Supplementary Materials for
Distribution-Consistent Modal Recovering for Incomplete Multimodal Learning

Yuanzhi Wang, Zhen Cui*, Yong Li*
PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.
{yuanzhiwang, zhen.cui, yong.li}@njust.edu.cn

1. Overview

In this supplementary material, we present the details of the normalizing flows in Sec. 2. In Sec. 3, we provide the detailed neural network architecture as well as the hyperparameter settings in DiCMoR.

2. Details of the normalizing flows

Let us take the normalizing flow $\mathcal{F}^{(m)}$ of modality m as an example, $\mathcal{F}^{(m)}$ consists of N invertible layers: $\mathcal{F}^{(m)} = f_1 \circ f_2 \circ \cdots \circ f_N$, and accordingly, $(\mathcal{F}^{(m)})^{-1} = (f_N)^{-1} \circ (f_{N-1})^{-1} \circ \cdots \circ (f_1)^{-1}$. We adopt the classical affine coupling layer [1] as the basic invertible layer. The architecture of the basic invertible layer is shown in Fig. 1, we assume that the input of the affine coupling layer is x that is split into two parts along the channel axis, denoted as x_a and x_b. The forward process can be formulated as:

$$y_a = x_a,\ y_b = s \times x_b + t, \tag{1}$$

where s and t is obtained by a complex nonlinear neural network NN$(\cdot)$, as shown in Fig. 1. y_a and y_b are concatenated as the final output y of the basic invertible layer. For the reverse process, since $y_a = x_a$, s and t can be obtained by feeding y_a into NN$(\cdot)$. The reverse process can be formulated as:

$$x_a = y_a,\ x_b = (y_b - t)/s. \tag{2}$$

By stacking several affine coupling layers, we construct a modality-specific normalizing flow for each modality.

3. Settings in DiCMoR

Tab. 1 illustrates the network architecture and the hyperparameter settings in DiCMoR. We explain the involved neural network components in DiCMoR as follows.

* The corresponding authors.

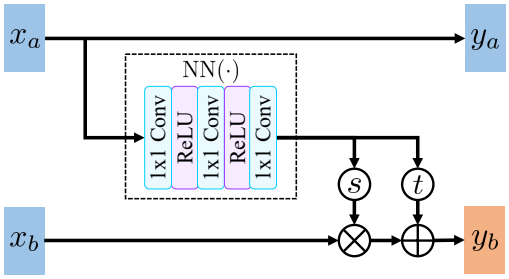

Figure 1: The architecture of affine coupling layer.

Table 1: Hyperparameter settings in DiCMoR.

Hyperparameter	CMU-MOSI	CMU-MOSEI
Shallow feature extractor		
Kernel size for $\mathcal{C}^{(m)}$	5	5
Hidden dimension for $\mathcal{C}^{(m)}$	32	64
Cross-modal distribution transfer		
Hidden dimension for $\mathcal{F}^{(m)}$	64	128
Number of invertible layers for $\mathcal{F}^{(m)}$	32	32
Hidden dimension for $\mathcal{D}^{(m)}$	64	128
Number of RCAB for $\mathcal{D}^{(m)}$	20	20
Multimodal fusion and prediction		
Hidden dimension for $\mathcal{T}^{(m)}$	32	64
Number of attention heads for $\mathcal{T}^{(m)}$	8	8
Layers of transformer for $\mathcal{T}^{(m)}$	4	6

DiCMoR mainly consists of three parts: **shallow feature extractor**, **cross-modal distribution transfer**, and **multimodal fusion and prediction**. First, shallow feature extractor encodes multimodal shallow features $\mathbf{X}^{(m)}$ via separate 1D temporal convolutions $\mathcal{C}^{(m)}$, where $m \in \{L, V, A\}$. Second, we build a cross-modal distribution transfer to learn the distribution space of each modality via normalizing flows $\mathcal{F}^{(m)}$, and conduct cross-modal distribution transformation to estimate the distribution of missing modality. Then, we use the reconstruction modules $\mathcal{D}^{(m)}$ to recover

the final missed data. Each $\mathcal{D}^{(m)}$ is composed of several residual channel attention blocks (RCAB) [3], where the 2D convolutional layers are replaced with 1D temporal convolutional layers to fit the temporal features. Finally, the recovered modalities together with available modalities could be jointly fed into multimodal transformers $\mathcal{T}^{(m)}$ [2] for multimodal fusion and prediction.

References

[1] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 1

[2] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 2

[3] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2