ABSTRACT

Schizophrenia is a serious disease that can not ordinarily express real thinking. Psychological counseling is currently a diagnosis approach that doctors mainly exploit visual and speech perception to estimate the mental state. From visual perception, a sick person may not show wealthily emotion and behave inconsistently through the time series, thus, leading to an inaccurate diagnosis. With the recent remarkable progress of machine learning, we are thus motivated to design a deep learning algorithm to model the visual emotion signal to assist the doctor in inferring the mental state of the patient. Particularly, we propose a bottom-up approach that can effectively model complex temporal relations to generate a precise emotional representation. Our method first exploits graph convolution network (GCN) to captures the consistent emotional signal from each video frame to better describe the local dependency (face - body) and then draws on the transformer to integrate each temporal representation for the final prediction. With the impressive power of GCN, we can effectively handle the dense data structure and produce a better local feature. On the other hand, the transformer can endure the capability to deal with various lengths of video data and integrate a better global representation for final prediction. In addition to the effort, we also design a depression module to evaluate the patient state to provide an additional reference for doctors further. In the experiments, we demonstrate the effectiveness and efficiency of our approach to mental disorder diagnosis.

**Keywords:** Deep learning, Emotion Recognition, Context-aware Emotion Recognition, Depression Recognition, Graph Convolutional Network