# Learning Modality-Specific and -Agnostic Representations for Asynchronous Multimodal Language Sequences: Supplementary Material

Dingkang Yang
Academy for Engineering and Technology, Fudan University

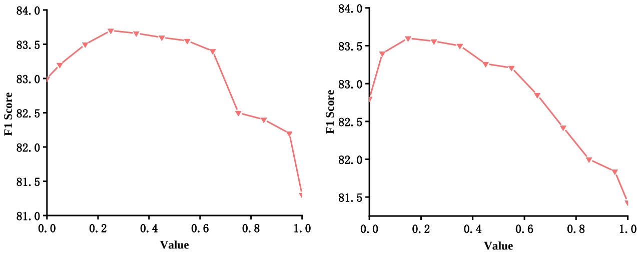
Haopeng Kuang
Academy for Engineering and Technology, Fudan University

Shuai Huang
Academy for Engineering and Technology, Fudan University
Engineering Research Center of AI and Robotics, Ministry of Education
Artifical Intelligence and Unmanned Systems Engineering Research Center of Jilin Province

Lihua Zhang*
Academy for Engineering and Technology, Fudan University
Jilin Provincial Key Laboratory of Intelligence Science and Engineering
Ji Hua Laboratory
lihuazhang@fudan.edu.cn

***Analysis of the coefficient*** $\mu$. In Figure 1, we analyze the coefficient $\mu$ used in the Predictive Self-Attention (PSA) module to balance the convolution-based prediction chain and the dot-product attention branch on the CMU-MOSI and CMU-MOSEI datasets. Specifically, the values are varied from 0 to 1 to observe the trend in the $F1$ score of the model. We find that as the values increase, the performances rise first and then start to drop significantly. Our model achieves the best performance when the coefficient $\mu$ of the CMU-MOSI and CMU-MOSEI datasets are set to 0.25 and 0.15, which aligns with the values adopted in the experiments. This observation suggests that it is beneficial and essential to introduce predictive attention maps with the appropriate balance. In addition, the model achieves the worst performance when $\mu$ is set to 1. This fact demonstrates the dominant role that dot product attention still plays in the PSA module.

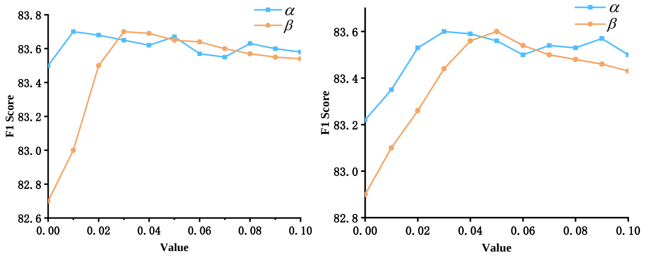
***Analysis of the trade-off parameters*** $\alpha$ ***and*** $\beta$. As shown in Figure 2, we conduct sensitivity analysis of the hyper-parameters on the CMU-MOSI and CMU-MOSEI datasets. The tested hyper-parameters include the trade-off parameter $\alpha$ for the separation loss $\mathcal{L}_{sep}$ and the trade-off parameter $\beta$ for the adversarial loss $\mathcal{L}_{agn} + \mathcal{L}_{spe}$. Specifically, the sensitivity analysis is conducted by varying the value of the corresponding hyper-parameter, while fixing the other hyper-parameters to the values adopted in the experiments. For both datasets, the $F1$ scores increase first and then start to decrease. Moreover, our approach achieves the best performance when the trade-off parameters $\alpha$ and $\beta$ of the CMU-MOSI and CMU-MOSEI datasets are set to $\{1e^{-2}, 3e^{-2}\}$ and $\{3e^{-2}, 5e^{-2}\}$, respectively. These appropriate balances are consistent with the values adopted in the experiments. Overall, we observe that the model is not sensitive to the parameters when the range of values is approximately between 0.03 and 0.1.



(a)CMU-MOSI

(b) CMU-MOSEI

**Figure 1: Analysis of the effect of the coefficient $\mu$ on performance by controlling for different values. The results are obtained by varying the value of the corresponding hyper-parameter, while fixing the other hyper-parameters to the values adopted in the experiments.**



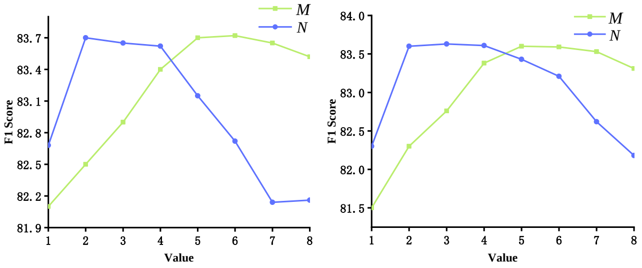
(a)CMU-MOSI

(b) CMU-MOSEI

**Figure 2: Sensitivity analysis of the trade-off parameters $\alpha$ and $\beta$ on the CMU-MOSI (a) and CMU-MOSEI (b) datasets. The results are obtained by varying the value of the corresponding hyper-parameter, while fixing the other hyper-parameters to the values adopted in the experiments.**

***Analysis of the number of layers in PSA and HCA modules.*** Figure 3 shows the effect of varying the number of layers in the PSA and HCA modules on performance. For the PSA module, we observe that the performances gradually increase and then stabilize as the number of layers rises on both datasets. Our model can achieve stable and optimal gain when $M$ is 5. For the HCA module, we find that the performances decrease significantly when the number of layers $N$ exceeds 2 on both datasets. A potential reason is that a

[*] indicates corresponding author.



(a)CMU-MOSI (b) CMU-MOSEI

**Figure 3: Sensitivity analysis of the number of layers $M$ and $N$ in the PSA (green line) and HCA (blue line) modules. The results are obtained by varying the value of the corresponding hyper-parameter, while fixing the other hyper-parameters to the values adopted in the experiments.**

densely hierarchical structure can increase complexity, limiting the model's performance.