The supplementary of "Attention is not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion"

Tao Liang[1,2] Guosheng Lin[3] Lei Feng[3] Yan Zhang[4] Fengmao Lv[1,5]
[1] Southwest Jiaotong University
[2] Engineering Productivity & Quality Assurance of IES, Bytedance
[3] Nanyang Technological University
[4] University of Electronic Science and Technology of China
[5] Center of Statistical Research, Southwestern University of Finance and Economics

Table A1. Performance (%) by controlling the number of epochs and transformer layers. The superscripts † and ‡ indicate using the parameter settings of MulT and MICA, respectively.

Model	Benchmark	Epoch	Layer	Acc_7	Acc_2	F1
$\mathrm{MulT}^\dagger$	MOSI	100	4	39.1	81.1	81.0
$\mathrm{MulT}^\ddagger$	MOSI	120	4	38.7	80.8	80.6
$\mathrm{MulT}^\dagger$	MOSEI	20	4	50.7	81.6	81.6
$\mathrm{MulT}^\ddagger$	MOSEI	120	6	50.5	81.4	81.5
MICA	MOSEI	120	4	52.2	83.4	83.1

Table A2. Performance (%) of ablation study on CMU-MOSEI by applying the alignment losses to the standard MulT. In each row, the corresponding design is progressively considered into the model. The number of epochs (120) and transformer layers (6) are fixed across all the rows. The results are averaged over 5 runs.

Model design	Acc_7	Acc_2	F1
MulT w/o align	50.5±0.22	81.4±0.37	81.5±0.40
+ MMD align	51.5±0.24	82.7±0.40	82.6±0.45
+ PE align (full model)	51.9±0.21	83.2±0.37	83.0±0.35

The proposed MICA approach adopts different parameter settings with the ones in the original MulT model. To validate that the performance improvement of the proposed MICA approach against MulT is not caused by using more epochs or transformer layers, we further conduct the comparisons by controlling the corresponding parameters. The performance is not improved when we use more epochs or transformer layers in MulT (see row 1-4 of Table A1). Moreover, our approach still outperforms MulT when we reduce the number of transformer layers to 4 in our approach (see the last row of Table A1). We use more epochs in MICA since the introduced alignment losses require more iterations to converge.

We also conduct the ablation study results on the standard MulT backbone. From Table A2, it is clear that the alignment losses are also effective on the standard MulT backbone. The improvements of both "MMD align" and "PE align" in the ablation study are significant for all the metrics ($p < 0.01$).