DeepSuM: A Deep Sufficient and Efficient Modality Learning Framework

Zhe Gao[1], Jian Huang[2], **Ting Li[3*]**, Xueqin Wang[1]

[1]*School of Management, University of Science and Technology of China, Hefei, Anhui, People's Republic of China*

[2]*Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, People's Republic of China*

[3]*Department of Statistics & Data Science, Southern University of Science & Technology, Shenzhen, People's Republic of China*

[*]*Address for correspondence. Ting Li, Southern University of Science & Technology, Shenzhen, 518055, China. liting@sustech.edu.cn*

ABSTRACT

Multimodal learning has become a pivotal approach in developing robust learning models with applications spanning multimedia, robotics, large language models, and healthcare. The efficiency of multimodal systems is a critical concern, given the varying costs and resource demands of different modalities. This underscores the necessity for effective modality selection to balance performance gains against resource expenditures. In this study, we propose a novel framework for modality selection that independently learns the representation of each modality. This approach allows for the assessment of each modality's significance within its unique representation space, enabling the development of tailored encoders and facilitating the joint analysis of modalities with distinct characteristics. Our framework aims to enhance the sufficiency and effectiveness of multimodal learning by optimizing modality integration and selection.

Keywords: Multimodal learning, Representation, Modality integration, Modality selection