Abstract

In the field of human-computer interaction, human intention detection is a challenging problem and a key link to achieving barrier-free communication between humans and machines. With the rapid evolution of multimedia technology and the widespread use of social media platforms, the detection of user intention has become increasingly challenging. Traditional unimodal approaches, especially those relying solely on either textual or visual information, often fall short 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 intention detection.

Compared with single-modal data such as images and text, multimodal data can contain more information and can more accurately identify user intentions. At present, there are few studies on intention detection based on image and text fusion, and they mainly consider how to integrate modal features in the feature fusion stage, while in the feature extraction stage, only a single-modality pre-trained model is used, which not only increases the complexity of the system, which increases the computational cost and may also limit the model's comprehensive understanding of the global context, making it difficult to effectively capture the correlation information between modalities.

This research endeavors to design a new intention detection framework, which includes two equally important stages of multimodal representation and fusion, to explore the integration of image and text data to enhance the accuracy and robustness of intention detection in multimedia content. In the feature representation part, the CLIP multimodal large-scale pre-training model is used to simultaneously extract text and image features, which simplifies system integration and saves computing resources while learning the associated information between modalities. In the feature fusion part, due to the different importance of text and pictures, an attention-based cross-modal fusion method is designed, which enables the model to dynamically adjust the attention to different modalities at each moment and capture important features, while reducing noise. Finally, to verify the effectiveness of the proposed model, this study developed an intention detection framework and carried out experimental verification on the intention detection dataset.

Key Words: intention detection, multimodal technology, CLIP, feature representation, multimodal fusion