}},\tag{4}$$

- FGW Barycenter Problem

$$Z_B,\{T_m^*\}_{m=1}^{M}=\arg\min_{Z}\sum_{m=1}^{M}\mathrm{FGW}(Z,Z_m;\alpha),\tag{5}$$

- Stochastic Mixing

$$\tilde{Z}_B=N_B\sum_{i=1}^{M}M_m\odot(T_m^*Z_m),\text{ with }\sum_{m=1}^{M}M_m=1_{N_B\times d},\tag{6}$$

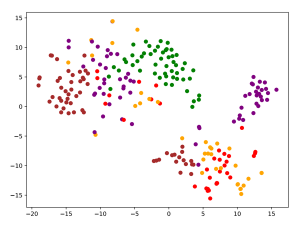
- Augmented Latent Code

$$\tilde{Z}_m=(T_m^*)^\top\tilde{Z}_B,\tag{7}$$

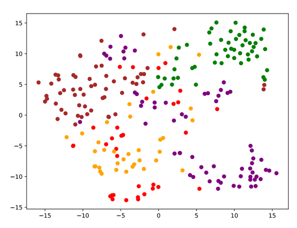
Numerical Comparisons

- Clustering Performance

Data type	Datasets / Algorithms	Caltech 7 Purity	ORL Purity	Movies Purity	Prokaryotic Purity
Aligned	MCCA	0.5313	0.3475	0.0989	0.5620
	DCCAE	0.4110	0.5625	0.1572	0.5070
	AttnAE	0.4600	0.4600	0.1880	0.5390
	MVKSC	0.5196	0.3013	0.2285	**0.6188**
	MultiNMF	0.4525	**0.6900**	0.1726	0.5771
	MWAE+OTM	<u>0.6072</u>	<u>0.6563</u>	<u>0.3177</u>	<u>0.5952</u>
	MWAE+OTM(WB)	**0.6097**	0.6525	**0.3184**	0.5541
Unaligned	MVC-UM	0.3112	0.5431	0.1841	0.4451
	GWMAC	0.3568	0.5118	0.1928	**0.5479**
	MWAE+OTM	**0.5788**	**0.6550**	**0.2925**	<u>0.5438</u>
	MWAE+OTM(WB)	<u>0.5667</u>	<u>0.6350</u>	<u>0.2860</u>	0.5299



Aligned



Fused

- Classification & Regression Performance

Dataset	Method	Result	Task
AV-MNIST	Late fusion	0.7295	Classification
	Late fusion + OTM	**0.7316**	
ENRICO	MI matrix	0.4815	Classification
	MI matrix + OTM	**0.5034**	
	Tensor matrix	0.4814	
	Tensor matrix + OTM	**0.4911**	
CMU-MOSI	Late fusion	0.5194	Classification
	Late fusion + OTM	**0.5368**	
	LRTF	0.5245	
	LRTF + OTM	**0.5327**	
	MFM	0.5391	
	MFM + OTM	**0.5410**	
	Late fusion	1.3710	Regression
	Late fusion + OTM	**1.3630**	
	Tenser fusion	1.3691	
	Tenser fusion + OTM	**1.3644**	
MUJOCO	Tensor fusion	1.583×10^{-3}	Regression
	Tensor fusion + OTM	$\mathbf{1.369}\times10^{-3}$	

Dataset (Task)	Method	Selected Modality 1	2	3
Prokaryotic (Clustering)	MWAE	0.4936	**0.6261**	0.4791
	MWAE+OTM	**0.5554**	0.5209	**0.5426**
CMU-MOSI (Classification)	Late fusion	0.5369	**0.5373**	0.5163
	Late fusion+OTM	**0.5428**	0.5328	**0.5268**
	LRTF	**0.5190**	**0.5241**	0.5131
	LRTF+OTM	0.5131	0.5222	**0.5209**
CMU-MOSI (Regression)	Late fusion	**1.3721**	**1.3581**	1.4055
	Late fusion+OTM	1.3804	1.3629	**1.3698**
	Tenser fusion	1.3684	1.3680	1.3853
	Tenser fusion+OTM	**1.3674**	**1.3669**	**1.3716**

Single Modality

Conclusion & Future Work

- We propose a novel optimal transport-based mixer (OTM) that achieves data alignment and augmentation for robust multi-modal learning.
- In the future, we plan to test our method in real-world applications, e.g., federated learning for healthcare data modeling.

Github Repo

Email:
hongtengxu@ruc.edu.cn
fengjiaogong2021@ruc.edu.cn
angxiaoyue@ruc.edu.cn