NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

Deep Learning for Detection of Depression

by Combining Audio and Visual Images

Student: Wong Xiaoqing Supervisor: Professor Jagath Chandana Rajapakse

Motivation

Depression is a common mental disorder that affects approximately 280 million people worldwide. It could cause great suffering and impair one's ability to function daily at work, or even lead to suicide. However, in most cases, symptoms can be improved with early intervention from healthcare professionals. Therefore, it is critical that patients receive the necessary diagnosis and treatments.

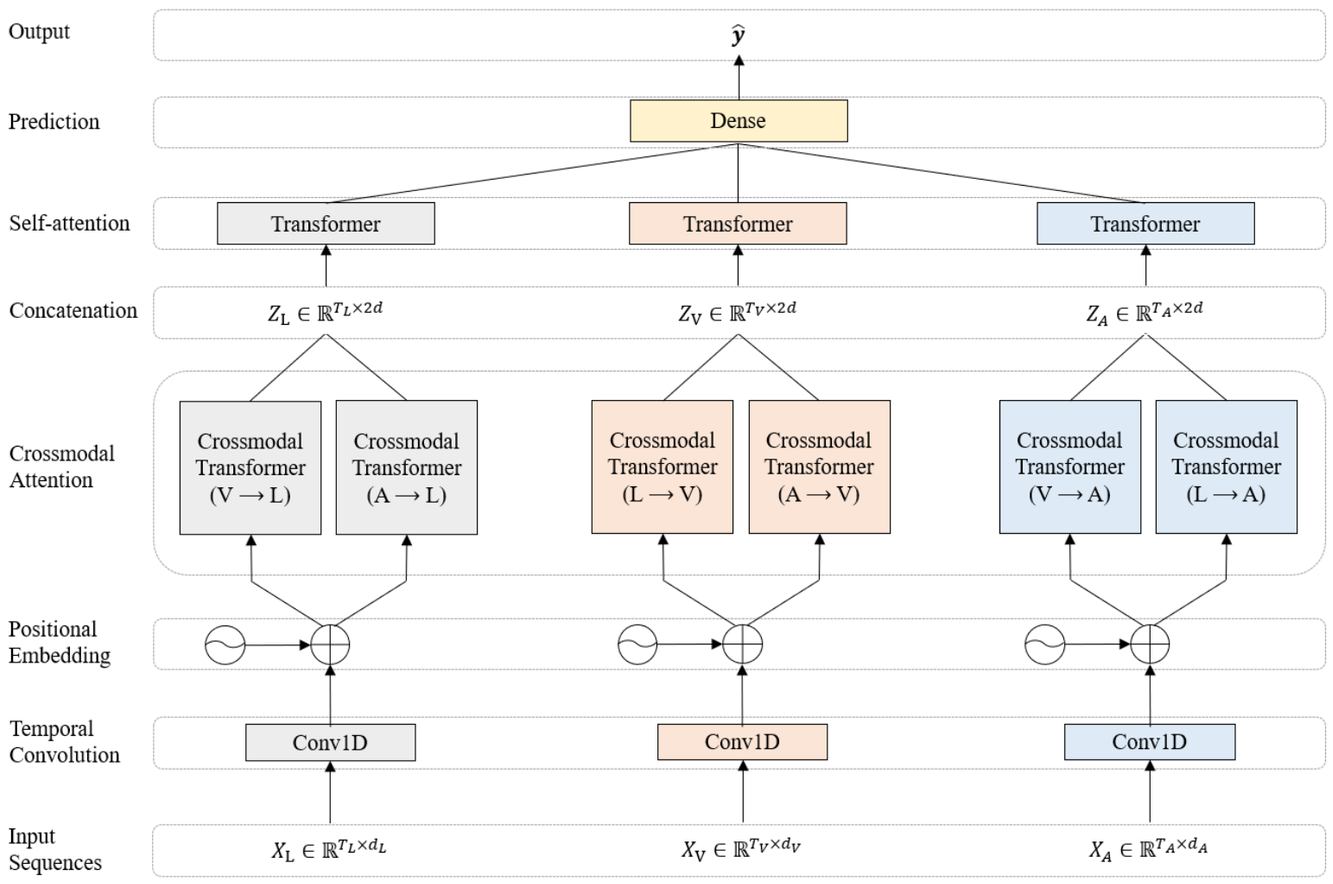
Objective

This project aims to develop deep learning models to **classify levels of depression**. The main focus is to **combine information from audio and video features** of clinical interviews to predict depression ratings of the patient – None, Mild, Moderate, Moderately Severe and Severe. Transformer based architectures and attention mechanisms are used to identify interactions between different features from audio and video and perform fusion of the multimodal features.

Deep Learning Models

Unimodal Classification + Multimodal Transformer → Comparison

Multimodal Transformer Architecture



Multimodal Transformer Results

Model Hyperparameters	
Batch Size	32
Number of Epochs	30
Layers of Crossmodal Transformer Blocks	5
Layers of Multi-head Attention Module	5
Model Results	
Class	***F1 Score*** / ***Accuracy***
None	0.840079 / 89.1304%
Mild	0.441915 / 58.6304%
Moderate	0.687168 / 78.2608%
Moderately Severe	0.840079 / 89.1304%
Severe	0.778005 / 84.7826%

Conclusion

From experiment results, the multimodal transformer can learn from multimodal data and perform better than audio and visual unimodal classifiers, while text classifier outperforms it due to limitations.