Conversational Semantic- and Knowledge-guided Graph Convolutional Network for Multimodal Emotion Recognition

1910275 FU Yahui

Emotion recognition in conversation (ERC) has received significant attention in recent years and become a new frontier of natural language processing research due to its widespread applications in diverse areas, such as social media, health care, education, and artificial intelligence interactions. Therefore, the effective and scalable conversational emotion recognition algorithms are of great significance.

It is challenging to enable machines to understand emotions in conversations, as humans often rely on the contextual interaction and commonsense knowledge to express emotion. Therefore, both context and incorporating external commonsense knowledge are essential for the task of ERC. Graph convolutional neural network (GCN) technologies have been widely applied in the contextual information extraction due to its ability in learning complex structures. Most studies which utilize GCN only consider the syntactic information between utterances. Thus, for implicit emotional texts that do not contain obvious emotional terms, and the words are relatively objective and neutral, it is difficult to correctly distinguish the emotions if only the semantics of the utterances are considered. Commonsense knowledge is the basis for understanding contextual dialogues and generating appropriate responses in human-robot interaction, however, it has not well been explored for ERC. Previous studies either focused on extracting features from a single sentence and ignored contextual semantics, or only considered semantic information when constructing the graph,ignoring the relatedness between the tokens. We hypothesize that both semantic contexts and commonsense knowledge are essential for machine to analyze emotion in conversations.

To further tackle the above problems, we propose a new multimodal Semantic- and Knowledge-guided Graph Convolutional Network (ConSK-GCN) to effectively structure the semantic-sensitive and knowledge-sensitive contextual dependence in each conversation. On one hand, we construct models capturing the contextual interaction and intradependence of the interlocutors via a conversational semantic-guided GCN (ConS-GCN). In this context graph, each utterance can be seen as a single node, and the relational edges between a pair of nodes/utterances represent the dependence between the speakers of these utterances. On the other hand, we incorporate an external knowledge base that is fundamental to understand conversations and generate appropriate responses to enrich the semantic meaning of the tokens in the utterance via a conversational knowledge-guided GCN (ConK-GCN).

Furthermore, we introduce an affective lexicon into knowledge graph construction to enrich the emotional polarity of each concept. We leverage the semantic edge weights and affect-enriched-knowledge edge weights to construct a new adjacency matrix of our ConSK-GCN for better performance in the ERC task. In addition, we focus on multimodal emotion recognition using the acoustic and textual representations, because both text and prosody convey emotions when communicating in conversations.

Experimentation on the multimodal corpus IEMOCAP and MELD show that our methodology could effectively utilize the contextual dependence of the utterances in a conversation for emotion recognition. Moreover, detailed evaluation indicates that our approach is superior than several state-of-the-art baselines in both uni-modality and multi-modality emotion recognition.