Deep Multimodal Clustering with Cross Reconstruction

Xianchao Zhang, Xiaorui Tang, Linlin Zong, Xinyue Liu, Jie Mu

Introduction

Deep Clustering

- Two-stage methods
- End-to-end methods

Multimodal Clustering

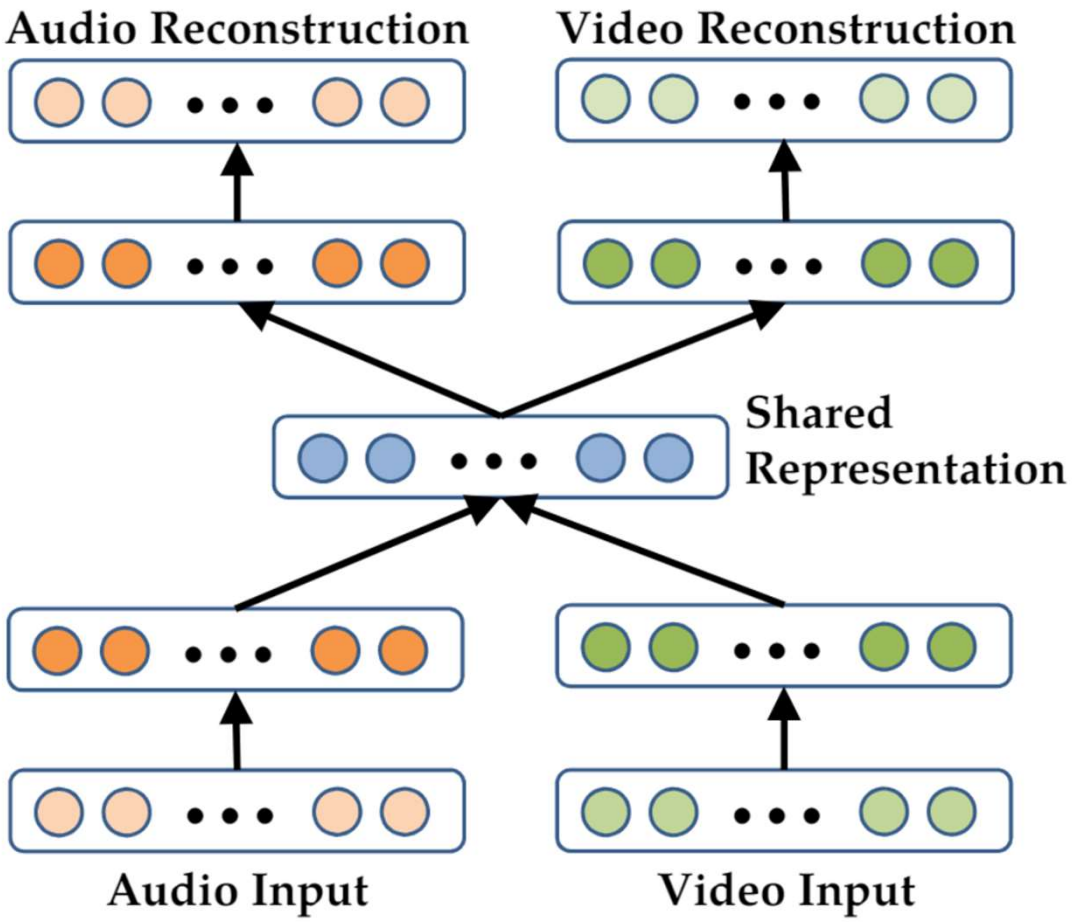
- Traditional clustering based methods
- Deep clustering based methods

Deep Multimodal Clustering

- Autoencoder based methods
- Deep Bolzmann Machine(DBM) based methods
- Deep canonical correlation analysis(DCCA) based methods

Introduction

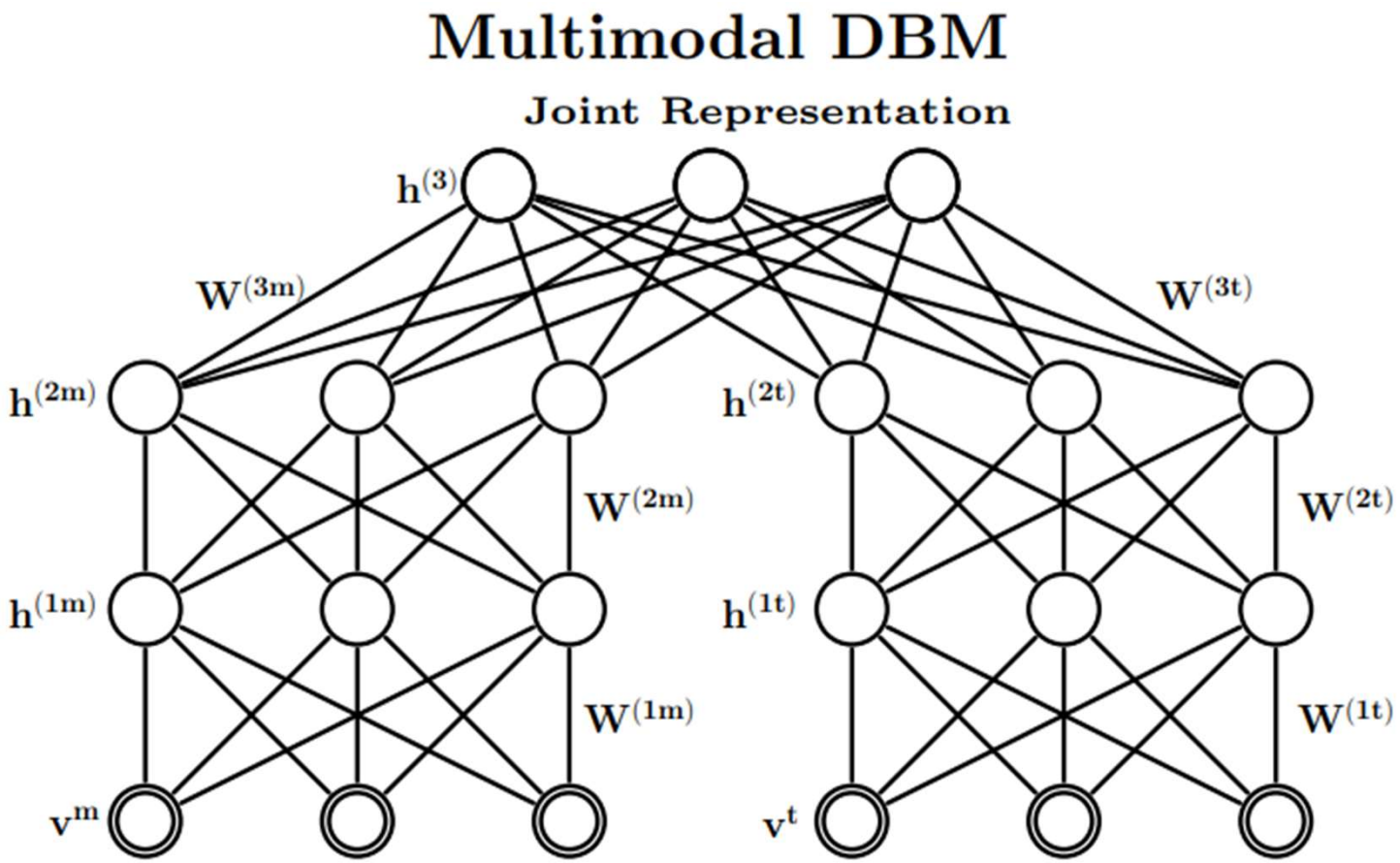
Autoencoder based methods



Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A.Y. (2011). Multimodal Deep Learning. ICML.

Introduction

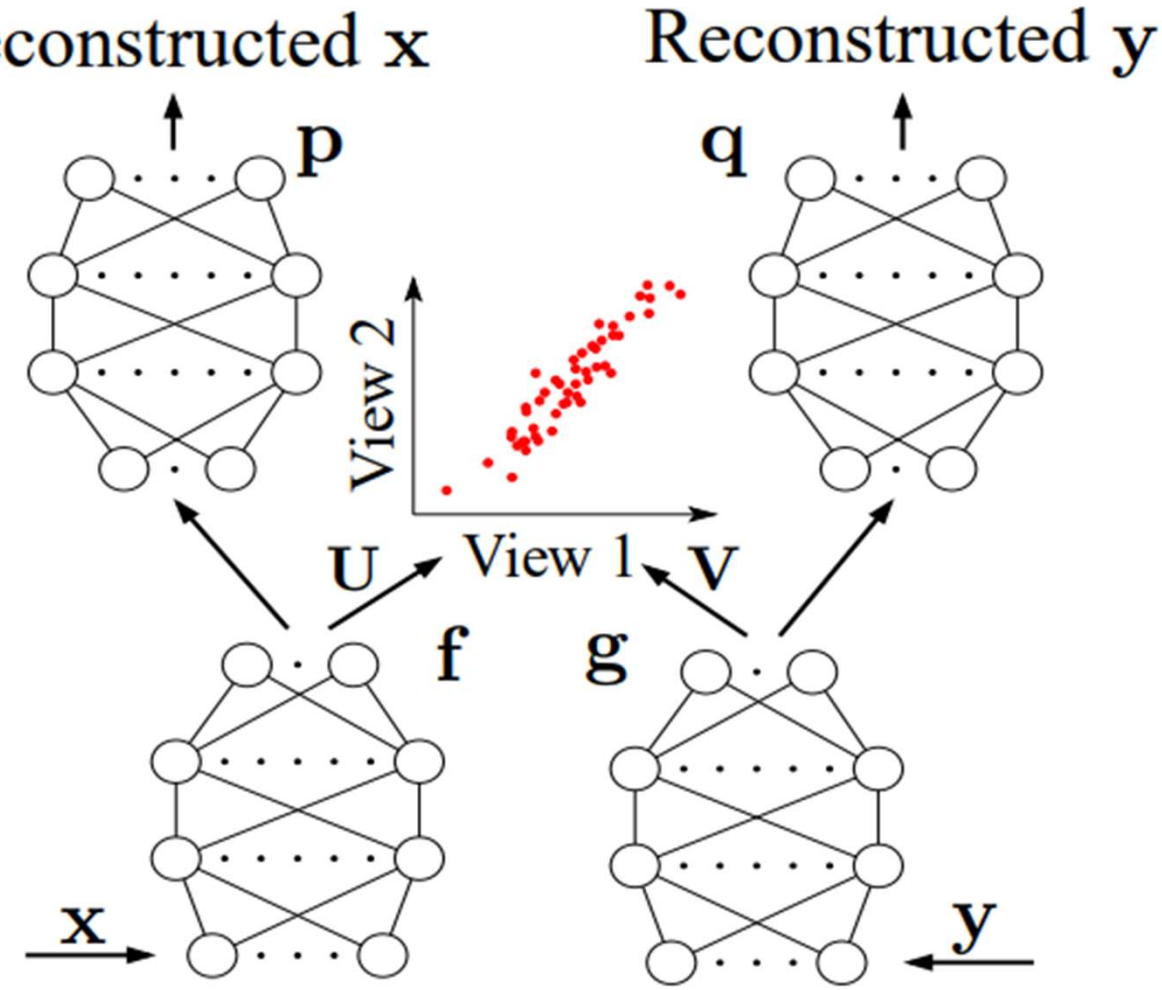
DBM based methods



Srivastava, N., & Salakhutdinov, R. (2012). Multimodal learning with deep Boltzmann machines. J. Mach. Learn. Res., 15, 2949-2980.

Introduction

DCCA based methods



Wang, W., Arora, R., Livescu, K., & Bilmes, J.A. (2015). On Deep Multi-View Representation Learning. ICML.

Introduction

Autoencoder based methods

Autoencoders do not discover the similarity of the common feature distributions.

DBM based methods

Due to the high computational costs in high-dimensional data space, the DBM based methods have not been widely studied in recent years.

DCCA based methods

DCCA based methods lack the analysis of probability theory, which makes it difficult to measure the distribution differences of different modalities.

Introduction

Contributions

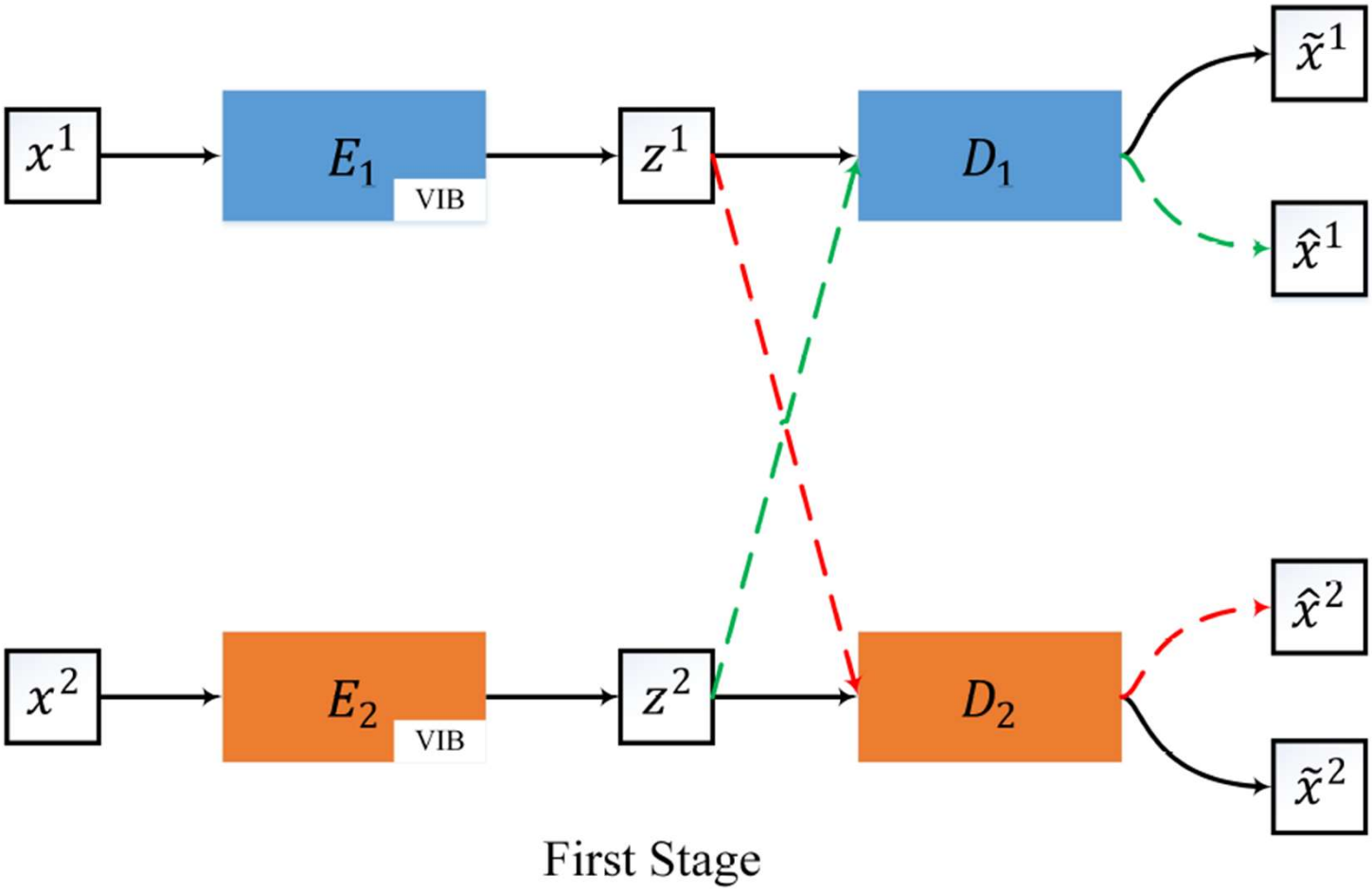
① We propose a novel deep multimodal clustering algorithm, which can effectively reduce the distribution differences among different modalities in feature space.

② We provide a theoretical analysis to prove that the proposed cross reconstruction method effectively reduce the distribution difference of different modalities in feature space.

③ Experiments show obviously improvement over state-of-the-art multimodal clustering methods on six benchmark multimodal datasets.

The Proposed Algorithm

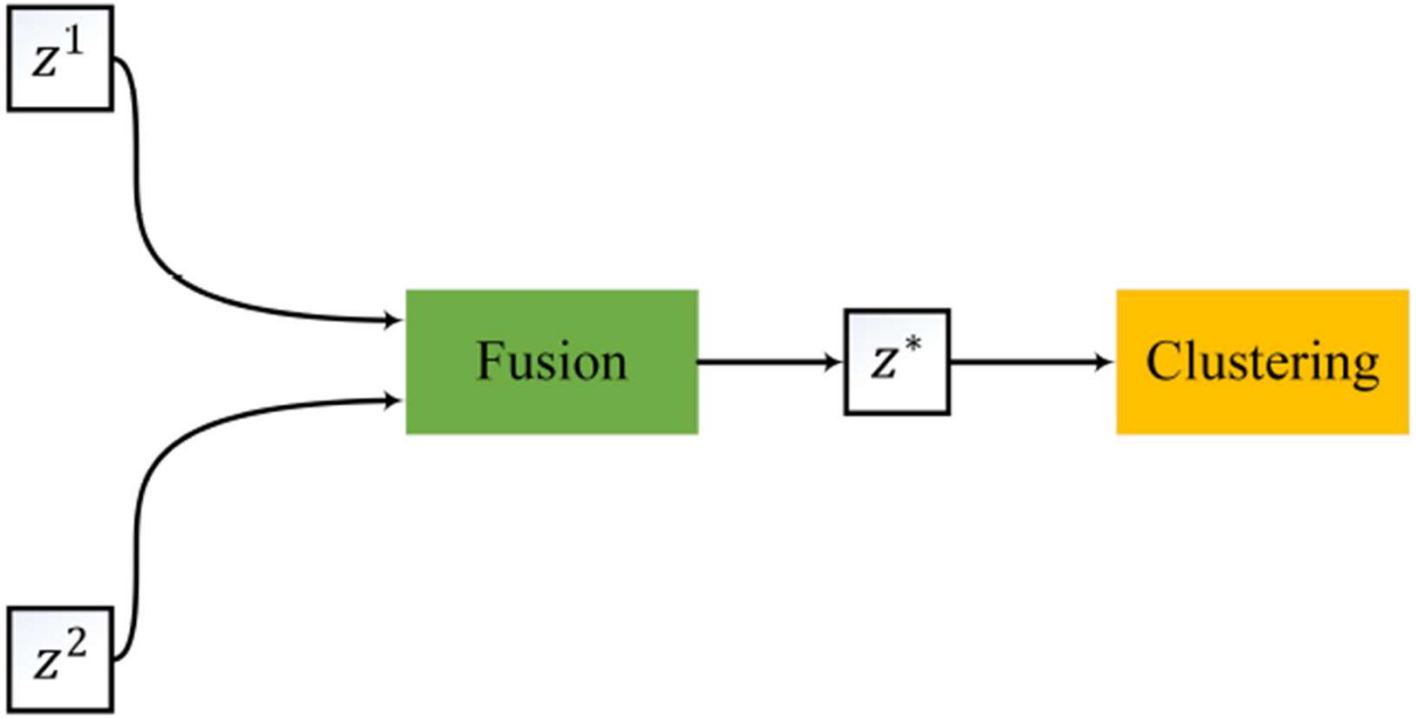
First stage of DMCR algorithm



First Stage

The Proposed Algorithm

Second stage of DMCR algorithm



Second Stage

The Proposed Algorithm

Loss Function of the first stage

$$\min_{\theta^i,\varphi^i,\varphi^j} E_{x^i\sim p(x^i)}[-E_{z^i\sim p(z^i|x^i;\theta^i)}[\log q(\tilde{x}^i|z^i;\varphi^i)] + \beta KL\big(p(z^i|x^i;\theta^i), q(z^i)\big) + \gamma[\sum_{j=1}^{M,j\neq i} -E_{z^i\sim p(z^i|x^i;\theta^i)}[\log q\big(\hat{x}^j|z^i;\varphi^j\big)]]]$$

Loss Function of the senond stage

$$\min_{\eta} \sum_{i=1}^{m} \left\| z^* - z^i \right\|^2$$

Theoretical Analysis

We prove that the cross reconstruction method can minimize the Wasserstein distance of feature distributions among different modalities.

$$W\left(p\left(z^i\middle|x^i;\theta^i\right),p\left(z^j\middle|x^j;\theta^j\right)\right)$$

$$= inf_{\epsilon\in\Pi\left(p\left(z^i|x^i;\theta^i\right),p\left(z^j|x^j;\theta^j\right)\right)}E_{\left(z^i,z^j\right)\sim\epsilon}\left[\left\|z^i-z^j\right\|\right]$$

Experiments

Description of Datasets

	m	d_1	d_2	d_3	n	k
Digits	3	76	216	240	2000	10
CNN	2	35719	996	–	2107	7
AwA	3	2000	2000	2000	5814	10
Caltech101	3	254	512	256	712	10
LUse-21	3	254	512	256	3000	15
Scene-15	3	254	512	256	2100	21

Experiments

Comparing methods

Single modal clustering methods

DEC, IDEC and JULE.

Multimodal clustering methods

MMSC, RMKMC, DIMSC, LT-MSC, ECMSC and DCCAE.

The simplified DMCR

DMC

Experiments

Experiment results

Table 1. Clustering ACC (%)

	DEC	IDEC	JULE	DCCAE	MMSC	RMKMC	LT-MSC	DIMSC	ECMSC	DMC	DMCR
Digits	44.07	46.70	76.36	66.19	54.00	73.30	73.35	52.10	77.40	85.55	**90.75**
CNN	36.96	31.69	40.00	34.27	32.02	34.21	21.26	21.64	25.34	55.52	**66.68**
AwA	16.10	20.08	23.05	24.31	20.28	23.17	24.17	20.79	22.19	26.23	**28.67**
Cal101	35.74	43.54	63.20	62.85	49.02	51.97	61.35	39.47	53.90	62.50	**66.57**
LUse-21	12.34	25.14	27.14	23.72	20.38	27.24	30.86	22.81	25.29	30.14	**31.62**
Scene-15	16.97	25.03	38.10	34.95	18.20	40.03	43.93	23.87	42.70	44.60	**54.00**

Table 3. Clustering Purity (%)

	DEC	IDEC	JULE	DCCAE	MMSC	RMKMC	LT-MSC	DIMSC	ECMSC	DMC	DMCR
Digits	47.97	49.00	77.44	69.92	58.25	76.40	77.35	53.50	78.13	85.55	**90.75**
CNN	40.84	37.66	40.80	35.19	32.21	34.87	25.63	24.54	25.49	56.95	**61.70**
AwA	17.96	21.87	23.58	25.68	20.64	24.70	27.35	23.89	23.24	27.95	**31.84**
Cal101	37.88	47.19	**68.68**	66.29	54.21	58.85	64.58	47.33	61.94	62.50	64.46
LUse-21	14.29	27.52	30.57	27.49	21.81	29.00	32.24	24.43	28.33	34.43	**35.05**
Scene-15	17.65	26.43	38.97	38.69	18.60	41.40	44.70	25.87	44.80	46.57	**56.60**

Conclusion

In this paper, we propose a novel deep multimodal clustering framework called DMCR. Firstly, we control the scale of feature using VIB. Secondly, we reduce the distribution differences among multimodal features using cross reconstruction. Thirdly, we fuse the extracted features to common features. Finally, we cluster the common features. In addition, we prove that the proposed cross reconstruction method effectively reduce the distribution differences of multimodal features. We compare our DMCR algorithm with the state-of-the-art multimodal methods on many multimodal datasets. Experimental results show that our algorithm achieves obviously improvement on multimodal clustering task.

Thanks