Multimodal Representation Learning with Homogeneous and Diverse Constraints Using Multi-emotional Audiovisual Features for Depression Detection

Shiyu Teng, Jiaqing Liu, Shurong Chai, Tomoko Tateyama, Xinyin Huang, Lanfen Lin, and Yen-Wei Chen*

Depression is a prevalent mental disorder affecting a significant portion of the global population, leading to consiferable disability and contributing to the overall burden of disease. Consequently, designing efficient and robust automated methods for depression detection has become imperative. Recently, deep learning methods, especially multimodal fusion methods, have been increasingly used in computer-aided depression detection. Importantly, individuals with depression and those without respond differently to various emotional stimuli, providing valuable information for detecting depression. Building on these observations, we propose an intra- and inter-emotion transformer-based fusion model to effectively extract depression-related features. The intra-emotion fusion framework aims to prioritize different modalities, capitalizing on their diversity and complementarity for depression detection. The inter-emotion model maps each emotion onto both invariant and specific subspaces using individual invariant and specific encoders. The emotion-invariant subspace facilitates efficient information sharing and integration across different emotions, while the emotion-specific subspace seeks to enhance diversity and capture the distinct characteristics of individual emotions. Our proposed intra- and inter-emotion fusion model effectively integrates multimodal data under various emotional stimuli, providing a comprehensive representation that allows accurate task predictions in the context of depression detection. We evaluate the proposed model on the Chinese Soochow University depressive severity dataset, and the results outperform state-of-the-art models in terms of concordance correlation coefficient (CCC) and root mean squared error (RMSE).