Realize Generative yet Complete Latent Representation for Incomplete Multi-view Learning

Xin Gao

2024.1.30

Content

- Introduction (Multi-view Learning)
- Vanilla Multi-view VAE (VMVAE)
- Complete Multi-view VAE (CMVAE)
- Experiments and results

Realize Generative yet Complete Latent Representation for Incomplete Multi-view Learning

Hongmin Cai[1], *Senior Member, IEEE*, Weitian Huang[1], Sirui Yang[1], Siqi Ding[1], Yue Zhang[2], Bin Hu[3], *Fellow, IEEE*, Fa Zhang[3], Yiu-ming Cheung[4,*], *Fellow, IEEE*

南方科技大学
北京理工大学
广州技术师范大学
香港浸会大学

Multi-view Learning

Multi-view data

- News in different languages;
- Describing events with images, audio and text;
- Detecting organs through different imaging mechanisms to obtain multi-modal medical images.

Multi-view learning

- These semantically coherent multi-view samples are connected by a consensus representation.
- Individual views containing insufficient information, while different views can complement each other

Incomplete Multi-view learning

Challenges and Solutions:

- Multi-view data with random missing views
- Missing filling / Grouping-and-learning / Cross-view learning
- Goal: making full use of existing data to predict the hidden features of missing views and integrate them into a unified representation.

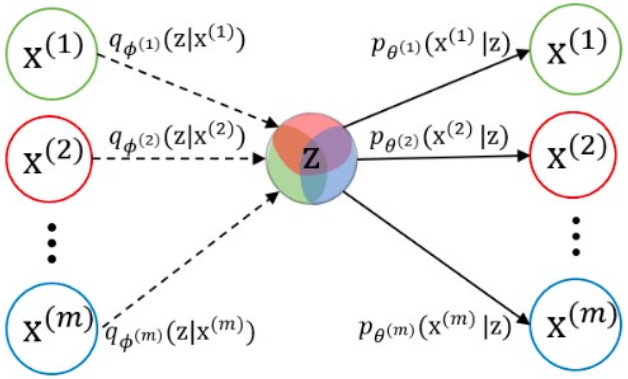
Desiderata:

- Completeness: The learned multi-view representation contains complete information.
- Cross-generative: Multi-view representation learned from available views has the ability to generate missing sample.
- **Solidative**: There exists conservative information inherent to multi-view sampling that is not altered by the absence of views, e.g., the weight of views, the intrinsic correlations between views.

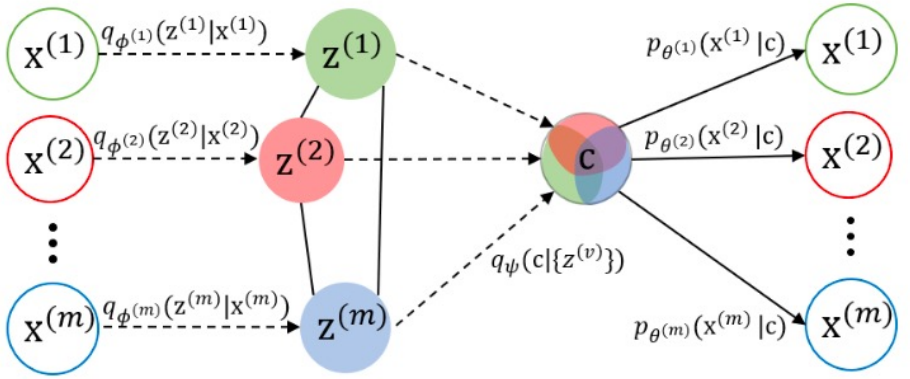
There is something not changed.

Overview

- A deep generative model to learn view-peculiar and complete latent representation.
- Model the generation of the multiple views from a complete latent variable represented by a mixture of Gaussian distributions.
- Integrate view-invariant information into posterior inference to enhance the solidative of the learned latent representation.
- The intrinsic correlations between views are mined to seek cross-view generality.



(a) Vanilla Multi-view VAE



(b) Complete Multi-view VAE

Recap

- **Variational Inference:** This approach uses an approximate posterior distribution in a family to estimate the true posterior distribution by maximizing a variational lower bound.
- **Objective:** Use approximate posterior $q(\theta)$ to estimate the true posterior distribution $P(\theta|X)$

$$\log P(X) = \int q(\theta) \log P(X) d\theta$$

evidence
与θ无关

$$= \int q(\theta) \log \frac{P(X \mid \theta) P(\theta)}{P(\theta \mid X)} \frac{q(\theta)}{q(\theta)} d\theta$$

$$= \int q(\theta) \log P(X \mid \theta) d\theta - KL(q(\theta), P(\theta)) + KL(q(\theta), P(\theta \mid X))$$

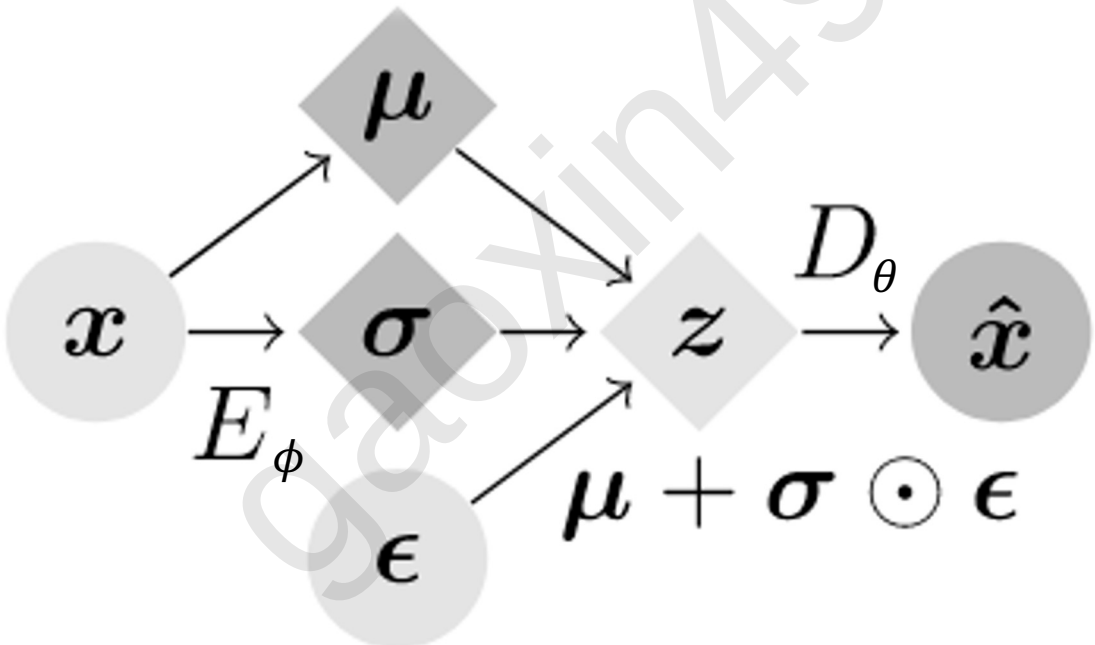
ELBO (Evidence lower bound)　　KL divergence

Our goal: Minimize the KL divergence between the approximate posterior and the true posterior

⇒ **Maximize the ELBO**

Recap

ELBO $\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}\right) = -D_{KL}\left(q_{\boldsymbol{\phi}}\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right) \| p(\mathbf{z})\right) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)} \mid \mathbf{z}\right)\right]$



Vanilla Multi-view VAE

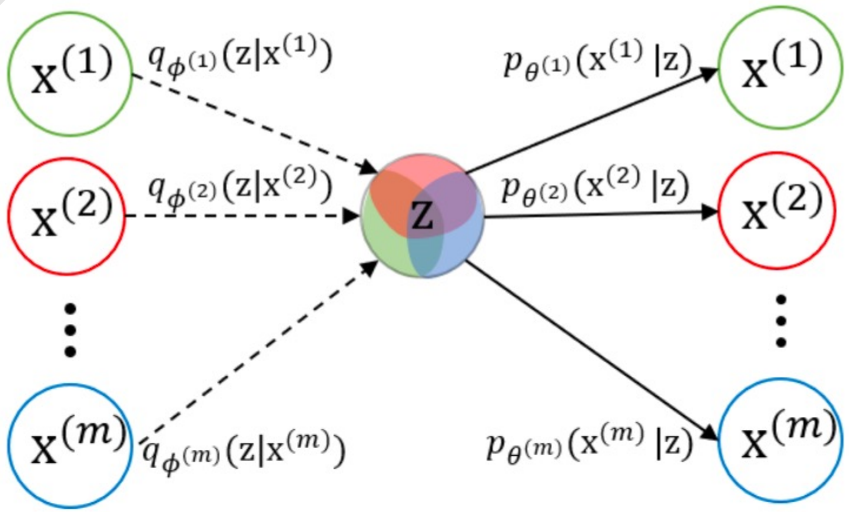
omit the subscript i and denote $\{*^{(v)}\}_{v=1}^{m}$ as $\{*^{(v)}\}$

a m views dataset $\mathcal{X} = \{\mathbf{X}^{(v)} \in \mathbb{R}^{n\times d_v}\}_{v=1}^{m}$, $\{\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, ..., \mathbf{x}_n^{(v)}\}$.

Under the assumption that the density $p(\{x^{(v)}\})$ is achieved through the marginalization of a shared latent:

$$p(\{\mathbf{x}^{(v)}\}) = \int \prod_{v=1}^{m} p_{\theta^{(v)}}(\mathbf{x}^{(v)}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \quad (2)$$

$$\begin{aligned}\log p(\{\mathbf{x}^{(v)}\}) \geq & \mathcal{L}_{\text{ELBO}}(\{\mathbf{x}^{(v)}\}) \\ = & - D_{\text{KL}}\left(q(\mathbf{z}|\{\mathbf{x}^{(v)}\})\|p(\mathbf{z})\right) \\ & + \mathbb{E}_{q(\mathbf{z}|\{\mathbf{x}^{(v)}\})}\left[\log p(\{\mathbf{x}^{(v)}\}|\mathbf{z})\right]. \quad (3)\end{aligned}$$



(a) Vanilla Multi-view VAE

It is crucial to choose a highly expressive and easily computable density as the joint variational posterior.

Vanilla Multi-view VAE

Introducing a mixture of Gaussian distributions as the joint variational posterior

$$q(\mathbf{z}|\{\mathbf{x}^{(v)}\}) = \sum_{v=1}^{m} \lambda_v q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)}) = \sum_{v=1}^{m} \lambda_v \mathcal{N}(\mathbf{z};\ \boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}),\ \boldsymbol{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)}))$$

$$\mathcal{L}_{\text{ELBO}} = -D_{\text{KL}}\left(\sum_{v=1}^{m} \lambda_v q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)}) \| p(\mathbf{z})\right) + \sum_{v=1}^{m} \lambda_v \mathbb{E}_{q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)})}\left[\sum_{j=1}^{m} \log p_{\theta^{(v)}}(\mathbf{x}^{(j)}|\mathbf{z})\right]$$

- The summation of KL-divergences drives the unimodal variational posteriors to approach the prior individually
- the summation of their expectations reveals alternatively variational inference followed by full view generation.

$$\mathcal{L}_{\text{VMVAE}} = -\sum_{v=1}^{m} \lambda_v D_{\text{KL}}\left(q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)}) \| p(\mathbf{z})\right) + \sum_{v=1}^{m} \lambda_v \mathbb{E}_{q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)})}\left[\sum_{j=1}^{m} \log p_{\theta^{(v)}}(\mathbf{x}^{(j)}|\mathbf{z})\right] \leq \mathcal{L}_{\text{ELBO}}.$$

Hint: $f_i(x)$, $\int f_i(x)\,dx = 1$, $g(x) = \sum \lambda_i f_i(x)$

Continuous Jensen Inequality: $\int f_i(x) \log \frac{g(x)}{f_i(x)}\,dx \leq \log \int f_i(x) \frac{g(x)}{f_i(x)}\,dx = 0$

Complete Multi-view VAE

- Exploit the intrinsic correlations between views => Missing views
- Let $(\{z^{(v)}\}, c)$ denote the view-peculiar and complete generative latent variables
- Modelling $\{z^{(v)}\}$ by a linear transformation $z^{(w)} = z^{(v)} C_{vw}$

For a random variable obeying the Gaussian distribution, given $y \sim N(\mu, \Sigma)$, whose linear transformation distribution is $yC \sim N(\mu C, C^T \Sigma C)$ under the statistical principle.

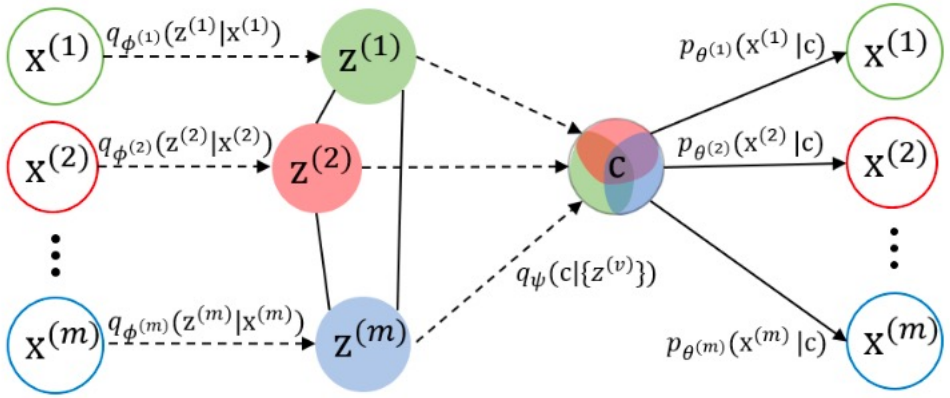
$$q(\{\mathbf{z}^{(v)}\}, \mathbf{c} | \{\mathbf{x}^{(v)}\})$$

$$= \sum_{v=1}^{m} \lambda_v q(\{\mathbf{z}^{(v)}\}, \mathbf{c} | \mathbf{x}^{(v)})$$

$$= \sum_{v=1}^{m} \lambda_v q_\psi(\mathbf{c} | \{\mathbf{z}^{(v)}\}) \prod_{w \neq v}^{m} q(\mathbf{z}^{(w)} | \mathbf{z}^{(v)}) q_{\phi^{(v)}}(\mathbf{z}^{(v)} | \mathbf{x}^{(v)})$$

$$q_{\phi^{(v)}}\left(\mathbf{z}^{(v)} \mid \mathbf{x}^{(v)}\right) \sim \mathcal{N}(\mathbf{z}^{(v)}; \mu_{\phi^{(v)}}(\mathbf{x}^{(v)}), \mathbf{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)}))$$

$$q\left(\mathbf{z}^{(w)} \mid \mathbf{z}^{(v)}\right) \sim \mathcal{N}(\mathbf{z}^{(w)}; \mu_{\phi^{(v)}}(\mathbf{x}^{(v)}) C_{vw}, C_{vw}^T \mathbf{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)}) C_{vw})$$



(b) Complete Multi-view VAE

Complete Multi-view VAE

$$q\left(\{z^{(v)}\}, c \mid \{x^{(v)}\}\right) = \sum_{v=1}^{m} \lambda_v q(c \mid \{z^{(v)}\}) q(\{z^{(v)}\} \mid x^{(v)})$$

$$\log p(\{x^{(v)}\}) = \int q(\{z^{(v)}\}, c \mid \{x^{(v)}\}) \log \frac{p(\{x^{(v)}\} \mid c, \{z^{(v)}\}) p(c) p(\{x^{(v)}\})}{p(\{z^{(v)}\}, c \mid \{x^{(v)}\})} \frac{q(\{z^{(v)}\}, c \mid \{x^{(v)}\})}{q(\{z^{(v)}\}, c \mid \{x^{(v)}\})} d\{z^{(v)}\} dc$$

$$= KL(q(\{z^{(v)}\}, c \mid \{x^{(v)}\}) \| p(\{z^{(v)}\}, c \mid \{x^{(v)}\}))$$

$$- KL(q(c \mid \{z^{(v)}\}) \| p(c)) - KL(\sum_{v=1}^{m} \lambda_v q(\{z^{(v)}\} \mid x^{(v)}) \| p(\{z^{(v)}\})) + \sum_{v=1}^{m} \lambda_v \mathbb{E}_{q(\{z^{(v)}\}, c \mid x^{(v)})} (\sum_{j=1}^{m} \log p(x^{(j)} \mid c))$$

$$\geq -KL(q(c \mid \{z^{(v)}\}) \| p(c)) - \sum_{v=1}^{m} \lambda_v KL(q(\{z^{(v)}\} \mid x^{(v)}) \| p(\{z^{(v)}\})) + \sum_{v=1}^{m} \lambda_v \mathbb{E}_{q(\{z^{(v)}\}, c \mid x^{(v)})} (\sum_{j=1}^{m} \log p(x^{(j)} \mid c))$$

$$= -KL(q(c \mid \{z^{(v)}\}) \| p(c)) - \sum_{v=1}^{m} \lambda_v KL\left[\prod_{w \neq v}^{m} q(z^{(w)} \mid z^{(v)}) q(z^{(v)} \mid x^{(v)}) \| \prod_{w \neq v}^{m} p(z^{(w)} \mid z^{(v)}) p(z^{(v)})\right] + \dots$$

$$= -KL(q(c \mid \{z^{(v)}\}) \| p(c)) - \sum_{v=1}^{m} \lambda_v \sum_{w \neq v}^{m} KL\left[\mathcal{N}(z^{(w)}; \mu^{(w)}, \Sigma^{(w)}) \| \mathcal{N}(z^{(w)}; \mu^{(v)} C_{vw}, C_{vw}^T \Sigma^{(v)} C_{vw})\right]$$

$$+ \sum_{v=1}^{m} \lambda_v \mathbb{E}_{q(c \mid \{z^{(v)}\})} (\sum_{j=1}^{m} \log p(x^{(j)} \mid c))$$

$$p\left(z^{(v)}\right) = 1$$

Complete Multi-view VAE

$$-KL(q(c \mid \{z^{(v)}\}) \| p(c)) - \sum_{v=1}^{m} \lambda_v \sum_{w \neq v}^{m} KL\left[\mathcal{N}(z^{(w)}; \mu^{(w)}, \Sigma^{(w)}) \| \mathcal{N}(z^{(w)}; \mu^{(v)} C_{vw}, C_{vw}^{T} \Sigma^{(v)} C_{vw})\right] \quad (12)$$

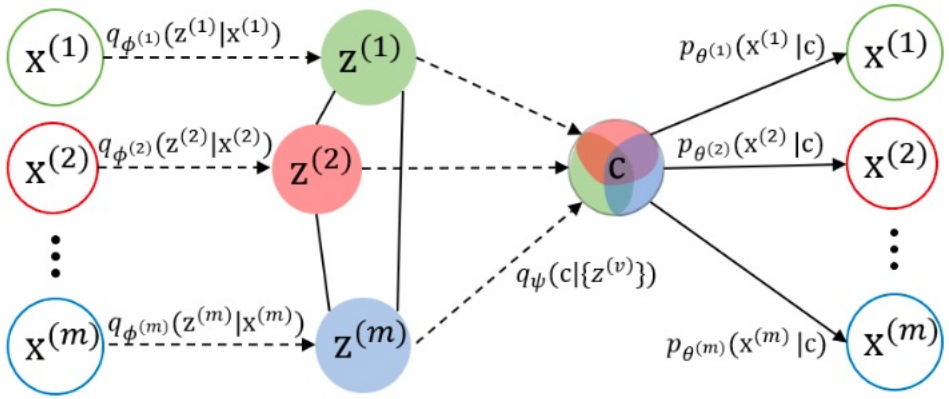
$$+ \sum_{v=1}^{m} \lambda_v \mathbb{E}_{q(c|\{z^{(v)}\})} \left(\sum_{j=1}^{m} \log p(x^{(j)} \mid c) \right)$$

① Prior on the complete latent variable

② Prior on the correlation between different view-peculiar latent variables

Note that the second term of Eq. (12) calculates only the available paired views.

③ Reconstruction loss



(b) Complete Multi-view VAE

Complete Multi-view VAE

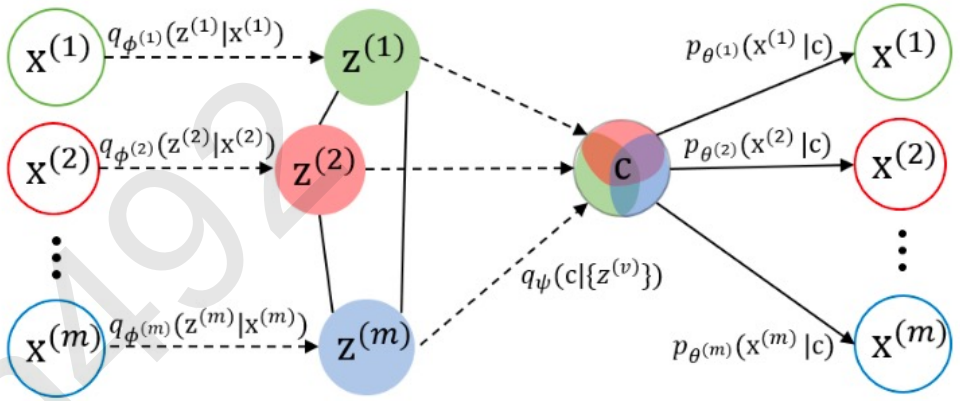
Algorithm 1 Optimization Procedure of CMVAE

Input: Multi-view dataset $\mathcal{X}$; Statistical model of the prior $p(\mathbf{c}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$; Setting $T = 1$ and the dimensionality of latent variables.

Parameter: Initialize parameters $\{\phi^{(v)}\}$, $\{\theta^{(v)}\}$, ψ with random values, $\lambda_v = \frac{1}{m}$ and C_{vw} with identity matrix.

1: **while** not reaching the maximal epochs **do**
2: **for** v in m views **do**
3: Calculate $(\boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}), \boldsymbol{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)}))$ through v-th encoder and then sample $\mathbf{z}_t^{(v)}$ by Eq. (10);
 Implement $\mathbf{z}^{(w)} = \mathbf{z}^{(v)} C_{vw}, \forall w \in 1, 2, ..., m, \ w \neq v$;
 Calculate $(\boldsymbol{\mu}_{\psi}(\{\mathbf{z}^{(v)}\}), \boldsymbol{\Sigma}_{\psi}(\{\mathbf{z}^{(v)}\}))$ through the fusion network and then sample $\mathbf{c}_t$ by Eq. (11);
4: **for** j in m views **do**
5: Generate $\{\mathbf{x}^{(v)}\}$ by m decoders.
6: **end for**
7: **end for**
8: Update $\{\phi^{(v)}\}$, $\{\theta^{(v)}\}$, ψ, C_{vw}, λ_v by maximizing Eq. (12).
9: **end while**

Output: The complete generative latent representation $\mathbf{c}$.



(b) Complete Multi-view VAE

$$\mathbf{z}_t^{(v)} = \boldsymbol{\mu}_{\phi^{(v)}} + \boldsymbol{R}_{\phi^{(v)}} \boldsymbol{\epsilon}_t^{(v)} \quad (10)$$

$$\mathbf{c}_t = \boldsymbol{\mu}_{\psi} + \boldsymbol{R}_{\psi} \boldsymbol{\epsilon}_t \quad (11)$$

- The intrinsic correlations between views are explicitly modeled, ensuring that they are not affected by the absence of certain views.
- As a result, the weight of the Gaussian Mixture Model (GMM) remains preserved.

Experiments

- Clustering, Classification, Cross-view generation, Bioinformatic Data
- **Model setup**: $q_{\phi^{(v)}}\left(\mathbf{z}^{(v)} \mid \mathbf{x}^{(v)}\right)$, $p_{\theta^{(v)}}\left(\mathbf{x}^{(v)} \mid \mathbf{c}\right)$

  The fusion network $q_{(\psi)}\left(\mathbf{c} \mid \left\{\mathbf{z}^{(v)}\right\}\right)$ concatenates multiple view-peculiar latent representations, followed by a fully connected layer with dimensionality D.

- **Multi-view in different datasets**:

  Different image features (HOG, LBP, Gabor features…)
  Different deep representation
  Different writing styles…

Datasets	# Samples	# Views	# Classes	Dimensionality
MSRC-V1	240	5	7	24,576,512,256,254
Notting-Hill	550	3	5	2000,3304,6750
Handwritten	2000	5	10	240,76,216,47,64
Caltech101-20	2386	6	20	48,40,254,1984,512,928
BDGP	2500	2	5	1750,79
Animal	10158	2	50	4096,4096
PolyMNIST	60000	5	10	784,784,784,784,784
Multiome PBMC	11909	2	11	36601,108377
Multiome BMNC	69249	2	22	13431,116490

- **Incomplete data construction**:

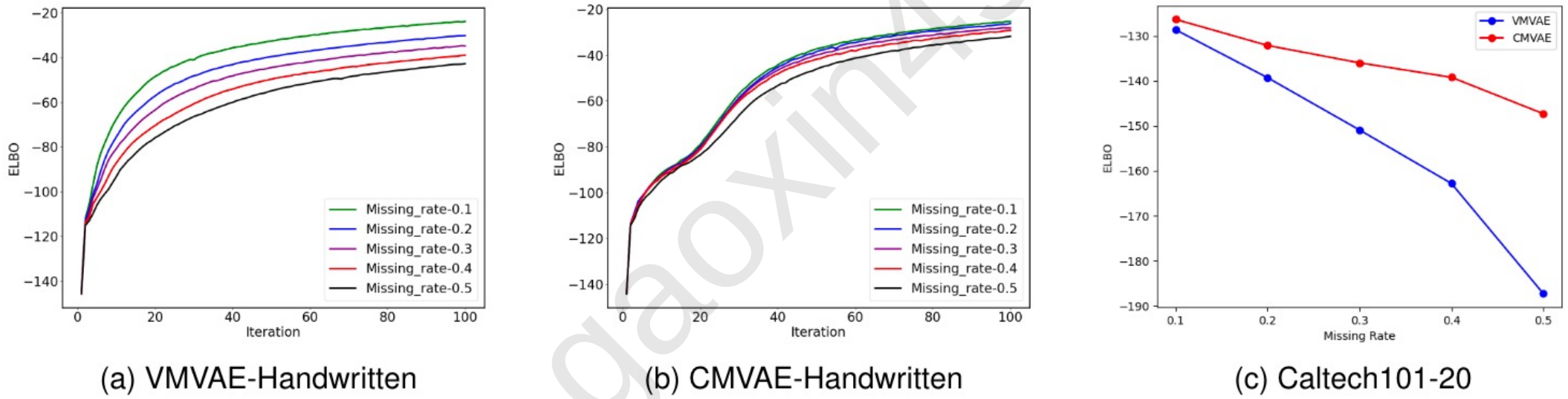
  All samples were guaranteed to retain at least one view
  Missing rate η = {0, 0.1, 0.2, 0.3, 0.4, 0.5}, randomly selected η×n×m samples as missing data, then random instances were removed from each view

Joint Likelihood Approximation

$$\log p\left(\left\{\mathbf{x}^{(v)}\right\}\right) \geq \mathcal{L}_{\text{VMVAE}}$$

$$\log p\left(\left\{\mathbf{x}^{(v)}\right\}\right) \geq \mathcal{L}_{\text{CMVAE}}$$

The value of the variational lower bounds affects the portrayal of the data distribution, as well as the accuracy of the inference of the posterior.

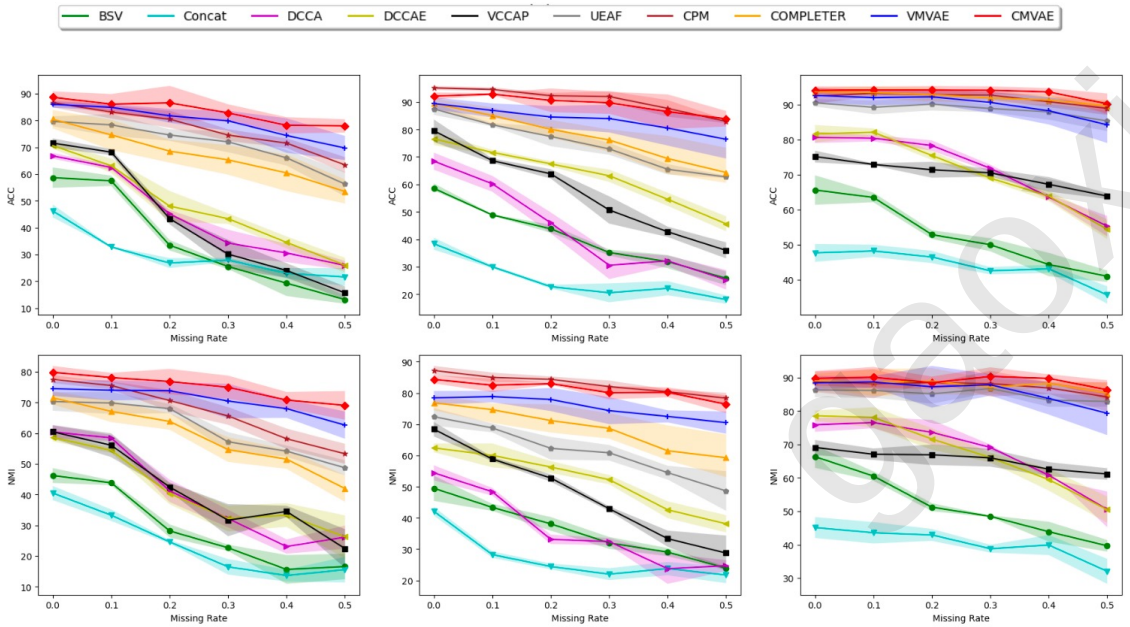


(a) VMVAE-Handwritten (b) CMVAE-Handwritten (c) Caltech101-20

Fig. 2. The variation of the objective values in terms of training iteration for (a) VMVAE, and (b) CMVAE, on the Handwritten dataset. The convergence values reached by ELBO decrease as the missing rate increases, while CMVAE ultimately achieves a higher ELBO value. (c) The pronounced difference in ELBO values on the Caltech101-20 dataset verifies that CMVAE has a tighter lower bound, especially with large amounts of information missing.

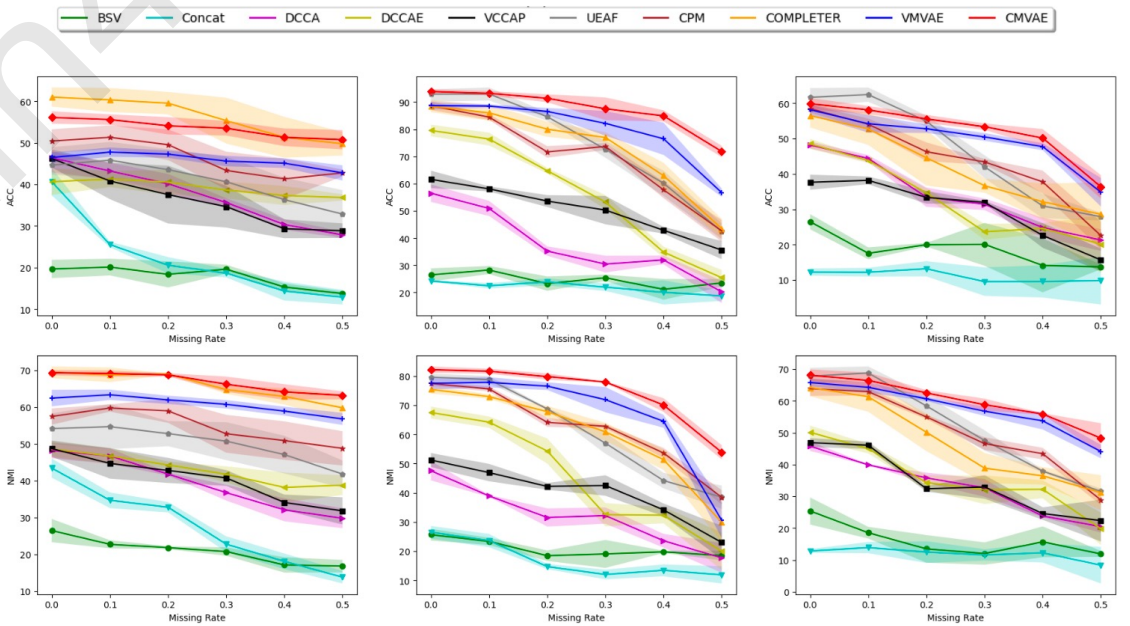
Results: Clustering

- Impute Missing Data with Mean (methods which cannot handle missing views directly)
- Evaluate Two-View Combinations (methods which can only handle two view data)
- Conducting k-means directly on the latent representation z and c, respectively



(a) MSRC-V1 (b) Notting-Hill (c) Handwritten

Fig. 3. Clustering performance comparison in terms of NMI and Accuracy by tested ten methods under different missing rates, on (a) MSRC-V1, (b) Notting-Hill, and (c) Handwritten.



(a) Caltech101-20 (b) BDGP (c) Animal

Fig. 4. Clustering performance comparison in terms of NMI and Accuracy by tested ten methods under different missing rates, on (a) Caltech101-20, (b) BDGP, and (c) Animal.

Results: Classification

- The multi-view unified latent representations z and c are respectively fed into fully connected layers with the softmax activator.
- Network parameters are jointly optimized by adding cross-entropy loss.

TABLE 3
Classification accuracy comparison under different missing rate on three datasets (mean±standard deviation). Higher values indicate better performance. The optimal and suboptimal results are in bold and underlined, respectively.

Datasets	Methods	0	0.1	0.2	0.3	0.4	0.5
Caltech101-20	BSV	52.45±2.21	50.95±1.62	53.30±1.27	45.70±1.15	39.28±2.41	30.01±2.55
	Concat	67.91±2.08	65.54±1.06	55.43±1.04	50.31±1.41	45.34±0.60	32.79±0.32
	DCCA [2]	56.78±1.92	52.00±0.54	52.16±1.55	46.97±1.64	41.67±1.86	36.78±1.43
	DCCAE [22]	56.60±1.35	57.20±1.46	56.82±0.62	55.47±0.89	49.36±1.14	50.12±1.18
	VCCAP [3]	58.17±3.18	50.56±0.54	47.80±0.58	44.20±0.57	46.77±1.21	44.49±1.07
	UEAF [52]	76.15± 0.95	74.67±1.07	72.63±1.05	71.86±1.19	68.96±1.62	66.83±2.20
	CPM [17]	90.84±0.52	91.10±1.28	90.85±0.98	89.40±1.24	87.23±1.18	84.39±2.38
	COMPLETER [33]	91.48±0.84	89.65±0.55	88.68±0.84	86.15±1.74	85.14±1.47	84.80±1.65
	VMVAE	**92.58±0.78**	90.10±1.60	89.65±0.65	87.88±1.85	86.68±2.07	84.51±3.32
	CMVAE	92.48±0.65	**92.21±0.45**	**91.45±0.64**	**90.45±0.55**	**89.55±0.90**	**87.65±1.58**

Six datasets

Mining the correlation between views and making full use of view invariant information is helpful for learning complete latent representations in the absence of views.

Results: Cross-view Generation

- Five digits [0, 4, 6, 7, 9]: ① digit 0 containing only view 1. ② digit 4 containing only view 2. ③ digit 6 containing only view 3. ④ digit 9 containing only view 4 and ⑤ digit 7 containing views 2, 3, 4, 5.

(a) Samples with missing view (b) MoPoE-VAE (c) VMVAE (d) CMVAE

Fig. 5. Visualization on cross-view image generation. (a) For each sample of 0, 4, 6, and 9, there are only one view are observed, while the others are missing. For the sample 7, there are only one missing view. The observed samples are used for generating the remaining view images by (b) MoPoE-VAE, (c) VMVAE, and (d) CMVAE. As can be seen, CMVAE shows the best detail in terms of figures structure and background, which is clearly contrasted in the first view, highlighted by the green box.

Application to Bioinformatic Data

- Bioinformatic multi-omics data: PBMC, BMMC datasets
- Multiomics single-cell genomic data

(1) Multiome PBMC. Human peripheral blood mononuclear cell (PBMC) profiles generated by the 10× Genomics Multiome ATAC and RNA kit with 11,909 cells.

(2) Multiome BMMC. Single-cell multi-omics data collected from bone marrow mononuclear cells (BMMC) from 12 healthy human donors. Half of the samples were measured using paired RNA and ATAC kits, and half were measured using single-cell gene expression kits only, for a total of 69,249 cells.

TABLE 5
Performance comparison of cell typing. Larger values indicate better performance. The optimal and suboptimal results are in bold and underlined, respectively.

Datasets	Methods	NMI	ARI	ASW
Multiome PBMC	MOFA+ [54]	79.12	71.58	62.14
	Seurat-v4 [55]	<u>81.68</u>	<u>74.34</u>	<u>60.59</u>
	MultiVI [56]	77.68	65.37	59.48
	CMVAE	**81.85**	**75.53**	**62.80**
Multiome BMMC	MOFA+ [54]	60.63	25.86	53.40
	Seruat-v4 [55]	73.67	<u>61.19</u>	58.98
	MultiVI [56]	<u>75.10</u>	60.27	<u>59.28</u>
	CMVAE	**78.56**	**68.17**	**59.89**

Application to Bioinformatic Data

CMVAE clearly divides the NK cell population into two clusters, which implies that NK can be classified into two subtypes, as in [1]

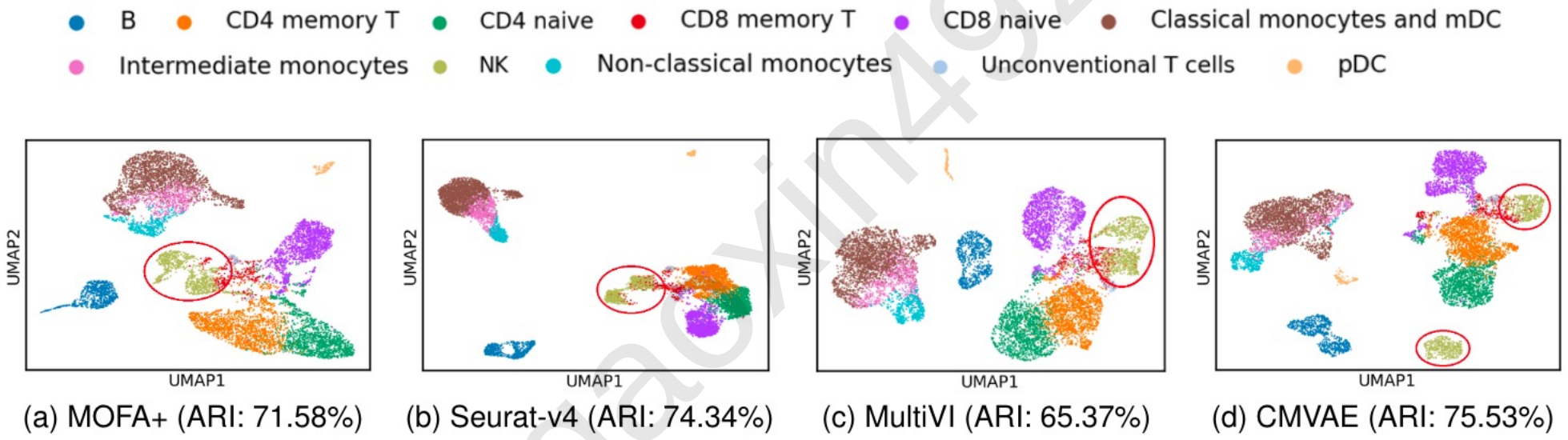


Fig. 7. Visualization results of multi-view latent representations using UMAP on Multiome PBMC dataset. Different colors represent different cell types. Through CMVAE, NK cells are more distinctly divided into two clusters in the embedding space, and the best cell typing performance for the ARI indicator is obtained.

[1] S. Ghazanfar, C. Guibentif and J.C. Marioni. Stabilized mosaic single-cell data integration using unshared features. Nature Biotechnology, 1-9, 2023.

Conclusion

Advantages:

1. Introduction of view-invariant information mining enables compensation for missing view information.
2. The variational inference process incorporates the exploration of intrinsic transformations between views for interconversion, ensuring that view weights remain invariant to prevent misrepresentation of the latent variable.
3. Practical significance demonstrated through application to real bioinformatics data.

Disadvantage:

1. Limited proof demonstration; modifications have been made in the PowerPoint presentation.