**Computational Image and Text Fusion Modeling for Multimodal Representation and Intent Prediction**

Wang Meng1, Suzan J. Obaiys1\*, Nurul Fazmidar Binti Mohd Noor1, Yeliz Karaca2 and Hui Zi1

1 Department of Computer System & Technology, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

2 University of Massachusetts (UMass) Chan Medical School, 55 Lake Avenue North, Worcester, MA 01655, USA

**Abstract**

The detection of user intention has become challenging upon the ever-evolving advances in multimedia technology as well as the widespread use of social media platforms. Intent recognition is considered to be one of the core components of mindreading, which is a critical process in social cognition. In terms of complexity, unimodal Artificial Intelligence (AI) models are significantly less complex than the multimodal counterparts which are characterized by intricate structures comprising multiple modalities as well as other systems.  Traditional unimodal approaches, especially those relying solely on either textual or visual information, often may manifest the deficiency in terms of capturing the intricacies of user intentions in multimedia content. To address this limitation, the fusion of image and text modalities using multimodal technology has emerged as a promising solution for intent detection. Multimodal technologies enable the summarization of the information retrieved from multitude of modalities so that complementary information is employed as an accumulative structure while redundant components of the modalities are filtered out accordingly. Owing to the heterogenous nature of data, the challenges emerging could arise from different sorts of noise, alignment of modalities and techniques of handling the missing data. Furthermore, compared with single-modal data such as images and text, multimodal data can contain more information while more accurately identify user intentions. Thus, we propose a new intent detection method, which includes two important stages of multimodal representation and fusion, for the exploration of the integration concerning the image and text data so that the accuracy of intent detection in multimedia content can be enhanced. The effectiveness of the approach proposed in this study, for intent detection based on image and text fusion, is verified by comparative experiments with the baseline model on the public multimodal intention dataset. Consequently, the proposed scheme within a revolutionary and transformative angle can enable the integration of information from diverse sources, both uni- and multimodal for a more profound and plausible understanding of the data under consideration, which in turn ensures the revealing of new insights and support towards a more extensive range of actionable applications.

**Keywords:** Artificial Intelligence (AI); Complexity; Intent detection; Multimodal technology; CLIP; Feature representation; Computational image and text fusion; Dynamic feature extraction; Hyperparameter setting; Multimodal adaptation gate architecture; Multimodal fusion.