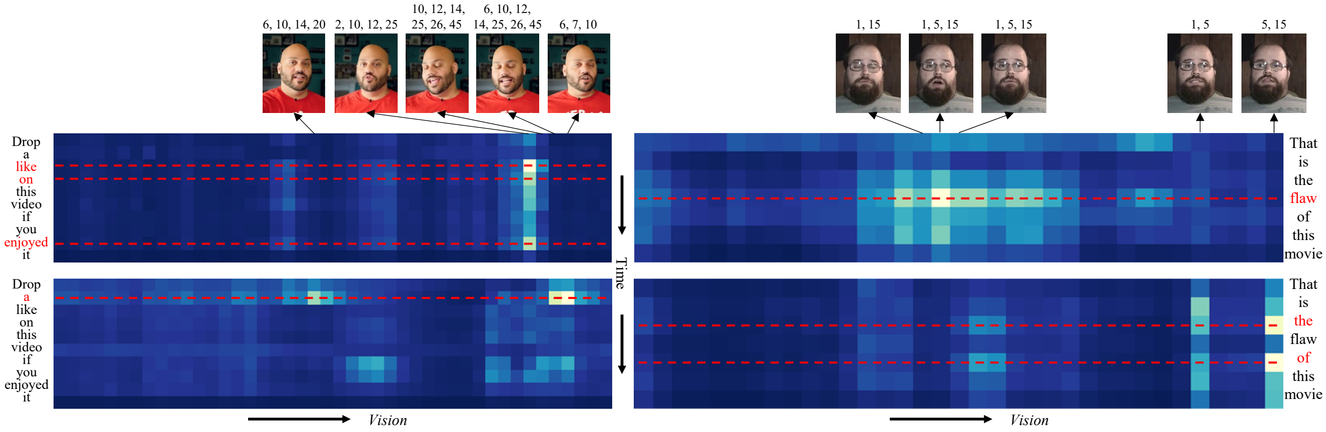
Supplementary Materials for
Decoupled Multimodal Distilling for Emotion Recognition

Yong Li, Yuanzhi Wang, Zhen Cui[*]
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.
{yong.li, yuanzhiwang, zhen.cui}@njust.edu.cn



(a) Top: **DMD** (**Ours**). Bottom: MulT

(b) Top: **DMD** (**Ours**). Bottom: MulT

Figure 1. Visualization of the cross-modal attention matrix activation for the proposed DMD and the classical MulT [1] on the CMU-MOSEI dataset. The language words that are closely related to emotion are highlighted in red. The numbers above the video frames denote the corresponding activated facial action units. Compared with MulT, DMD perceives reliable correlations between elements of different modalities. (a) and (b) illustrate a multimodal sample labelled with *positive* and *negative*, respectively.

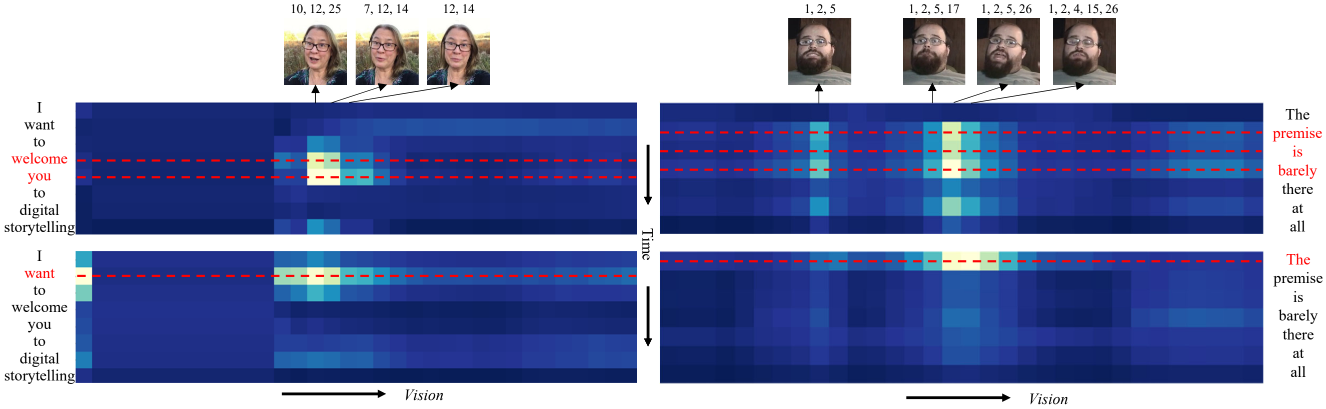
1. Overview

In this supplementary material, we present more experimental results and analysis in Sec. 2 to further prove the effectiveness of the proposed Decoupled Multimodal Distillation (DMD). In Sec. 3, we conduct experiments to prove the consistencies of the graph edges and MER. In Sec. 4, we provide the detailed neural network architecture as well as the hyperparameter settings in DMD. In Sec. 5, we conduct ethics discussion.

2. Visualization

Fig. 1 and Fig. 2 show the visualizations of the cross-modal attention matrix in our proposed DMD and MulT [1] on the CMU-MOSEI dataset. Obviously, our proposed DMD is capable of perceiving more reliable correlations between the modalities. As shown in Fig. 1 (a), DMD successfully correlates the positive words, e.g., "like" and "enjoyed", with the faces that show the corresponding action units, e.g., AU6 (Cheek Raiser), AU10 (Upper Lip Raiser), AU12 (Lip Corner Puller), AU25 (Lips part). In Fig. 1 (b), DMD is capable of building reliable crossmodal correlations where faces in the visual modality show few activated AUs. The visualizations in Fig. 1 and Fig. 2 are reasonable because the features in DMD are refined and decoupled. The visualization results also suggest building modality adaptation with the decoupled heterogeneous features is beneficial as they have significantly reduced the information redundancy. It is worth noting that the comparisons in Fig. 1 and Fig. 2 are consistent with Tab. 4 in the main paper, where DMD outperforms MulT (w/ GD) under various evaluation metrics.

[*] The corresponding author.



(a) Top: **DMD (Ours)**. Bottom: MulT

(b) Top: **DMD (Ours)**. Bottom: MulT

Figure 2. Visualization comparisons between our proposed DMD and MulT [1]. DMD builds superior crossmodal correlations. In (a), the positive word "welcome" is correlated with the faces that show AU10 (Upper Lip Raiser), AU12 (Lip Corner Puller), AU14 (Dimpler). In (b), the negative word "barely" is correlated with the faces that show AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU5 (Upper Lid Raiser). (a) and (b) illustrate a multimodal sample labelled with *positive* and *negative*, respectively.

3. The consistencies of the graph edges and MER

To explore the consistencies of the graph edges and MER, we repeat five times experiments with random network initialization and data shuffling on different seeds. The mean and STD of graph edges on MOSEI dataset are shown in Tab. 1. For ACC$_2$ and F1 of MER, their mean/STD are 84.7%/0.123 and 84.65%/0.150. Thus, the consistencies can be guaranteed.

Table 1. Mean and STD of **graph edges** for the five runs.

Metrics	$V \to L$	$A \to L$	$L \to V$	$A \to V$	$L \to A$	$V \to A$
	homo/hete	homo/hete	homo/hete	homo/hete	homo/hete	homo/hete
Mean	0.063/0.076	0.063/0.023	0.330/0.276	0.107/0.081	0.330/0.276	0.107/0.268
STD	0.007/0.005	0.007/0.003	0.014/0.006	0.007/0.004	0.014/0.005	0.007/0.004

4. Settings in DMD

Tab. 2 illustrates the network architecture and the hyperparameter settings in DMD. Code as well as the pre-trained models will be publicly available. We explain the involved neural network components in DMD as follows.

For DMD, given the input multimodal data, DMD encodes their respective shallow features $\widetilde{\mathbf{X}}_m$ via the separate 1D temporal convolutions $\mathcal{C}_m$, where $m \in \{L, V, A\}$. For *feature decoupling*, DMD exploits the decoupled homo-/heterogeneous multimodal features $\mathbf{X}_m^{\text{com}}$ / $\mathbf{X}_m^{\text{prt}}$ via the shared encoder ($\mathcal{E}^{com}$) and exclusive encoders ($\mathcal{E}_m^{prt}$), respectively. Subsequently, $\mathbf{X}_m^{\text{com}}$ will be fed into a GD-Unit for adaptive knowledge distillation in *HomoGD*. For *HeteroGD*, $\mathbf{X}_m^{\text{prt}}$ are reinforced to $\mathbf{Z}_{\to m}^{\text{prt}}$ via multimodal transformers to bridge the distribution gap. The GD-Unit in *HeteroGD* takes $\mathbf{Z}_{\to m}^{\text{prt}}$ as input for multimodal distillation. The GD-Units in *HomoGD* and *HeteroGD* share the same architecture where the hidden dimension was set as 32.

Table 2. Hyperparameter settings in DMD.

Hyperparameter	CMU-MOSI	CMU-MOSEI
Feature decoupling		
Kernel size for $\mathcal{C}_L, \mathcal{C}_V, \mathcal{C}_A$	5,5,5	5,3,3
Kernel size in $\mathcal{E}^{\text{com}}$	1	1
Kernel size for $\mathcal{E}_L^{\text{prt}}, \mathcal{E}_V^{\text{prt}}, \mathcal{E}_A^{\text{prt}}$	1,1,1	1,1,1
Kernel size for $\mathcal{D}_L, \mathcal{D}_V, \mathcal{D}_A$	1,1,1	1,1,1
Margin: α	0.1	0.1
HomoGD & HeteroGD		
Hidden dimension for a CA unit	50	30
Number of attention heads	10	6
Layers of transformer	4	4

5. Ethics discussion

Recognizing human emotion from multiple modalities helps build better human-computer interactions. The MER technology facilitates machines to address a wider range of human needs. Albeit the obvious benefits, this technology introduces potential risks, including the manipulation of privacy and the potential racial bias, and potentially unexpected economic impact due to rapid technological developments.

References

[1] 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*, 2019. 1, 2