Imperial College London

MEng Individual Project

Imperial College London

Department of Computing

Emotion Recognition Using a Multi-Modal Approach

Author:
Yihui Ilona Zhu

Supervisor:
Dr. Anandha Gopalan
Prof. Abbas Edalat

Second Marker:
Dr. Ronald Clark

June 22, 2022

Abstract

In recent years, mental health issues are becoming more prevalent around the world, and the COVID-19 pandemic has worsened the situation. Although many psychotherapeutic interventions are useful, the availability of these treatments are often limited. The development of self-administrable psychotherapy with advanced technology like virtual reality (VR) provides patients with easier access to effective therapies and an immersive experience at the same time. As the VR platform provides suggestions based on the emotion detected, an accurate emotion recognition model is essential. Since humans express emotions through different modalities, previous students have developed models that target different modalities. The drawback is that these models operate separately.

This project aims to implement a robust fusion mechanism that combines the existing models and thus devise a single model that predicts the strongest emotion experienced by the human subject.

The final model developed is trained on CMU-MOSEI and successfully integrates features from the video, audio and text modalities together. It outperforms the baseline model established for predicting the four basic emotions (i.e. happy, sad, angry and disgusted). The model also maintains its performance on human trial data, suggesting that it generalises quite well to unseen data. Additionally, it has been demonstrated that the text modality is the most crucial for a high-performing model.

Acknowledgements

First and foremost, I would like to express my profound gratitude to Dr. Anandha Gopalan for his invaluable guidance and support throughout the project. I'm very grateful that he gave me the opportunity to work on a very interesting project. He has also been very understanding and encouraging from the day I started the project. I would like to extend my gratitude to Prof. Abbas Edalat for co-supervising me and specifically clarifying many questions I had at the beginning of the project.

Secondly, I would like to thank Neophytos Polydorou for his generous help with many aspects of the project – from setting up the project and the human trial, to giving me suggestions on how to improve the model's performance and interesting areas to explore for a more insightful evaluation.

Thirdly, I would like to thank Dr. Ronald Clark for providing helpful advice and feedback on the structure of my project.

Additionally, I want to thank all my friends for their companionship and support over the past four challenging years. In particular, I want to thank them for their help on the human trial in this project.

Finally and most importantly, I would like to thank my family, especially my sister, for always encouraging me and believing in me. None of these would have been possible without their love and support. Thanks for always being by my side and inspiring me to become a better person.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

Mental health disorders have always been a challenging issue faced by many people around the world. A study by WHO [2] estimates that mental health conditions account for 20% of years lived with disability (YLDs) globally and cause more than 1 trillion USD economic loss per year. The COVID-19 pandemic has further worsened the situation and there has been a significant increase in the number of people facing mental health issues and negative emotions. Studies have estimated that in 2020 there were around 76.2 million more cases of anxiety disorders and 53.2 million more cases of major depressive disorder (MDD) globally due to the COVID-19 pandemic [3]. Although psychotherapeutic interventions can help reduce psychiatric morbidity [4], they often require the presence of a therapist, who may not always be available, especially when mental health problems are highly prevalent worldwide.

To address this issue, particularly the limited access to psychotherapists, the Algorithmic Human Development (AHD) research group at Imperial College London has developed a self-administrable psychotherapy known as Self-Attachment Technique (SAT) [5]. This therapy is more scalable as it involves almost no interactions with a psychotherapist. With advances in technologies such as Artificial Intelligence (AI) and Virtual Reality (VR), the accessibility of the therapy can be further increased. Particularly, the AHD group has also developed a VR platform that helps deliver SAT sub-protocols to the user interactively [6]. The VR platform uses an emotion recognition algorithm and provides the user with suitable SAT sub-protocols based on the emotion detected. Therefore, it is vital to ensure good performance of the emotion recognition algorithm used.

In real life, humans express and perceive emotions through a number of behaviours, such as facial expressions, body gestures, speech cues, text, physiological signals and so on [7]. As a result, emotion recognition algorithms largely rely on these modalities to classify human emotions. In recent years, the AHD group has supervised a few projects that developed various emotion recognition algorithms based on different modalities: audio-text, visual (i.e. video) and heart rate (HR). However, the existing emotion recognition algorithms operate separately and only use the specific modality input it was designed for.

1.2 Objectives

It has been shown that emotion recognition based on single modality (uni-modal approach) has its limitations and that multi-modal emotion recognition that integrates input from

multiple modalities has become a popular research area in recent years [8]. Hence, the aim of this project is to enhance the existing emotion recognition algorithms by implementing a fusion procedure such that the final algorithm is able to utilise more or all modalities. Specifically, this involves the following:

- Adapt existing models to work with predicting a single emotion, which can then be the basis for suggesting suitable SAT sub-protocols.
- Develop an overall model that utilises the full-face video together with the other modalities. The model will be based on the existing models with a fusion mechanism, which should be carefully designed and implemented based on research of state of the art models and their adopted fusion strategies.
- Adapt the model developed to work in a setting with facial occlusions, specifically the occlusion caused by wearing a VR headset.

1.3 Contributions

The main contributions of this project can be summarised as follows:

- Adapted and fine-tuned (when necessary) the existing models from individual modalities to act as the feature extraction component of the overall model.
- Designed and implemented an overall model that combines features from audio, text and video modality using a fusion mechanism based on cross-modal attention. A baseline has also been established using a simpler fusion mechanism.
- Adapted the final model to a setting with facial occlusions, where the area around the eyes is covered.
- Evaluated the performance of the model in different settings using different fusion mechanisms and with respect to the improvement from existing models. A detailed analysis on the different combination of modalities is also provided, concluding that the text modality can be considered the most important in multi-modal emotion recognition.

Chapter 2

Background

2.1 Mental Health

2.1.1 Self-Attachment Technique (SAT)

Attachment theory introduced by Bowlby [9] presents the idea that children develop a strong attachment – an emotional bond – with their primary caregivers as they grow up. It is believed that the earliest attachment formed can heavily influence the way children perceive the social world and their behaviour. The impact on children's emotion and personality continues into adulthood and later stages of their lives [5]. There are four main types of attachment: secure, avoidant, ambivalent and disorganised [10]. Children who are securely attached to their parents can often self-regulate negative emotions and maintain a stable mindset; however, having the three types of insecure attachment can negatively affect their development and thus expose them at the risk of mental health problems [5].

SAT is developed based on attachment theory. Instead of building a secure attachment with new caregivers such as the psychotherapist, the individual practising SAT acts as the parent for a child-self representation of the individual [5]. The child-self resembles the individual with insecure attachment in early life and the adult-self of the individual would try and form a secure attachment with the child-self. As a result, the child-self no longer experiences insecure attachment and can grow mentally and emotionally [5]. By forming this secure attachment with themselves, the individual can now self-regulate emotions, especially ones resulting from attachment insecurities, which can help improve mental health conditions.

2.1.2 Virtual Reality Therapy

Virtual Reality (VR) has recently become a popular technology that is capable of simulating immersive and interactive 3D environments. It is believed that VR can help solve the issue of limited access to useful psychological interventions in many ways: automated delivery of treatments, directly engaging patients in situations that may cause psychological distress, as well as providing a safe zone for patients to try different responses and transfer their learning to the real world [11].

Therefore, VR has been utilised extensively in clinical treatments. With regards to psychological interventions, various studies have experimented with using VR to treat psychiatric disorders such as PTSD, phobias, autism, anxiety and stress disorders [12]. Results have demonstrated that VR can potentially help improve mental health and address the disorders by exposing patients in the environments where distress stems from [12].

As SAT is self-administrable, an interactive VR platform [6] has been built by members of the Algorithmic Human Development (AHD) research group to support the user perform SAT sub-protocols in an immersive environment. The platform consists of a virtual agent avatar, a child avatar representing the childhood-self of the user and a real-time emotion recognition algorithm. The virtual agent guides and assists the user through the process of completing self-attachment interventions; the user interacts with the child avatar to bring more positive emotion to the child avatar and thus the user. The emotion recognition algorithm is used to detect the user's emotion and adapt the child avatar's emotion accordingly. Currently, the emotion recognition algorithm takes audio input and classifies the emotion into one of these categories: happy, sad, angry, surprised, disgusted and fearful [6]. Separate emotion recognition algorithms that use videos of the user wearing the VR headset and heart rate data have been developed, but they have yet to be integrated.

2.2 Machine Learning

Machine Learning (ML) is a branch of computer science that enables computers to improve performance at different tasks like predictions by 'learning' the underlying complex patterns from data, often referred to as the training set [13].

Supervised learning, unsupervised learning and reinforcement learning are the most common types of ML algorithms. In particular, supervised learning uses training set that includes both the input data and the outcome [14]. In other words, the model knows what the ground truth is and can utilise that to better capture the pattern from data. For the purpose of this project, supervised learning is used: labelled emotion data is used to train an ML model that classifies the VR platform user's emotion.

The following sections provide an overview of various ML architectures and frameworks that serve as the foundation of existing emotion recognition models.

2.2.1 Artificial Neural Network (ANN)

ANN, often simply referred to as 'neural network', resembles the structure of a human brain, where there are many layers of interconnected neurons that take an input signal, process it and output a result signal [15].

In a typical hidden layer, every neuron is connected to all the neurons in the next layer with an associated weight matrix. Each neuron also has a bias term associated. The output of the layer is computed as the product of the neuron input values and the weight matrix, summed together with the bias terms. Different activation functions are often used to scale the output of a neuron within a given range that is the most suitable for the specific ML task.

2.2.2 Convolutional Neural Network (CNN)

CNN is a type of neural network specifically designed to take images as the input by introducing a convolutional layer on top of other common layers, such as fully connected layers [16]. In a CNN, each neuron only sees a small region of the image input and extracts useful features from the local region. The local information is then combined to allow the network to form a global understanding of the input.

A typical convolutional operation involves sliding a kernel across each pixel. For each source pixel, the new output pixel value is obtained by element-wise multiplication of the weights associated with the kernel by the local region pixels and weighted summation [16].

Pooling layer is often used with convolutional layers to reduce the dimension of feature representations – similar to the effect of downsampling the image.

2.2.3 Recurrent Neural Network (RNN)

RNN is a type of neural network that contains a memory unit, which makes it very useful and applicable to tasks that require processing sequential data. As RNN considers the outputs from previous timestamps when computing the current output, it stores the previously seen information and uses it to make future predictions [17]. This makes RNN suitable for the task of emotion recognition as the input to the model is often a continuous sequence of timeseries data and RNN is capable of modelling the dependencies between different utterances of the input. In particular, RNN is often used to incorporate the context of the input sequence and thus used in Natural Language Processing (NLP) tasks.

2.2.4 Long Short-Term Memory (LSTM)

One disadvantage of RNN is that it suffers from vanishing and exploding gradient of the loss function during training, which limits its ability to model the dependency between inputs separated by a long time interval [17]. LSTM was introduced to mitigate this problem [18]. LSTM introduces cell state that is capable of storing information from an earlier point in time and gates that controls the extent to which the network forgets past memory. This way, the network can selectively keep useful and important information in the cell state flow and forget information that is no longer relevant. For the task of emotion recognition, this additional feature may be more powerful for capturing the context, especially in a real-time setting where the model takes long sequences of input.

2.2.5 Gated Recurrent Unit (GRU)

GRU is a variation of LSTM introduced by Cho et al. [19]. It combines the cell state introduced by LSTM with the hidden state of the network and only has two gates. Moreover, fewer operations are involved in GRU due to its simpler structure compared to LSTM, making it faster to train. As there has not been empirical evidence that suggests which of LSTM and GRU has better performance, both are RNN variations that are commonly used, and the suitability often depends on the specific task [20].

2.2.6 Attention

Bahdanau et al. [21] first introduced the concept of attention as a way of addressing the issue of RNNs suffering from short-term memory and helping the model keep track of long input sequences. In particular, incorporating attention in the network architecture allows the model to recognise important features or sections of input sequences. The model can then rely more heavily on these highlighted features.

In essence, the attention mechanism aims to obtain a weight distribution that puts more emphasis on relevant input data. The weight values can be computed by generating scalar values that score the alignment between inputs and outputs and normalising the scalar values by a softmax function [22]. This alignment score is often obtained by checking the compatibility between a query-key pair using its corresponding value output.

2.2.7 Transformer

Transformer introduced by Vaswani et al. [23] is a new architecture that can process sequential data efficiently without using any recurrence networks. Instead, the transformer

model relies on multi-head attention and a small feed-forward neural network to capture the dependencies between input and output data. As transformer allows more parallelisation when processing data, it can be trained faster than normal RNN [23].

The transformer model adopts an encoder-decoder structure, where the positional encoding of input and output are included in the feature embedding to compensate for the removal of recurrent layers. Multi-head attention is used extensively in the model: the encoder and decoder each contains a multi-head self-attention module, and a separate multi-head attention layer is added between the encode and decoder [23]. In particular, the encoder self-attention is unmasked, which removes the position dependency and allows the transformer to utilise input data at all positions, making it bidirectional and preferred over the original RNNs.

2.2.8 Natural Language Processing (NLP)

NLP is a branch of Artificial Intelligence (AI) that often involves writing programs that enable computers to process human natural language. In this project, the ML model needs to process and interpret human emotions from textual input and NLP is involved in this respect. Previous research in this field has established many powerful language models for NLP tasks, which can be specifically applied to emotion recognition.

Bidirectional Encoder Representations from Transformers (BERT)

BERT [24] is a widely used model for various NLP tasks based on the transformer architecture. Instead of adopting an encoder-decoder structure, BERT consists of multiple encoders stacked together without any decoder. Novel pre-training tasks have been introduced for BERT, including Masked Language Model (MLM) and Next Sentence Prediction (NSP) [25]. During training, input text is first converted to tokens, and a subset of tokens is randomly selected and hidden from the model. The goal of the MLM task is to predict the original word and output a numerical feature representation vector once training finishes [24]. The model is able to learn from the left and the right context of the tokens, showing that the learning becomes bidirectional. On the other hand, the NSP task helps the model learn meaningful pairwise sentence relationship. Combined with billions of training data and TPU for computational power, the pre-training tasks help BERT achieve its performance improvements on 11 of the most common NLP tasks [25]. As BERT is a good-performing pre-trained model, it can be fine-tuned to adapt to specific tasks including emotion recognition. The commonly used BERT models include BERT-base and BERT-large.

XLNet

As BERT uses a masking technique as part of the pre-training tasks, the dependency between masked words is ignored. XLNet is a bidirectional autoregressive pre-trained transformer introduced to overcome the shortcomings of BERT [26]. Instead of processing and learning from inputs with fixed sequential order, XLNet considers all permutations of the factorisation order of the input sequence and attempts to maximise the expected likelihood of the sequence [26]. Due to the usage of random ordering, XLNet can capture the context information across the whole sequence and learn the dependencies between words [27]. In terms of its performance, XLNet achieves better performance than BERT on 20 common NLP tasks [27] and is often selected for NLP tasks due to its advantage in handling dependencies better.

RoBERTa

RoBERTa stands for Robustly optimised BERT approach and has been proposed to further improve BERT's performance by modifying its training procedure [28]. The main modifications made to the original training method are removing the NSP objective and changing the masking pattern to be dynamic (i.e. changing masked tokens during training) [27]. Furthermore, the RoBERTa has been trained for longer on longer sequences of input and larger batch sizes [28].

Compared to BERT and XLNet, RoBERTa has been trained on more data and the results demonstrate that RoBERTa achieves state of the art performance and consistently outperforms BERT-large and XLNet-large when using the same masked words objective as BERT-large [28]. Therefore, RoBERTa may be preferred over BERT for its better performance, although it should be noted that the training time for RoBERTa is longer.

T5

The Text-to-Text Transfer Transformer (T5) [29] has been introduced recently as a language model that utilises transfer learning. As opposed to BERT-like models that only generates a single label or a restricted span over input as the output, T5 is a text-to-text model that always returns text strings. This offers T5 much greater flexibility to adapt to many NLP tasks, ranging from sentiment classification, document summerisation to regression tasks, where the model is trained to output the string representation of desired numerical values [29]. In fact, Roberts et al. [30] claim that T5 is capable of being adapted to any NLP task with minimal changes to the network parameters required. Experimentation results have shown that T5 can achieve the state of the art performance and outperform previous state of the art models in certain NLP tasks [30].

2.3 Datasets

The goal of the project is to implement an effective fusion strategy for an ML model that is capable of classifying human emotions leveraging multiple modalities. As a result, different modalities of the input data need to be provided for the model training process. Therefore, the focus is on finding suitable multi-modal datasets that can be used to train and evaluate the model.

IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [31] is one of the most commonly used datasets for emotion classification tasks. It consists of video recording, audio recording, text transcripts, face motion capture and hand movements data from 10 actors in scripted and improvised scenarios [31]. The entire dataset contains a total of approximately 12 hours of data, split into 5 sessions, each with a pair of actors.

IEMOCAP is labelled by multiple annotators with discrete emotion classes, as well as continuous sentiment values. For the categorical emotion labels, each utterance is given a class out of the following: happy, sad, angry, disgusted, fearful, surprised, frustrated, excited and neutral; for the sentiment values, each utterance is assigned valence, arousal and dominance values.

The main disadvantage of IEMOCAP is that although it contains many emotion classes, the dataset is imbalanced – much fewer samples of surprise, anger and fear are included compared with the other classes [31]. This implies that a model trained on IEMOCAP may

not be able to distinguish the underrepresented emotions due to the inadequate amount of training samples it has seen from these emotions.

CMU-MOSEI

CMU Multi-modal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [32] is the largest multi-modal dataset for emotion recognition and sentiment analysis to date. Different from IEMOCAP that consists of recordings of actors, CMU-MOSEI is collected from 3228 online videos of 1000 distinct speakers covering about 250 distinct topics. The 3228 videos are further split into sentence level and a final set of 23453 sentences are selected and included in the dataset, which sums to around 65 hours of data [32].

Each sentence in CMU-MOSEI has a video clip, audio recording and a text transcript associated with it. In terms of the labels provided, each sentence is annotated with a sentiment score from -3 to 3 and 6 emotion intensity scores. Each emotion score ranges from 0 to 3 and represents how strong the emotion presence is.

CMU-MOSEI also has the drawback that the dataset is imbalanced, and this is a common problem faced by most multi-modal datasets found. However, due to the large size of CMU-MOSEI, the dataset imbalance issue would be less severe compared to IEMOCAP.

MELD

The Multimodal EmotionLines Dataset (MELD) dataset [33] is an extended version of EmotionLines dataset [34]. It contains audio, visual and text modality data from approximately 1400 dialogues with multiple speakers, which can be further split into over 13000 utterances, extracted from the TV series 'Friends' [33].

The dataset is annotated by 5 workers using a majority vote scheme that determines a sentiment label and a discrete emotion class for each utterance. The sentiment label is one of neutral, positive and negative; the discrete emotion label is given as one of joy, sadness, anger, disgust, fear, surprise and neutral. If the majority vote system cannot agree on a common label, the utterance is not included in the dataset [33].

The main limitation of MELD is that the utterances included are all scripted. Training the model on scripted input only can limit its ability to generalise to unseen data from the real world as people may express certain emotions differently in real life.

RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [35] is a widely used dataset for speech emotion recognition. It is an audio-visual dataset that contains 7356 utterances collected from 12 female and 12 male actors performing speeches and songs in one of the asked emotions. All recordings in RAVDESS have been acted in North-American accent and the reliability of the performance is ensured by ratings produced by 247 research participants [35].

The displayed emotions for speeches are happy, sad, angry, fearful, surprise, disgusted and calm, and the ones for songs are happy, sad, angry, fearful and calm. Moreover, each emotion is acted at two different intensities – normal and strong.

The disadvantage of using RAVDESS is that the text modality is not provided. Although there exists many powerful engines and frameworks that can generate the transcript for a speech file, doing so introduces more uncertainties during the learning process. Ideally,

the text input to the model should match the audio recording exactly. Although one can manually check that the transcripts are correct, this process may be tedious and not time-efficient for 7356 utterances.

K-EmoCon

K-EmoCon [36] is a newly introduced multi-modal dataset that contains the widest range of modalities. For instance, the dataset contains video recording of a subject's facial expression and gesture, audio recording, as well as data from biosignals including electroencephalography (EEG) and electrocardiography (ECG). The data is collected from 16 sessions of paired debates, and the distinct feature of K-EmoCon is the different perspectives it considers: labels of data come from the subject, debate partner and external observers [36].

Labels from multiple perspectives are provided for K-EmoCon on the continuous arousal and valence value (on a scale of 0 to 5) and the emotion intensity score. In particular, there are 5 emotion classes included (i.e. happy, sad, angry, nervous and cheerful) and each emotion score lies in the range of 1 to 4.

The biggest limitation of K-EmoCon is the small dataset size – the overall length of the dataset is only around 173 minutes. Although K-EmoCon contains most of the modalities desired for the training of the model of this project, after splitting the dataset into train/validate/test folds, the training set would be too small for the model to learn useful features from all emotions. Furthermore, the discrete emotion classes are not a standard representation of human basic emotions. In particular, fear is not included, but it is a common emotion felt by people who have mental disorders. Thus, it is important that the model can detect fear when the user is interacting with the VR platform.

2.4 Related Work

2.4.1 Simple Fusion

End-to-End Multimodal Emotion Recognition Using Deep Neural Networks

An end-to-end emotion recognition model featuring a deep learning approach and using the audio and video modality has been proposed [37]. The model takes raw audio and video data as input and extracts features from them using a CNN and ResNet-50 respectively. Several LSTM stacks are added on top of the feature extraction networks to capture the contextual information from the input data. The final output of the model is the predicted arousal and valence values.

The model is evaluated on RECOLA using concordance correlation coefficient (CCC) between the ground truth values and the predicted arousal and valence. Since the experiments with uni-modal models are not of interest for the purpose of this project, the focus is on the performance of the multi-modal model. Results show that the proposed model outperforms the other existing models in both arousal and valence prediction, with the arousal prediction improving by a large margin. Experiments have also been conducted to make use of extracted facial landmarks instead of raw pixels and the result demonstrates that the model may benefit from utilising more accurate facial landmark predictions.

This model is not directly applicable to this project as it only uses two modalities and predict continuous sentiment values instead of categorical emotion labels. However, its performance suggests that using RNNs, particularly LSTMs, may be beneficial when predicting emotions. As simple concatenation on feature level is used as the fusion mechanism, it only allows the

model to utilise the input data from both modalities. The more complicated information such as context and dependencies between input data is likely modelled by the LSTM stacks.

Multimodal Sentiment Analysis using Hierarchical Fusion with Context Modelling

Majumder et al. [38] present an unconventional feature fusion strategy for combining multiple modalities – in this case audio, visual and text. Uni-modal feature vectors for the three modalities are first extracted using openSMILE, 3D-CNN and a deep CNN respectively. After transforming the feature vectors to equal dimensionality, each pair of modality vectors are first fused together by fully connected layers. The three bi-modal fusion output vectors are then fused together to form the final tri-modal representation vector. Therefore, the proposed method distinguishes itself from the common fusion method like simple concatenation and element-wise summation, and approaches the fusion in a hierarchical way. Furthermore, to ensure that context information is incorporated into the learning process, each layer of the hierarchical fusion (i.e. input feature, bi-modal fusion and tri-modal fusion) includes a GRU for each representation vector.

The model is evaluated on IEMOCAP and CMU-MOSI using accuracy and F1 score. It should be noted that since CMU-MOSI contains seniment labels only, the 2-class accuracy results are not applicable in this setting where the task is on discrete emotion classification. Looking at the F1 score on predictions of a subset of emotions (i.e. anger, happiness, sadness and neutral) from IEMOCAP, the multi-modal approach outperforms the baseline fusion methods and the tri-modal model has the best performance. Adding in the contextual information also helped to increase the F1 score for 1-2%.

The advantage of the hierarchical approach is that the model is able to consider the full joint relationship between all three modalities. The idea of extracting information from each pair of modality inputs is also present in other literature, which will be discussed next. This suggests that the joint relationship between all input modalities should be considered.

2.4.2 Similarity-Based Fusion

Multi-Modal Emotion Recognition by Fusing Correlation Features of Speech-video

The end-to-end emotion recognition model proposed is based on correlation features of speech and video input; the model also incorporates the class information for more accurate classification [39]. The speech features are extracted using a 2D-CNN and the video features are extracted using a 3D-CNN. These features are then fused together by correlation analysis with class information included. The final emotion classification is completed by support vector machines (SVM) with Gaussian kernel. The model is trained and evaluated on RML, eNTERFACE05 and BAUM-1s using the recognition rate. The proposed model outperforms previous state of the art models, achieving recognition rate higher than 90%.

With regards to the fusion mechanism adopted, a novel correlation analysis algorithm based on canonical correlation analysis (CCA) is used. Instead of maximising the correlation between projections of the feature vectors on their basis vectors in standard CCA [40], the proposed algorithm adds a weight matrix representing the similarity between classes and an inter-class divergence matrix to the objective function. Adding both matrices allows the model to effectively distinguish similar classes and improves the performance of the model.

In the context of this project, the main limitation of this model is that it only uses the

audio and video modality and the correlation analysis designed is quite specific to the bi-modal architecture. This increases the difficulty of extending the model to utilise more input modalities. Moreover, the model is trained and evaluated on less commonly used datasets compared to the other recent models in the field. The evaluation metric 'recognition rate' is also defined vaguely, making it difficult to directly compare this model's performance with others. However, the idea of using correlation between features is interesting and new. As previous studies have shown that when emotions are correctly predicted, each of the modality input has correlation with other modality signals [41], it may be worth exploring a fusion mechanism that uses correlation between feature vectors.

M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues

M3ER is another novel multi-modal emotion recognition model developed that uses the correlation between feature vectors from audio, video and text modality [42]. M3ER first introduces a modality check step using Canonical Correlation Analysis (CCA) to distinguish effective and ineffective modalities. The effective feature vectors are then passed into the fusion and classification network. Specifically, each feature vector is passed through an LSTM and into a shared attention module with a shared memory variable. The memory variable and inputs to the LSTMs are updated every few iterations. The outputs from the LSTMs are fused together by multiplicative fusion [43] and then used to make the final predictions. A modified loss function is also used to boost the modalities that make correct predictions and ignore the wrong predictions.

The model is trained on CMU-MOSEI and evaluated on IEMOCAP and CMU-MOSEI using mean accuracy and F1 score. M3ER achieves high performance on both test sets, both of which are improvements from previous state of the art models. Another interesting component of the model is the generation of proxy features for ineffective modalities during test time. The main idea is that there exists a linear transformation that can approximate the ineffective modality feature vector using the feature vector from an effective modality.

Different from the model by Chen and Zeng [39], M3ER uses CCA in early stages as opposed to using it as a fusion strategy. The advantage of incorporating the correlation early in the modality check step is that the model is more robust to noisy data as it can identify ineffective modalities. The multiplicative fusion mechanism used also helps the model to put more weights on reliable modalities. Although the model has some limitations such as it often gets confused between 'angry' and 'happy', the model is more extensible and it leverages more modalities. Additionally, the idea of multiplicative fusion at final stages of the classification network may help improve the performance of fusion compared to normal additive fusion.

2.4.3 Attention and Transformer-Based Fusion

Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation

Multilogue-Net is a RNN architecture proposed for both emotion detection and sentiment analysis, specifically in a conversational setting, leveraging all of video, audio and text modality [44]. The model is designed based on the assumption that emotion and sentiment mainly depend on four factors: speaker state, speaker intent, preceding and future emotions, and the context of the conversation. In Multilogue-Net, every utterance has three independent feature representations corresponding to each modality. For each modality input, multiple GRUs are used to model the individual speaker's emotion and state as well

as the context of the overall conversation. The emotion vectors from different modalities are fused together using pairwise attention. A residual connection with the GRU outputs is also added at each timestamp. The final concatenated vector is passed through the final layers of the network to compute the emotion class probabilities.

Different versions of the model have been tested on CMU-MOSI and CMU-MOSEI using accuracy and F1 score. The results demonstrate that combining all of audio, video and text modality leads to the best performance. This tri-modal approach outperforms previous state of the art models and achieves 80.01% F1 score. On the other hand, the audio-video model performs the worst, which indicates the importance of the text modality.

From the fusion mechanism adopted, it has been shown that the pairwise attention adopted is very effective – it improves the performance of the model as long as the text modality is present. Another interesting observation made is that Multilogue-Net performs a lot better when trained and tested on CMU-MOSEI, suggesting that a larger dataset is beneficial for the model to learn complicated intermediate representations. The attempt to capture the context of the conversation is not directly applicable to this project as the focus is on classifying emotion of the individual user instead of in a conversation.

A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis

After the introduction of the transformer architecture, emotion recognition utilising transformers has attracted many researchers' attention. The proposed transformer-based joint-encoding (TBJE) uses one transformer for each modality and jointly encodes the transformer outputs by modular co-attention and a glimpse layer [45]. The model makes use of pre-extracted features for audio, video and text modality; each feature is passed through the dedicated transformer and the co-attention mechanism is applied. In the proposed model, the text modality acts as the primary modality and it is used to modulate the self-attention of the other two modalities. The glimpse layer is added at the end for each modality to compute a new projection representation and the three modality outputs are fused together by element-wise summation. This vector is then used to make the final emotion prediction.

The model is evaluated on CMU-MOSEI test set using accuracy and F1 score and the performance on accuracy outperforms previous state of the art models. However, since accuracy is largely influenced by the majority classes in the dataset used and emotion recognition datasets are often imbalanced, other metrics should be considered along with accuracy. From the perspective of F1 score, TBJE performs slightly worse than Multilogue-Net [44] except the 'surprise' class. As both models use some variation of the attention mechanism applied to pairs of modalities, the similar performance obtained shows that using attention as the fusion mechanism can lead to a robust performance.

Another interesting observation made is that the model using all of audio, video and text performs worse than using only audio and text. This limits the extent to which the same model architecture can be applied to the task of this project as the aim includes integrating the video modality into the final model. Nevertheless, the performance of the tri-modal approach may be influenced by factors other than the fusion strategy such as the feature extraction network used. Therefore, considering that both Multilogue-Net and TBJE achieve good performance using attention to fuse different modalities, it is worth exploring using attention on pairs of modalities as the fusion mechanism.

Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion

Xie et al. [46] presents a robust multi-modal emotion recognition architecture in conversations by combining the state of the art cross-modal transformer fusion [47] with another robust multi-modal fusion architecture called EmbraceNet [48]. The architecture takes audio, video and text input and uses recurrent networks and a large langualge model like GPT to extract feature representation vectors. Each pair of the input features is then passed through a cross-modal transformer block. For each modality, the two attention outputs (i.e. two pairs formed with the other two modalities) are multiplied together to form the final modality representation. EmbraceNet is used as the final robust fusion mechanism that outputs the emotion embedding vector that is used to make the final emotion predictions.

The idea behind using cross-modal transfomers is to capture the correlation and inter-modality connections between all input modalities, allowing the model to learn important characteristics. On the other hand, EmbraceNet is added to ensure that the feature embedding vectors can be robustly fused together and the model's performance will not be affected by occasional absence of data in certain modalities.

The model is trained and evaluated on MELD dataset using various classification metrics including accuracy and F1 score. The results demonstrate that the multi-modal model outperforms all uni-modal models. The final model achieves a small improvement from the previous state of the art performance and puts an additional emphasis on the robustness in the emotion recognition task, which is desirable.

Since using transformer as the fusion method in multi-modal emotion recognition has become popular in recent years, it is worth experimenting with this mechanism as it helps achieve the current state of the art performance. Another advantage of the architecture proposed is that it can be easily extended to include more inputs of different modalities. Adopting a similar architecture can help make the final model more scalable, especially making it easier to incorporate new modalities in the future. The main limitation of the proposed model is the model size. With the limited resources available for this project such as GPU memory, adjustments to the fusion mechanism might be required to not exceed the memory limit.

2.5 Ethical and Professional Considerations

For the context of this project – training an ML model that classifies human emotions, there are some wider social, ethical, legal and professional implications that should be considered.

Firstly, experimenting with different fusion strategies involves long training and testing times for large ML networks, especially when video inputs are used, which require large amount of computational resources. As the process takes place on a shared server, this may reduce the amount of resources available to other researchers. Furthermore, there may also be potential environmental issues due to the large amount of electricity needed to keep the server running. To mitigate this, code has been tested on sample data first before the running on the whole dataset. Moreover, the video inputs from the datasets are pre-processed and saved separately, so that the training time and evaluation time can be reduced.

Secondly, it is important to recognise that the final model may have potential biases when predicting emotions. As the learning process depends heavily on the training data, a varied representation of the population is needed to ensure good performance and generalisation of the model. For instance, people with different demographic backgrounds (e.g. different

culture or languages spoken) may express emotions differently. It is therefore important to choose a training set that contains diverse data, so that the model can learn meaningful features in multiple settings. As the most common multi-modal emotion datasets used are all in English, inevitably there are some biases in the training process of the ML models.

Thirdly, human trials are planned to fully evaluate the performance of the model, particularly when generalising to real world situations. This means that personal video and audio data of participants need to be recorded. To ensure the participants are aware of this, a consent form has been sent out to get their approval and informed consent before any data collection takes place. The personal data collected is also strictly used for the purpose of testing the ML model only. Furthermore, when handling sensitive data collected such as facial expressions, the data is anonymised as much as possible, and the participants can request the deletion of their data once the evaluation process finishes. In addition, under the COVID-19 situation, safety and comfort of participants will be prioritised. For instance, the participants are allowed to record videos remotely at home at their convenience.

Lastly, it should be noted that ethical approval from Imperial College Research Ethics Committee has been granted before the human trials take place. The Ethics Checklist and the ethical approval statement are attached in Appendix A. Any individual or corporation that wishes to use the emotion recognition model developed in this project in the future should seek for ethical approval from a professional or academic organisation.

Chapter 3

Dataset

3.1 Dataset Choice

Since the goal of the project is to integrate all the existing emotion recognition algorithms developed by previous students together, the ideal dataset used for training the model should supply data from all of audio, video, text and heart rate (HR) modalities. There exists services like Google Speech Recognition that are able to transcribe audio files and thus convert an audio-visual dataset to one with text data too. However, using such services introduces more uncertainty in the training process. For instance, if the transcript does not accurately reflect the entire speech, the model may not be able to extract meaningful features from the text. This could lead to the model failing to accurately predict the target emotion and it would be difficult to identify whether it is due to the model design or inherent inaccuracies in the training set.

The only dataset that satisfies this requirement is K-EmoCon [36]. However, K-EmoCon is too small in sample size and class imbalance is a clear limitation mentioned by its creator. This suggests that training a model on K-EmoCon will not produce satisfying results as the model does not have enough data to learn from. Therefore, in this project, the focus is reduced to combine the audio, video and text modalities as there are many more available dataset options to choose from for these modalities. The model will be designed with extensibility such that if new multi-modal datasets are created in the future, the model can be easily adapted to include the HR modality data as well.

Out of all datasets with data for all of audio, video and text modalities, CMU-MOSEI [32] has been selected for several reasons. Firstly, CMU-MOSEI provides data from video, audio and text modalities, and thus no additional conversion between modalities is needed. Furthermore, it is the largest multi-modal emotion recognition and sentiment analysis dataset to date. Since dataset imbalance is a common issue for most multi-modal emotion datasets, the large total number of samples means that CMU-MOSEI would contain the higher number of samples in less-represented classes. This allows the model to train on sufficient number of samples from all classes and thus be with greater validity.

Secondly, CMU-MOSEI consists of video clips extracted from online YouTube videos, which are not specifically scripted or acted to depict the target emotions as in many other datasets. This has the advantage that the data used during training would more closely represent real-life situations, which are the intended use cases of the final emotion recognition model. Training the model using such data can allow for a better generalisation and help the model

achieve better performance when used on real-life data.

3.2 Dataset Overview

Since the goal is to develop an end-to-end emotion recognition model that can be integrated with the VR platform, the raw CMU-MOSEI dataset, instead of its pre-extracted features, will be used. The dataset contains `mp4` video files, audio files corresponding to each video in `wav` format and the text transcripts of the audio files. As data in CMU-MOSEI are labelled on the utterance level, each video file may contain multiple labelled samples, which need to be separated. The way to separate them is to use the timestamps for each transcribed sentence that are provided in the transcript text file and load the corresponding audio and video sub-clips.

Each utterance sample in CMU-MOSEI has been labelled with a sentiment score and six intensity scores – one for each of the six emotions (happy, sad, angry, surprised, disgusted and fearful). Taking the emotion with the highest intensity as the true emotion for a sample, the distribution of the dataset per class is shown in Figure 3.1 and Table 3.1. It should be noted that any sample with no intensity scores at all (not labelled with any of the six emotions) is not used.

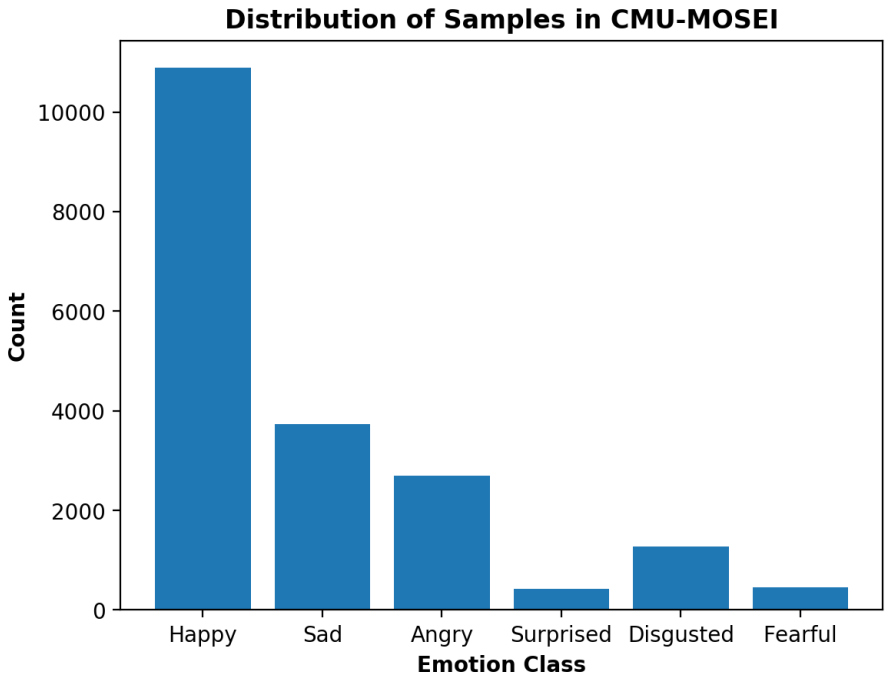


Figure 3.1: Distribution of samples per each emotion class in CMU-MOSEI

It is apparent that CMU-MOSEI is heavily imbalanced, with the 'happy' class making up more than 50% of the total samples and the smallest class ('surprised') only making up around 2%. The common ways to address class imbalance include assigning each class a weight when computing the loss and oversampling/downsampling the minority/majority classes during training. The effect of these approaches and the final selection will be explored more extensively in Section 4.3.2.

For the purpose of training and evaluating the model, CMU-MOSEI is split into three folds: train, validation and test folds. The train fold is the primary data that the model is trained on. The validation fold is used to tune the hyperparameters of the model and

Emotion	Count	Ratio (%)
Happy	10884	55.92
Sad	3731	19.17
Angry	2694	13.84
Surprised	428	2.20
Disgusted	1278	6.57
Fearful	450	2.31
Total	19465	/

Table 3.1: Number of samples per each emotion class in CMU-MOSEI

pick the desired model architecture. The test fold acts as the unseen data for evaluating the model's performance. The standard folds provided by the creator of CMU-MOSEI are used to obtain a random division of data and enable comparison with previous works as they have been done using these standard folds. Using this split, CMU-MOSEI is divided into train/validation/test folds roughly accounting for 70%/10%/20% of all the data. The per-class distribution of samples in each fold is shown in Table 3.2. It can be seen that the samples in all folds follow a similar class distribution, suggesting that they are suitable for training, validation and evaluating the final model.

Emotion	Train		Validation		Test	
	Count	Ratio (%)	Count	Ratio (%)	Count	Ratio (%)
Happy	7789	55.89	866	55.19	2229	56.29
Sad	2679	19.22	349	22.24	703	17.75
Angry	1911	13.71	188	11.98	595	15.03
Surprised	301	2.16	38	2.42	89	2.25
Disgusted	945	6.78	88	5.61	245	6.19
Fearful	311	2.23	40	2.55	99	2.50
Total	13936		1569		3960	

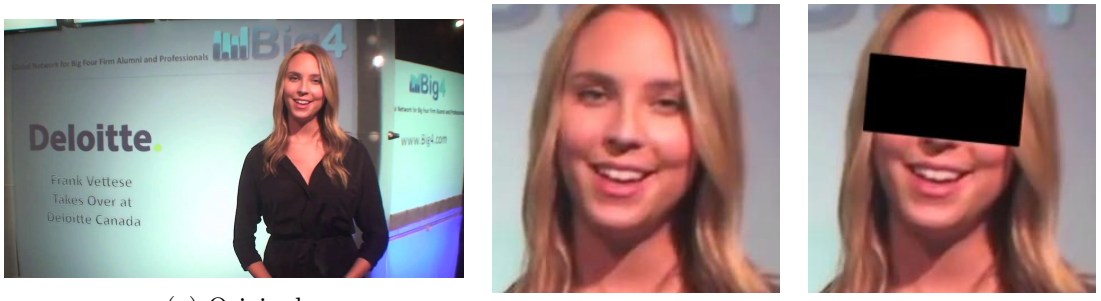
Table 3.2: Distribution of samples for each folds of CMU-MOSEI

3.3 Data Pre-Processing

As the raw data of CMU-MOSEI is used to train the model, they are pre-processed if needed as follows before input into the model. For the audio modality, all audio recordings are resampled to the same sampling rate. Furthermore, they are normalised using the mean μ and standard deviation σ computed across all samples in the training set by z-score normalisation $z = \frac{x-\mu}{\sigma}$.

For the video modality, the pre-processing procedure consists of facial detection, cropping and applying facial occlusion around the eye area. For each sample in the dataset, the frames of the clip are first obtained. Then, the face and facial landmarks in each frame are detected using a face alignment network implemented by Tzimiropoulos and Bulat [49]. Each frame is cropped centring around the detected face and saved as a separate image in a directory with the same name as the sample file name. In addition, the same cropped image with the eye area covered by a black rectangle is saved separately. An example frame is shown in Figure 3.2. As the user of the VR platform will be wearing a VR headset, this additional pre-processing is useful when adapting the model to a setting with the presence

of facial occlusions later. Since the face alignment network takes some time to run, for such a large training set, the extraction of frames for each video sample helps reduce training time as the video frames can be loaded very easily and efficiently during training.



(a) Original (b) Cropped (c) With Occlusion

Figure 3.2: Example frame before and after pre-processing

Chapter 4

Emotion Recognition Model

This chapter presents the steps taken to arrive at the final model (including its baseline) developed in this project. Firstly, the feature extraction network for each modality considered will be introduced. These networks will be adapted from existing models developed by previous students. The specific modifications to their models as part of this project will also be explained. Secondly, the fusion mechanism selected will be presented, along with a baseline simple fusion mechanism. Thirdly, experiments conducted to address issues occurred during the development process and the chosen training setup that optimises the model's performance will be outlined. Lastly, the process of adapting the final model developed in this project to work in a setting with facial occlusions around the eye area will be outlined.

4.1 Individual Modalities

Since previous students have implemented promising models using data from a subset of modalities, these models are adapted and used to extract features from different modalities. The resulting feature vectors are then fused together to make the final emotion prediction, which in this case is a single emotion label.

4.1.1 Audio

Simkanin [50] has developed a multi-modal emotion recognition model using audio and text data for multi-label emotion prediction. The model is trained on CMU-MOSEI and achieves competitive performance when compared to the state of the art models, and thus the corresponding network for each modality in this model will be used to extract audio and text features.

With regards to the model architecture, Simkanin based her model on one designed and implemented by Tavernor [51]. Based on their research, CNN is often used and performs well for the task of emotion recognition using audio data. Therefore, for the audio modality, a CNN model previously designed by Rizos et al. [52] is implemented. In addition, a bidirectional LSTM (BiLSTM) model is used on top of the CNN model to further process the features extracted and capture data dependencies in both directions. The detailed network architecture for the audio modality is shown in Figure 4.1.

4.1.2 Text

With more powerful language models being developed in recent years in the field of natural language processing, more useful and complex features from text data can be extracted.

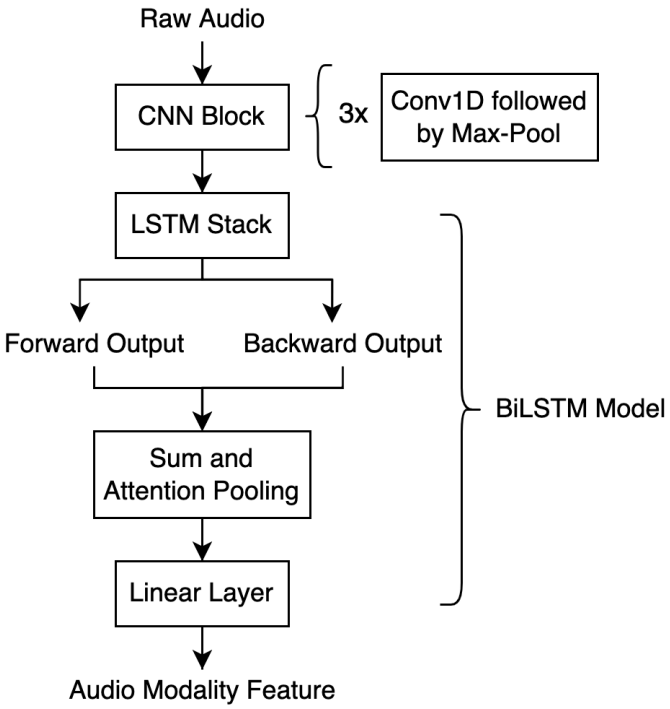


Figure 4.1: Model architecture of the audio model used

In particular, Tavernor and Simkanin have both established that fine-tuning a pre-trained BERT-based model is the most suitable for the task of emotion recognition using text data. Using text data from IEMOCAP [31], Tavernor has tested various variation of BERT ranging from BERT-base, BERT-large, RoBERTa to DistilBERT. He concluded that BERT-large achieves the best performance before the model starts to overfit to training data. Therefore, a pre-trained BERT-large model and its tokenizer are used to process and extract features from input text data. The architecture of this text modality model is shown in Figure 4.2.

The modification made to the existing text model in this project is to freeze the embedding layer of the BERT-large model. Since BERT-large is pre-trained on a large dataset, the embedding layer is generalised from a large vocabulary. Freezing the embedding layer can decrease the likelihood of the model overfitting to CMU-MOSEI text data. Furthermore, freezing the embedding layer also reduces the number of parameters, which is useful to reduce GPU memory consumption.

4.1.3 Video

Gotsman [53] has developed a novel model for facial emotion recognition, specifically with the presence of facial occlusion. In particular, the model is trained with images with the eye section blocked, representing the user's face when wearing a VR headset (as shown in Figure 3.2). In terms of its architecture, the model is adapted from a state of the art model EmoFAN [1] that predicts both the categorical emotion class and the continuous valence/arousal values. As shown in Figure 4.3, the model mainly consists of two hourglass-shaped convolutional blocks for feature extraction and facial landmark estimation, as well as a few final convolutional blocks for the joint prediction task. Since the goal for this project is to first develop a multi-modal emotion recognition model on unoccluded faces, the original EmoFAN model (without the final prediction layer) will be used to extract features from the video modality.

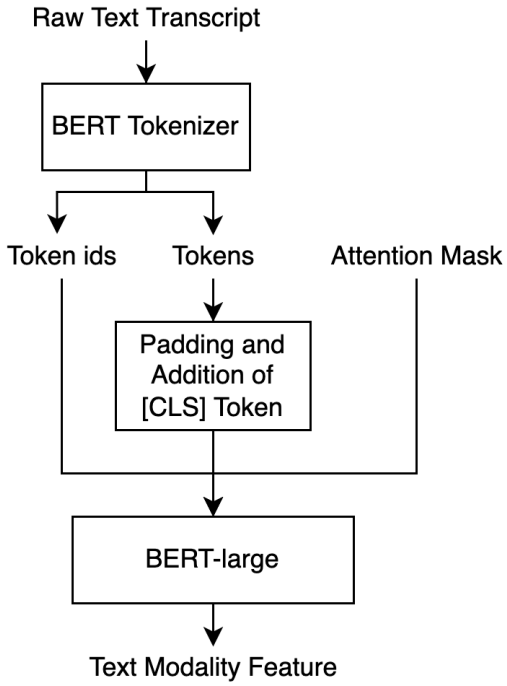


Figure 4.2: Model architecture of the text model used

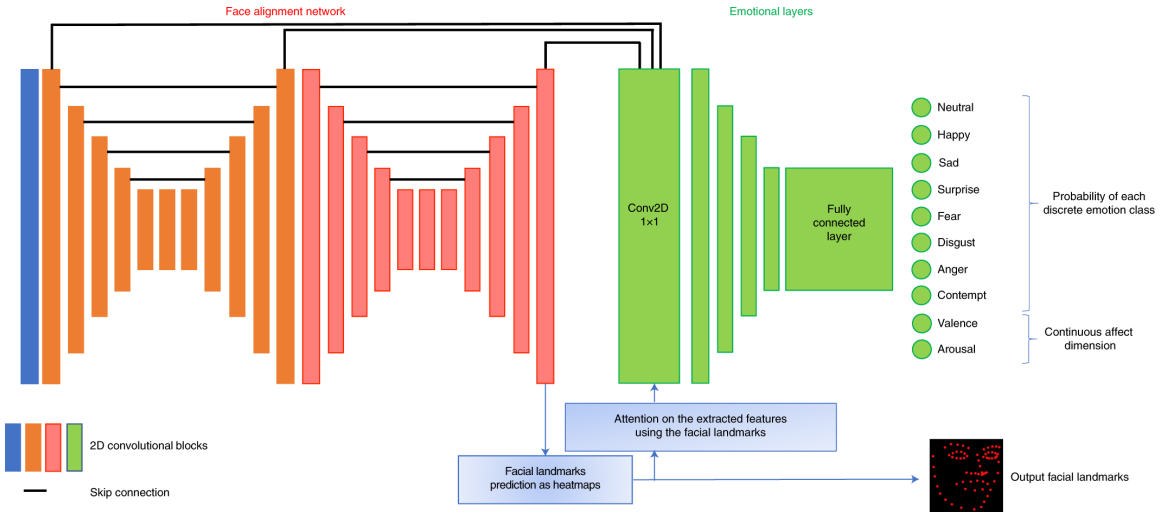


Figure 4.3: Model architecture of EmoFAN (the model used for video modality) [1]

Gotsman trained this model on AffectNet [54]. AffectNet is a dataset that consists of individual facial image, each with a discrete emotion label out of eight emotion classes and continuous annotated valence/arousal intensities. However, in this project, the training set selected is CMU-MOSEI. Samples in CMU-MOSEI are of video format, which can be seen as a batch of consecutive images. Moreover, only six basic emotions are considered for the discrete label and only a single sentiment value is annotated for each sample. Therefore, to utilise the model as part of the final model, the parameter controlling the output size is adjusted accordingly. Additionally, the model is modified such that a suitable way of handling multiple frames for one input is incorporated.

A simple option for handling multiple frames is to average all the frames in the video sample and input this single combined image into the model. However, the average number of frames across all samples in CMU-MOSEI is 699. Averaging this many frames will likely lead to a blurry frame that fails to capture the different facial expressions of different emotions. For instance, as shown in Figure 4.4, the average frame for the 'happy' sample is quite blurry and the person's facial features cannot be seen clearly. Furthermore, the average frames for the 'sad' sample, the 'angry' sample and the 'disgusted' sample are quite similar.

(a) Happy (b) Sad (c) Angry (d) Surprised (e) Disgusted (f) Fearful

Figure 4.4: Example averaged frames for each emotion class

Therefore, the preferred approach would be to select key frames as the input to the model – a fixed number of frames needs to be selected because the input feature size of the model needs to be constant. The number of key frames selected in this work is set to 100 to avoid exceeding the GPU memory limit. Two ways of selecting the key frames have been tested: (1) taking the consecutive centre frames of a video and (2) sampling frames uniformly across the entire video. Based on the performance on the validation set, the model that uses the centre frames slightly outperforms the uniform sampling one. A possible explanation for this is that since the video samples are quite long, selecting every seventh frame on average may cause the final batch of frames to lose track of the movement in facial landmarks, which may otherwise help distinguish between the different emotions. To enable the model to process multiple images as input, the first 2D convolutional layer for extracting shallow features (depicted as the blue block in Figure 4.3) is replaced by a 3D one. This modification is based on the finding that many previous work on multi-modal emotion recognition have also used various complex 3D-CNN to extract features from the video modality, possibly due to 3D-CNN being able to capture spatial dependencies from video data.

4.2 Fusion Mechanism

With the features from each modality extracted using the individual modality model, the key component of the final model is the fusion mechanism to combine the three feature vectors. To investigate the effect of using different fusion mechanisms, a simple model will be implemented first. The performance of this model is served as the baseline for analysing the performance of the final model that employs a much more complex fusion mechanism.

4.2.1 Simple Fusion

The baseline fusion mechanism is a late fusion mechanism that involves simple concatenation of features from all modalities. The reason for choosing this as the baseline is that most previous works investigating different fusion mechanisms have also established their baseline performance using a simple concatenation. The model designed by Tavernor [51] already combines the audio and text modality feature vectors together. The way he does this is a concatenation followed by applying self attention and attention pooling to the concatenated vector. On top of this, the baseline fusion used in this project concatenates the video feature vector at the end of the combined audio-text vector. This baseline fusion mechanism is a type of late fusion as the feature vectors are combined right before the final linear layer used to make the emotion prediction. Figure 4.5 shows the baseline fusion architecture.

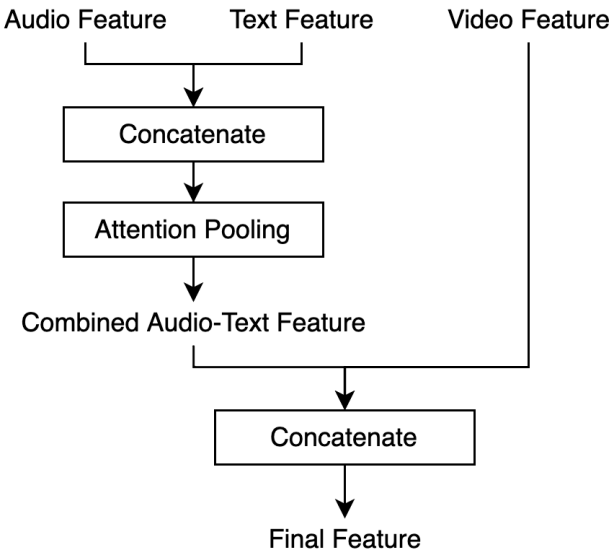


Figure 4.5: Architecture of simple fusion mechanism

4.2.2 Complex Fusion

Based on the background research conducted, attention and transformer-based fusion has attracted the attention of many researchers due to the superior improvements in model performance. Therefore, a similar approach to that proposed by Xie et al. [46] is adapted and used for the final model. This complex fusion mechanism can be split into two parts. The first part is the utilisation of cross-modal transformers [47] between pairs of modalities. The design of a cross-modal transformer is very similar to that of a traditional one, except the queries come from the projection of one modality input and the keys and values come from the projection of the other modality input. This also means that the cross-modal transformer output for the same pair of modalities is different depending on which modality is used to compute the queries. Therefore, to fully capture the inter-modality dependencies, two cross-modal transformers are needed per one pair of modalities.

In a typical transformer, multi-head attention is used, where multiple queries, keys and values are generated by different projections of the input. However, due to the large model size from the different feature extraction networks and the GPU memory limit, single-head attention is used here. To demonstrate how the cross-modal transformer between two modalities work, take passing the information from the video modality V to the text modality T as an example and denote the input feature vectors as X_V and X_T accordingly.

The query, keys and values are computed by projection of the input:

$$Q_V = X_V W_{Q_V} \in \mathbb{R}^{l_V \times d_h}$$
$$K_T = X_T W_{K_T} \in \mathbb{R}^{l_V \times d_h}$$
$$V_T = X_T W_{V_T} \in \mathbb{R}^{l_V \times d_h},$$

where d_h is the hidden size of the transformer and l is input length. The single-head attention output is then computed as:

$$O_{\text{attention}} = \text{softmax}\left(\frac{Q_V K_T^T}{\sqrt{d_h}}\right) V_T.$$

The rest of the cross-modal transformer architecture is identical to that of a typical transformer. A residual connection is first added and followed by a layer normalisation:

$$X = \text{LayerNorm}\left(O_{\text{attention}} + Q_V\right).$$

The resulting vector is passed through a small two-layer feed-forward network, which is then followed by another residual connection and layer normalisation, giving the final representation vector of passing information from modality V to T.

$$Z_1 = \text{Linear}\left(X\right)$$
$$A = \text{ReLU}\left(Z\right)$$
$$Z_2 = \text{Linear}\left(A\right)$$
$$O_{VT} = \text{LayerNorm}\left(Z_2 + X\right)$$

The final representation vector of an individual modality is computed by the Hadamard product of the two cross-modality vectors, which are obtained by passing the information from the other two modalities into this modality. Taking the video modality V as an example, the final feature vector X_V is obtained by:

$$X_V = O_{TV} \odot O_{AV}.$$

The second part of the fusion is to combine the newly obtained feature vectors for each modality. EmbraceNet [48] is used to combine the three feature vectors in a robust way. It consists of two main components, first of which are the docking layers that consider the different characteristics and correlations between the input modalities. The second part is the embracement layer, where a multinomial distribution is used to select parts of the feature vector (output from the docker layers) for each modality. This random selection also helps regularise the model and prevents it from overly relying on a specific modality. Since the author has provided a PyTorch implementation of EmbraceNet, it is directly used in the model.

The complete complex fusion architecture is shown in Figure 4.6.

4.3 Training and Experiments

4.3.1 Training Setup

The setup for training the model is similar to that in Simkanin's work [50], and the main challenge to overcome is the GPU out-of-memory issue. The model has a large number of

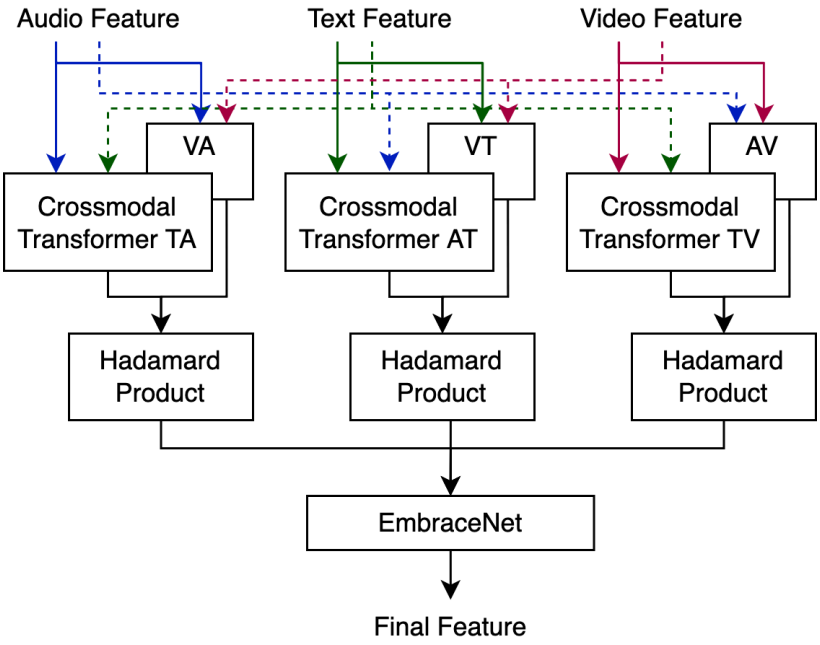


Figure 4.6: Architecture of complex fusion mechanism

parameters due to the inclusion of BERT-large as well as the addition of EmoFAN and the complex fusion mechanism. Furthermore, video frames need to be loaded for each sample, which takes much more space than only loading audio and text data. As a result, a maximum of 2 samples can be loaded and passed through the model at once. In order to match the desired update batch size, the losses from each small batch of 2 samples are backpropagated through the model but the model parameters are only updated when enough samples have been processed. Any sample that is too big to fit into the available GPU memory will be skipped and not counted into the number of samples already processed.

With regards to the loss function, the problem concerned in this project is a multi-class classification task, meaning that the model predicts one emotion label for each sample. This problem setup is selected as the VR platform agent relies on a single emotion predicted to give advice to the user, which is the intended use case of the model. Therefore, cross entropy is used as the loss function. As jointly predicting discrete emotion labels and continuous sentiment values is often shown to further improve the model's performance [1,51], its effect on performance has been tested. A separate EmoFAN model has been trained using a loss function that combines various continuous losses from sentiment prediction with the categorical cross entropy loss. The loss function used in this experiment is the same as in EmoFAN training, referred to as the 'shake-shake loss' [1]:

$$L\left(Y,\hat{Y}\right) = L_{\text{CE}}\left(Y,\hat{Y}\right) + \frac{\alpha}{\alpha+\beta+\gamma}L_{\text{RMSE}}\left(Y,\hat{Y}\right) + \frac{\beta}{\alpha+\beta+\gamma}L_{\text{PCC}}\left(Y,\hat{Y}\right) + \frac{\gamma}{\alpha+\beta+\gamma}L_{\text{CCC}}\left(Y,\hat{Y}\right),$$

where CE is cross entropy, RMSE is root mean square error, PCC is pearson correlation coefficient and CCC is concordance correlation coefficient between predicted value $\hat{Y}$ and ground truth Y. The constants α, β and γ are sampled randomly from the standard uniform distribution. They regularise the total loss, so that it does not penalise the model too heavily from specific continuous losses. The results of training EmoFAN with this summed loss function and cross entropy only are shown in Table 4.1.

Loss Function	Accuracy	Macro-F1
Discrete Only	**0.2779**	**0.1777**
Continuous and Discrete	0.0787	0.0777

Table 4.1: Classification results on validation set for different loss function tested

It is clear that for EmoFAN trained on CMU-MOSEI, using only discrete loss (cross entropy) yields a much better performance than including the continuous loss. Therefore, cross entropy is chosen as the loss function during training.

4.3.2 Addressing Class Imbalance

As discussed in Section 3.2, class imbalance is an apparent issue and it has negatively impacted the model's performance on the validation set. This is observed by inspecting the confusion matrix and the predictions made per emotion classes. For instance, in the extreme case, the model predicts all the samples in the validation set to be of the largest class ('happy' class). This is clearly an issue as the model will always output the same label regardless of the input, making it completely impractical to use for any emotion recognition task. To mitigate this issue, a common approach is to assign weights to the emotion classes based on their sizes – i.e. assign higher weights for less-represented classes and penalise the model more for making wrong predictions for these classes. Each class is assigned a weight proportional to $\frac{\text{total number of samples}}{\text{size of class}}$ and weighted cross entropy is computed as the loss.

The predictions made on the validation set have shown that using a weighted loss function has helped mitigate the model making highly skewed predictions. However, the model still completely ignores the two least-represented classes when making predictions. To further reduce the impact of class imbalance, a new experiment with random oversampling the minority classes and undersampling the majority classes has been run. The oversampling and undersampling is achieved by assigning each sample a weight according to its class label: sample weight $= \frac{1}{\text{class size}}$. Using PyTorch's weighted random sampler, it is expected that the model would see roughly equal number of samples from each class in one training epoch, where a sample from a less-represented class may be sampled repeatedly. In addition, to reduce model overfitting, L2 regularisation is also introduced in the loss computation. The validation results demonstrate that the sampling approach does result in the model making more predictions in the less-represented classes.

Although this is a desirable behaviour, the model's performance on the majority classes has dropped as a result of more samples being misclassified into the minority classes. A hypothesis for this behaviour is that the model keeps seeing the repeated samples from the least-represented classes and not enough unique samples from the more-represented classes, which cause it to overfit to the repeated samples. To check the validity of this hypothesis, the sample size of the weighted sampler has been doubled such that twice many samples from each class would be used during one epoch of training. However, this had no significant effect on the performance of the model and even made it worse slightly at the cost of increased training time. Therefore, this change has been removed and the original weighted random sampler is kept.

4.3.3 Hyperparameter Tuning

With the training process successfully implemented, the model's performance is optimised by tuning a few of its hyperparameters using the validation results as a reference. Ta-

ble 4.2 shows the macro-averaged F1 score obtained on the validation set for models with different parameter configurations. Macro-F1 is consulted here as it considers the model's performance across all classes.

Model	Macro-F1
A	0.2660
B	0.1537
C	0.2359
D	0.1462
E	0.2048
F	0.1093

Table 4.2: Performance of model with different hyperparameter values. The parameter specification for each model is detailed in Appendix C.

Since learning rate and batch size can often influence the model learning process greatly, the values for these two hyperparameters are tuned first. Theoretically, a large batch size allows for a higher learning rate and vice versa. Nevertheless, it has been observed during tuning that even when the batch size is set to 256, using a learning rate with magnitude higher than 10^{-5} results in the model ignoring the two smallest classes. This is shown through the confusion matrix obtained and thus not included in Table 4.2. Therefore, the learning rate is kept at 0.00001 to allow effective training. With regards to the batch size, since the learning rate is quite small, the model has been trained with batch size of 16 (Model B), 32 (Model A) and 64 (Model C) respectively for a few epochs and a batch size of 32 has shown the best performance.

Another hyperparameter to tune is the feature size in hidden layers of the feature extraction components of the model. For audio and text feature extraction, the model has been trained with the hidden size parameter set to 256 (Model D), 512 (Model E) and 1024 (Model A) respectively. From Table 4.2, the performance of the model is the worst in the 256 case (Model D). Since the macro-F1 increases by 0.06 when the hidden size increase from 512 (Model E) to 1024 (Model A), the hidden size is set to 1024.

The last hyperparameter explored is the embracement layer size in EmbraceNet at the end of the fusion. The initial embracement size is set as 256 (Model A), which is the same as the smallest feature size from the three modality inputs. When increasing the embracement size to 512 (Model F), the validation macro-F1 dropped significantly to 0.1093, suggesting that a embracement size too big negatively impacts the model's performance. Therefore, the embracement size is kept at 256.

4.3.4 Additional Tuning

Since the initial training of the video model has produced unsatisfactory results, additional experiments have been conducted to improve its performance. For instance, more 2D convolutional layers have been converted to 3D ones, with the aim of enabling the model to capture spatial and temporal dependencies between frames more effectively. The performance of the model is evaluated on the validation set and shown in Table 4.3.

It is apparent that introducing more 3D convolutional layers in the model does not improve its performance. A possible explanation for this is that having more 3D convolutional layers makes the model much more complex. With the increasing capacity, the model would overfit to training data, which then leads to lower performance on the validation set. Therefore,

Conv3D Layers	Accuracy	Macro-F1
Single	**0.2779**	**0.1777**
Multiple	0.2250	0.1548

Table 4.3: Experiment results when more 2D convolutional layers of the video model is converted to 3D

to keep the model simple and avoid overfitting, only the first layer of the video model is converted to a 3D convolutional layer.

4.4 Adaptation to With Facial Occlusion

When adapting the model to work with facial occlusions, only the video modality is affected. As discussed, Gotsman [53] experimented with various modifications made to EmoFAN when adapting it to work in a setting with the presence of facial occlusions. The final change that led to the best performance on occluded data was adding a dropout layer before the final prediction layer. The same change is introduced to adapt the model developed in this project to work with facial occlusions – a dropout layer is added before the final linear layer that outputs the logit for each emotion.

Chapter 5

Evaluation

To evaluate the performance of a classification model, the most commonly used metric is accuracy, defined as the ratio between the number of correctly predicted samples and the total number of samples. However, accuracy is often not reliable, especially for this project as accuracy does not take into account dataset imbalance. The main drawback of using accuracy as a metric is that accuracy is heavily influenced by the majority classes, making it unreliable and even misleading in some situations. For example, before introducing procedures to address class imbalance in Section 4.3.2, the complex fusion model simply predicts every sample to be from the same class. As that class contains the largest number of samples, the validation accuracy achieved was 0.58, which can be considered very descent, whereas the model has not actually learnt anything useful.

Therefore, an alternative accuracy metric – weighted accuracy introduced by Tong et al. [55] – has been chosen. Weighted accuracy (WA) weighs TP and TN according to the size of the positive and negative class and is computed by

$$\frac{1}{2N}\left(TP \times \frac{N}{P} + TN\right),$$

where T and P refer to the size of the positive and negative class respectively, and TP and TN refer to the number of correctly predicted positive and negative samples.

Other than accuracy, metrics like F1 score is useful. As F1 score is the harmonic mean of precision and recall, it considers both and is more informative compared to accuracy, which is only based on recall. In particular, macro-averaged F1 score (macro-F1) is chosen as it gives equal importance to each class, whereas micro-F1 is equivalent to accuracy in multi-class classification and thus not reliable.

In addition, per-class weighted accuracy and F1 as well as the confusion matrix can be referenced to fully understand how the model is performing. This will allow for a more in-depth analysis on the model's performance on specific classes, which can lead to meaningful explorations on changing the network and training pipeline to account for possible limitations.

5.1 Dataset Evaluation

The performance of the models developed in this project will first be evaluated on an existing dataset – the CMU-MOSEI test set.

5.1.1 All Emotions

The first section of Table 5.1 shows the overall metrics evaluated for the baseline fusion model and the final complex fusion model. It is shown that the complex fusion mechanism without oversampling outperforms the baseline simple fusion in all metrics. However, once oversampling from minority classes is added, the accuracy and macro-F1 decrease and fall below those for the baseline model. To investigate the possible reason for this decrease in performance, the per-class metrics results are computed and shown in Table 5.2.

	Accuracy	**Macro-F1**	**Average WA**
Baseline	0.524	0.293	0.593
Complex Fusion	**0.557**	**0.294**	**0.603**
Complex Fusion (With Oversampling)	0.406	0.266	0.600
Gotsman [53]	0.301	0.178	0.530
Simkanin [50]	0.542	**0.308**	0.598

Table 5.1: Overall metrics on CMU-MOSEI test set

	Happy		**Sad**		**Angry**		**Surprised**		**Disgusted**		**Fearful**	
	WA	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**
Baseline	0.71	0.73	**0.62**	**0.37**	0.55	0.21	0.54	0.10	0.62	0.28	0.52	0.07
Complex Fusion	**0.72**	**0.75**	0.61	0.36	**0.59**	**0.30**	0.50	0.00	0.70	**0.36**	0.50	0.00
Complex Fusion (With Oversampling)	0.66	0.60	0.54	0.26	0.54	0.21	**0.58**	**0.13**	**0.74**	0.31	**0.54**	**0.08**
Gotsman [53]	0.55	0.49	0.52	0.13	0.54	0.22	0.50	0.03	0.54	0.13	0.53	0.06
Simkanin [50]	0.70	0.74	0.59	0.32	0.58	0.28	0.54	0.12	0.66	0.32	0.52	0.07

Table 5.2: Per-class metrics on CMU-MOSEI test set

As shown in Table 5.2, the complex fusion without oversampling model completely ignores 'surprised' and 'fearful' (as mentioned in Section 4.3.2). When the per-class confusion matrices are consulted, the model makes zero prediction in any of these two classes, leading to a F1 score of 0. Furthermore, it can be observed that oversampling improves the model's performance on less-represented classes at the cost of worsening its performance on the majority classes. These observations suggest that the overall metrics may have been misleading as they are influenced by the class imbalance present in the test set.

Another interesting observation is that macro-F1 has also dropped although oversampling aims to balance the model's performance across all classes. To better understand the model's performance, the confusion matrix for the baseline model and the complex fusion model with oversampling are shown in Table 5.3. From the confusion matrix, the model with oversampling makes more balanced predictions across all classes, especially making more predictions in the minority classes. Although the number of correctly predicted samples in the minority classes increases, the model is also misclassifying many samples from the other classes to be in these minority classes. The increase in false positive rate in minority classes and the increase in false negative rate in the majority classes lead to a low precision and low recall respectively. This would limit the increase in F1 score or even decrease the F1 score, leading to a lower macro-F1.

This observation highlights a major limitation of the complex fusion model as the model

appears to easily overfit to the training data. Without oversampling, the model overfits its predictions to the majority classes and completely ignores the much smaller classes. On the other hand, with oversampling, the model overfits its predictions to the minority classes that it keeps seeing as a result of oversampling. Therefore, more measures of reducing overfitting such as more regularisation in the network could have been useful. Furthermore, the decrease in the model's performance when oversampling is added to training suggests that the model may not have learnt well distinct and useful features from individual classes. As it mixes samples between different classes quite easily, the feature extraction components of the model may be ineffective.

		Predicted					
		Happy	Sad	Angry	Surprised	Disgusted	Fearful
Actual	Happy	1541	447	84	53	56	48
	Sad	212	357	44	12	59	19
	Angry	156	247	89	20	75	8
	Surprised	34	38	1	10	6	0
	Disgusted	32	112	22	4	73	2
	Fearful	47	37	5	4	0	6

(a) Baseline

		Predicted					
		Happy	Sad	Angry	Surprised	Disgusted	Fearful
Actual	Happy	1088	492	264	79	202	104
	Sad	137	208	108	29	187	34
	Angry	86	121	120	33	225	10
	Surprised	13	21	13	17	22	3
	Disgusted	15	35	22	11	162	0
	Fearful	37	28	8	5	10	11

(b) Complex Fusion With Oversampling

Table 5.3: Confusion matrix for baseline model and complex fusion with oversampling model

5.1.2 Comparison With Previous Work

It is also worth comparing the performance of the final model developed in this project with those developed in previous works.

As shown in Table 5.4, the model performs much worse compared to Multilogue-Net [44]. Since the fusion mechanism used in Multilogue-Net is pairwise attention, which is similar to the complex fusion mechanism in this project, a possible reason for the big difference in model performance is the different feature extraction networks used. Multilogue-Net is not an end-to-end model as it uses pre-extracted features provided by CMU-MOSEI SDK, where multiple complex networks have been run to extract different features. In contrast, the end-to-end model developed in this project could not utilise complex feature extraction networks due to the GPU memory limit and the consideration that the integration of the model with the VR platform requires the model not to be too big.

When compared to Graph-MFN proposed by the creator of CMU-MOSEI [32], the model developed in this project outperforms Graph-MFN in weighted accuracy on 'happy', 'sur-

prised' and 'disgusted'. However, the per-class F1 score is much lower, indicating that the model developed in this project still performs worse than Graph-MFN overall. Similarly, Graph-MFN runs on pre-extracted features and requires constructing a graph between the modality features before the fusion can begin, which is quite different to the model developed in this project.

	Happy		**Sad**		**Angry**		**Surprised**		**Disgusted**		**Fearful**	
	WA	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**
Graph-MFN [32]	0.66	0.66	0.60	0.67	0.63	0.73	0.54	**0.86**	0.69	0.77	0.62	**0.90**
Multilogue-Net [44]	**0.70**	**0.68**	**0.76**	**0.75**	**0.83**	**0.81**	**0.87**	0.84	**0.90**	**0.87**	**0.90**	0.87
Complex Fusion (With Oversampling)	0.67	0.61	0.55	0.31	0.54	0.19	0.57	0.12	0.75	0.31	0.39	0.08

Table 5.4: Per-class metrics on CMU-MOSEI validation set in comparison with state of the art models

Since the model in this project is developed based on existing models developed by previous students, the final model's performance is also compared with that of their models. The second section in Table 5.1 and Table 5.2 shows the overall and per-class metrics computed on previous students' models after being adapted to multi-class classification using CMU-MOSEI.

Both the baseline and final model developed in this project outperforms Gotsman's video model [53] by a large margin, suggesting that a model on video modality only is not effective. In comparison with Simkanin's audio-text model [50], the complex fusion model slightly improves the performance in average weighted accuracy and achieves similar macro-F1. However, with oversampling, the complex fusion model achieves lower macro-F1 and similar average weighted accuracy. From the per-class metrics, this drop in performance is caused by the model trying to improve its performance on the smaller classes. For example, the complex fusion with oversampling achieves higher weighted accuracy in the three smallest classes.

The slight improvement in performance compared to the audio-text model may be explained by the inclusion of the video modality. Since the video-only model developed by Gotsman performs poorly on CMU-MOSEI, even after many rounds of experimentation and tuning, combining it with the audio-text model can only lead to very limited improvement of model performance. In particular, the complex fusion mechanism passes information from each modality to the other two for the purpose of cross-validating the most useful features. This means that if the video modality feature is not effective, combining it with the more effective audio and text features is not helpful and may even negatively impact the model's performance as demonstrated above.

5.1.3 Four Emotions

Since class imbalance has been a major challenge faced when training the model, it may be interesting to study the model's performance when the two smallest classes are removed from the prediction task. Table 5.5 and Table 5.6 detail the metrics computed when 'surprised' and 'fearful' are removed.

As shown, the model's performance improves significantly, which indicates that class imbalance has a big impact on the model training process. Similar to the case of six emotions, the complex fusion outperforms the baseline in all metrics overall and in most individual classes

	Accuracy	Macro-F1	Average WA
Baseline	0.598	0.434	0.633
Complex Fusion	**0.626**	**0.452**	**0.638**

Table 5.5: Overall metrics on CMU-MOSEI test set with the two least-represented classes removed

	Happy		Sad		Angry		Disgusted	
	WA	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**
Baseline	**0.72**	0.77	**0.62**	**0.38**	0.58	0.29	0.61	0.29
Complex Fusion	0.71	**0.79**	0.60	0.35	**0.61**	**0.34**	**0.63**	**0.32**

Table 5.6: Per-class metrics on CMU-MOSEI test set with the two least-represented classes removed

as well. The difference is that the oversampling approach is no longer needed in the four emotions setting, which again emphasises the challenge faced during training when significant class imbalance is present. The difference in models' performance demonstrates that complex fusion is more robust and more effective than the baseline simple fusion. Moreover, the final model demonstrates good performance in weighted accuracy, suggesting that it is effective when predicting samples from the four emotions. On the other hand, the relatively low F1 score in the minority classes suggests that class imbalance is still an issue and the predictions made are slightly skewed towards the largest class, which could be further investigated and improved in future works.

Since the model's performance on this subset of emotions is much more promising than on all six emotions, the four emotions setting will be the focus for further evaluation and analysis of the model's performance.

5.1.4 With Facial Occlusion

Other than the model that operates on the whole face, the performance of the model after being adapted to work with facial occlusions has also been evaluated. As outlined in Section 4.4, the adapted model is trained on CMU-MOSEI with facial occlusions applied to the input video frames with the same hyperparameter values.

As shown in Table 5.7 and Table 5.8, the complex fusion model achieves similar weighted accuracy to unoccluded video frames, but the F1 score is slightly lower. This is expected as humans often express emotions through eyes and eyebrow movements. When the eye section is blocked, the task of emotion recognition is much more difficult. The fact that the model is able to achieve similar performance to the unoccluded case shows that the adaptation strategy is quite effective and the final model developed can be used to predict the user's emotion in both settings (i.e. unoccluded and occluded scenarios).

5.2 Human Trial

A human trial has been conducted to better evaluate the model's performance in real-life settings. The trial has been reviewed by Imperial College Research Ethics Committee and no significant ethical concerns have been raised. The ethical approval is attached in Appendix A.

	Accuracy	Macro-F1	Average WA
Baseline	0.573	**0.456**	**0.655**
Complex Fusion	**0.585**	0.412	0.625

Table 5.7: Overall metrics on CMU-MOSEI test set with facial occlusions applied (four emotions)

	Happy		Sad		Angry		Disgusted	
	WA	F1	WA	F1	WA	F1	WA	F1
Baseline	**0.73**	0.74	**0.63**	**0.39**	**0.62**	**0.36**	**0.65**	**0.33**
Complex Fusion	**0.73**	**0.78**	0.58	0.32	0.59	0.30	0.60	0.25

Table 5.8: Per-class metrics on CMU-MOSEI test set with facial occlusions applied (four emotions)

The trial arrangement is outlined as follows:

- The goal of the trial is to record video and audio data of participants while they experience the six emotions present in CMU-MOSEI. These emotions are elicited from the participants as they watch a set of selected film clips. To ensure the health and safety of participants, target participants are normal population with no psychological disorders.

- Before recruiting the participants, 12 film clips (2 for each emotion) have been selected based on previous research and common triggers for eliciting the target emotions. The film clips and corresponding triggers are attached in Appendix B.

- The participants are asked to record their reactions while watching the videos as well as answering a few more descriptive questions immediately afterwards. The reason for this is to ensure enough text data can be collected. Although letting everyone record remotely means that the videos recorded will inevitably have some inconsistencies, this reflects the real-world situation that the model should be able to handle.

- Once the recordings have been received, they are processed to extract data for all modalities. Due to the limited number of participants recruited, some manual work had to be done. Specifically, to fully utilise the video samples, multiple clips are cropped out from individual video samples as people's reactions occur at various times while watching the same film clip. The length of the clips is kept short since when running the model real-time, the duration of the data sent to the model would be short. In addition, each reaction video used has been verified to represent the target emotion by checking whether the participant experienced that emotion after watching it.

The distribution of the human trial data per emotion is shown in Figure 5.1. The trial data is much more balanced than CMU-MOSEI, meaning that the evaluation results on the trial data will be influenced less by class imbalance in the test data. Therefore, the metrics computed are more informative and can aid a better understanding about the model's performance on all different emotions in real life.

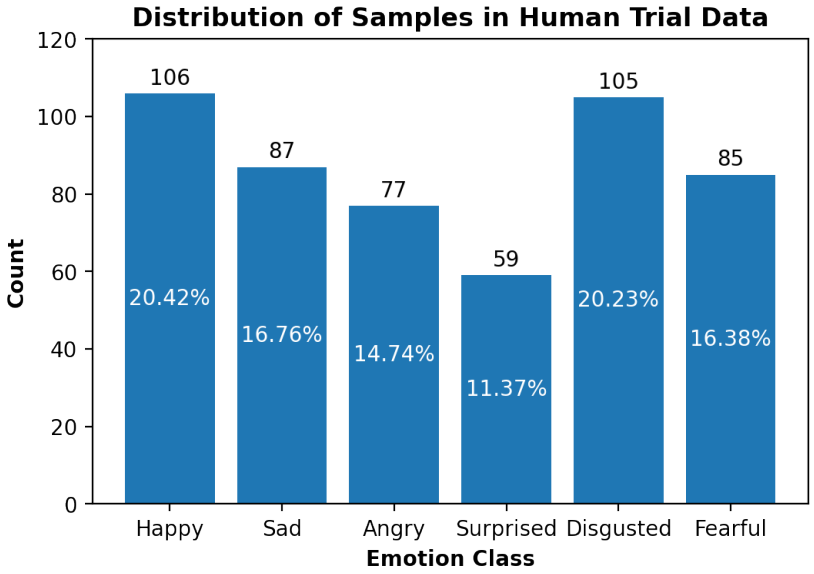


Figure 5.1: Distribution of trial data samples per each emotion class

5.2.1 Results

Table 5.9 and Table 5.10 show the trial results. It can be seen that the performance of the model drops in all metrics for both the baseline and the final model. The significant decrease in overall accuracy can be explained by the decrease in sample size. As there are much fewer samples and the test data is more balanced, a small number of misclassified samples could account for a large proportion of all the test data. As macro-F1 and average weighted accuracy are influenced less by the change in sample size, they are more meaningful.

The macro-F1 drops by around 6% for the baseline model and 10% for the final model, which is a big difference. From Table 5.10, the reason for this decrease appears to be the huge drop in F1 score on the 'happy' class. From the confusion matrix for this class, the final model makes 136 false positive predictions, which is a big number considering the small overall test data size. This reveals that the data imbalance in the training set has caused the model to make more predictions in the largest class. Since the CMU-MOSEI test set is also imbalanced, this behaviour results in good performance on the 'happy' class. Nevertheless, the trial data is more balanced and this behaviour results in a low precision on this class and thus a much lower F1 score. This again highlights that the model may have struggled to learn unique features from all classes during training because the training set is too imbalanced.

Another possible reason for the poor performance of the final model on the trial data is the validity of the trial data itself. As the audio recording of each sample is transcribed by the Google Speech Recognition API, the text transcript may be inaccurate. This is indeed the case as many samples have been manually inspected and the text transcripts do not reflect the actual audio recordings. As will be discussed later in Section 5.3, the text modality is the most impactful for the emotion prediction task; the poor quality of the trial text transcripts are likely to have impacted the model's performance negatively.

On the other hand, the model still achieves a decent weighted accuracy on average. This shows that the model developed in this project is fairly accurate on real-life samples, demonstrating that the model can be generalised to data with unseen distribution and patterns.

	Accuracy	Macro-F1	Average WA
Baseline	0.400	**0.372**	**0.605**
Complex Fusion	**0.408**	0.340	0.595

Table 5.9: Overall metrics on human trial data

	Happy		Sad		Angry		Disgusted	
	WA	F1	WA	F1	WA	F1	WA	F1
Baseline	0.67	0.51	**0.63**	**0.44**	**0.54**	0.24	**0.58**	**0.29**
Complex Fusion	**0.71**	**0.58**	0.57	0.32	**0.54**	**0.26**	0.56	0.21

Table 5.10: Per-class metrics on human trial data

Lastly, the key difference between the results on the trial data and on the CMU-MOSEI test set is that the baseline model can outperform the complex fusion model. For the 'sad' emotion, the baseline model consistently performs better as shown in both Table 5.6 and Table 5.10. With the trial data, the baseline model also achieves significantly higher F1 score. This observation links back to the overfitting and worse performance on more balanced test set discussion earlier. As the final model overfits to the imbalanced training data distribution, its ability to generalise is limited, which could lead to a decline in its usability.

5.2.2 With Facial Occlusion

Similarly to the CMU-MOSEI test set case, the performance on the trial data with facial occlusions applied has been evaluated. The overall metrics and per-class metrics are presented in Table 5.11 and Table 5.12 respectively. The macro-F1 and average weighted accuracy for the final complex fusion model decrease slightly as a result of the task becoming more difficult, but overall these two metrics are quite similar to the results shown in Table 5.9. The similar performance obtained (around 60% of average weighted accuracy) indicates that the model works decently even when facial occlusions are present, which is desirable as with minimal changes (i.e. adding a dropout layer) the model can work well in both occluded and unoccluded settings.

However, it is shown that the baseline model performs much better on occluded trial data, possibly because it is less likely to overfit to training data. This highlights limitation caused by the overfitting issue again: the usability of the whole VR platform could be negatively impacted if the model is integrated. Therefore, given more time other procedures to avoid overfitting or mitigate imbalance data distribution should be explored. These modifications to the training process would help improve the model's performance and make it more usable in the context of the VR platform setting.

	Accuracy	Macro-F1	Average WA
Baseline	**0.387**	**0.375**	**0.599**
Complex Fusion	0.352	0.335	0.574

Table 5.11: Overall metrics on human trial data with facial occlusions applied

	Happy		Sad		Angry		Disgusted	
	WA	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**
Baseline	**0.68**	**0.54**	**0.60**	**0.41**	**0.54**	**0.30**	**0.56**	**0.24**
Complex Fusion	**0.68**	**0.54**	0.56	0.38	0.51	0.25	0.54	0.17

Table 5.12: Per-class metrics on human trial data with facial occlusions applied

5.3 Ablation Studies and Other Insights

Since the final model with the robust fusion mechanism implemented in this project has been evaluated in various settings as discussed above, it is also worth investigating the effect of using different combinations of modalities. Table 5.13 shows the performance metrics computed for all different combinations of modalities. The per-class metrics have also been computed and shown in Table 5.14. They provide a reference to ensure that the high performance achieved by any model does not stem from the model completely ignoring the minority classes or making highly skewed predictions.

From the overall metrics, the final model developed in this project that utilises all of video, audio and text modality is the best-performing model. It has the highest accuracy and macro-F1, suggesting that the model achieves the best balance between recall and precision, thus outperforming all the other combination of modalities. In terms of the average WA, although the tri-modal model does not achieve the highest average WA, its performance is close to the best average WA obtained by the audio-text model. Furthermore, Table 5.14 shows that the final model performs the best (or close to the best) in all four emotion classes. Therefore, combining all metrics the final tri-modal case can still be considered the best combination of modalities.

	Accuracy	**Macro-F1**	**Average WA**
A + V+ T	**0.626**	**0.452**	0.638
A + V	0.575	0.236	0.515
A + T	0.563	0.434	**0.648**
V + T	0.586	0.431	0.638
A	0.577	0.250	0.525
V	0.299	0.269	0.555
T	0.606	0.418	0.538

Table 5.13: Overall metrics for different combination of modalities on CMU-MOSEI test set (A: audio, V: video, T: text)

Based on the metrics of the bi-modal and uni-modal models, the following important observations can be made:

- The audio-text model and the video-text model achieve the second highest performance on macro-F1, followed by the text-only model. All the other models have much lower macro-F1's.
- The bi-modal models with the text modality also achieves the highest average WA.
- The text-only model achieves the best performance when a single modality is used.

	Happy		Sad		Angry		Disgusted	
	WA	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**
A+V+T	0.71	**0.79**	0.60	0.35	**0.61**	**0.34**	0.63	**0.32**
A + V	0.53	0.73	0.53	0.20	0.50	0.01	0.50	0.00
A + T	**0.73**	0.74	**0.63**	**0.40**	0.58	0.29	**0.65**	0.31
V + T	0.72	0.76	**0.63**	0.39	0.57	0.26	0.63	0.31
A	0.55	0.74	0.51	0.04	0.54	0.22	0.50	0.00
V	0.57	0.40	0.53	0.26	0.56	0.27	0.56	0.15
T	0.70	0.78	0.46	0.34	0.53	0.28	0.46	0.27

Table 5.14: Per-class metrics for different combination of modalities on CMU-MOSEI test set (A: audio, V: video, T: text)

It can be deduced from these observations that the text modality is the most dominant modality for the task of emotion recognition. Moreover, combining the text modality with other modalities can further improve performance, but without text, the model would struggle with the classification task.

To further demonstrate that the text modality guides the model predictions, the trial data is divided into two sets: one set where text transcripts have been generated and the other with empty transcripts. One possible reason for empty transcripts is that some participants only show reactions through facial expressions instead of textual comments. Table 5.15 and Table 5.16 show the metrics computed.

The final model developed in this project (using all modalities) performs significantly better on samples with text detected across all metrics. On the per-class level, the model fails to recognise samples from two emotions when text input is empty. Therefore, the importance of the text modality is further emphasised by the superior performance the model obtains on real-life samples.

	Accuracy	**Macro-F1**	**Average WA**
With Text Detected	**0.495**	**0.467**	**0.673**
No Text Detected	0.333	0.159	0.512

Table 5.15: Overall metrics on human trial data separated by with and without text detected when generating the transcript

	Happy		Sad		Angry		Disgusted	
	WA	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**
With Text Detected	**0.85**	**0.71**	**0.65**	**0.47**	**0.61**	**0.38**	**0.59**	**0.31**
No Text Detected	0.54	0.51	0.51	0.13	0.50	0.00	0.50	0.00

Table 5.16: Per-class metrics on human trial data separated by with and without text detected when generating the transcript

Other than the different combinations of modalities, the ethnicity distribution of the trial participants can be obtained. Unfortunately, only a small number of participants have been successfully recruited and the ethnicity groups could not be controlled. The final ethnicity

distribution is shown in Table 5.17. Investigating the model's performance on different ethnicity groups can provide some insights into the potential bias that the model has.

	Participants		Samples	
	Count	**Ratio (%)**	**Count**	**Ratio(%)**
White	7	70%	119	31.73%
Asian	3	30%	256	68.27%
Total	10		375	

Table 5.17: Distribution of human trial data by ethnicity

As shown in Table 5.18 and Table 5.19, the final model performs worse on the Asian group of samples, especially when measured using macro-F1. Although this can be due to the model being biased towards non-Asian samples, the root of this behaviour may be the availability of the text input. After filtering the samples in each ethnicity group, 28 samples in the White samples and 131 samples in the Asian samples do not have text detected when the audio is transcribed. These samples make up around 23.5% and 51.2% of the samples in each group respectively. Since there is a larger proportion of samples without text in the Asian group, the worse performance is expected. Additionally, the average weighted accuracy achieved on both ethnicity groups do not differ by a large margin. Therefore, although the model may be slightly biased towards the White samples, it generalises decently to different ethnicity groups and the difference in performance observed can be explained by the lack of text data.

	Accuracy	**Macro-F1**	**Average WA**
White	**0.479**	**0.416**	**0.636**
Asian	0.406	0.349	0.603

Table 5.18: Overall metrics on human trial data by ethnicity

	Happy		**Sad**		**Angry**		**Disgusted**	
	WA	**F1**	**WA**	**F1**	**WA**	**F1**	**WA**	**F1**
White	**0.76**	**0.67**	0.59	0.36	**0.61**	**0.36**	**0.58**	**0.28**
Asian	0.69	0.54	**0.6**	**0.38**	0.57	0.3	0.54	0.17

Table 5.19: Per-class metrics on human trial data by ethnicity

5.4 Summary Discussion

In this chapter, the model developed in this project has been evaluated in various settings. On CMU-MOSEI, the final model with complex fusion outperforms the baseline, but the major limitation is that the model suffers from the imbalanced dataset and overfitting to training samples. With oversampling included during training to combat class imbalance, the final model performs worse and achieves similar performance as the baseline. When compared to previous work, the model performs worse than the state of the art model; however, the previous models are not end-to-end and have utilised much more complex feature extraction networks. This may have led to the difference in performance as the fusion mechanism adopted by their model and in this project are both based on attention between different pairs of modalities.

Since dataset imbalance poses a big challenge during training, the focus is shifted to the four emotions setting, where the two smallest classes ('surprised' and 'fearful') are removed. The final model with the complex fusion mechanism performs much better than the baseline in all metrics. This suggests that dataset imbalance in the training set has influenced the model's performance greatly. When testing the model on human trial data, the model can still achieve a decent average weighted accuracy, suggesting that it can generalise to unseen data from the real world. Furthermore, the model performs slightly worse when adapted to work with facial occlusion. This is expected as the task is much harder when the eye area is blocked. More importantly, the results demonstrate that the model architecture can be adapted to work with facial occlusions with minimal change, which is desirable.

Nevertheless, it has been observed that the final model's performance drops more than the baseline model when tested on real-life data. As the model makes more predictions on the more-represented classes as a result of training set imbalance and overfitting to the training samples, its performance is worse on a test set that is more balanced. Therefore, other approaches that address class imbalance and mitigate overfitting could be implemented to improve the model's performance.

Lastly, the text modality appears to be the most important for the task of emotion recognition. Including the text modality yields higher performance than relying on the audio and video modality. This can explain the fact that the final model developed in this project only makes marginal improvements from the existing audio-text model developed by Simkanin [50]. As the video modality performs very poorly on its own, combining it with the audio and text modality can only lead to very limited improvements. In addition, the decrease in model's performance when tested on human trial data can also be explained by the inaccuracies in the generated text transcripts. Since the model relies on the text modality input to make the final prediction, its performance may have been negatively impacted.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The main aim of this project is to develop an end-to-end multi-modal emotion recognition model that combines and enhances the existing models.

After adapting the existing models to extract features from corresponding modalities, a fusion mechanism based on the cross-modal transformer architecture and EmbraceNet has been implemented. Results have shown that this fusion mechanism performs better than using simple concatenation to combine features from different modalities. The final model also achieves descent results in a human trial conducted, demonstrating that the model developed is able to generalise to unseen data. Moreover, the adapted version of the model maintains its performance on video input with facial occlusions applied. For the basic emotions considered, the promising performance of the model on real-life data and occluded video input suggests that it can potentially be integrated with the bigger VR platform. In addition, it has been demonstrated that text is the most important modality in emotion recognition and the video modality performs the worst. Therefore, it is vital to ensure that enough text data is collected and passed to the model to obtain more accurate predictions.

The original aim of the project was to fuse all four modalities (i.e. audio, text, video and heart rate). However, this was not feasible within the duration of this project as there has not been a suitable training dataset that contains enough samples and provide data from all four modalities. Nevertheless, the fusion mechanism adopted in the final model is highly extensible; it has the potential to include multiple more modalities as long as training data is available. If one wishes to add a fourth modality into the model, assuming that the feature extraction network has been selected, they only need to add the pairwise cross-modal transformers for new modality pairs. The changes required for EmbraceNet are also very simple. In addition to the existing feature sizes, the feature size of the fourth modality also needs to be passed to EmbraceNet. Once these changes are applied, the final model obtained can utilise input from four modalities, although the overall performance of the model is bounded by the effectiveness of feature extraction from the individual modalities.

Nonetheless, there are a few limitations of the model developed in this project:

- The biggest challenge is the class imbalance in the training set and the model overfitting to the training data. When all emotions are considered, the model overfits to the majority classes, failing to capture information from the minority classes. Although the model's performance on a subset of emotions is much better, overfitting leads to

skewed predictions made and thus decreasing the performance when tested on real-life data.

- The model performs much worse compared to state of the art models. A possible explanation is that the model uses less complex feature extraction networks that are less effective. Compared to the existing audio-text model, very limited improvement has been achieved when all six emotions are included. This is possibly due to the inadequate performance of the video feature extraction model or wrong adaptions made when modifying EmoFAN to work with video input.

In addition, the data used for training and testing may have worsened the model's performance in the following ways:

- For each sample in CMU-MOSEI, a single emotion label is provided for the whole video, which on average has 699 frames, but the emotion may not be present at every frame. It is thus difficult to pick frames with high emotion intensity as there is no way to distinguish these frames based on the label. This could lead to inaccuracy in the training process.

- Emotion is highly subjective and often not mutual exclusive. For example, some participants have felt 'disgusted' and 'angry' when watching one of the clips that aim to elicit disgust. The 'fearful' clips also often include some jump-scare elements, causing some participants to feel 'surprised' at the same time. Therefore, providing a single label for each clip in the trial data makes it difficult for the model to distinguish between the different emotions.

- There have been inaccuracies when transcribing the trial data audio, leading to poor performance of the model as it relies heavily on the text input.

6.2 Future Work

With the limitations discussed, the following options may be explored to further understand the model's behaviour and improve its performance:

- Experiment with different feature extraction networks, especially for the video modality. As the video modality limits the possible performance improvements when combined with audio and text, a better performing video-only model can help boost the performance of the fused model.

- Enhance or implement additional measures that reduce overfitting during training. For instance, the weight decay parameter can be tuned to enforce stronger L2 regularisation on the model.

- Explore the model's performance on unaligned input data. Since the audio and text networks treat each input sample with input size of 1 and not its length after tokenization, the alignment of data for all modalities may not be strictly required. If the model can obtain decent performance on unaligned data, multiple datasets can be combined together to form a more balanced training set.

- Introduce more modalities to the model if a large enough multi-modal emotion dataset, including data from all input modalities, is made available.

References

[1] Toisoul A, Kossaifi J, Bulat A, Tzimiropoulos G, Pantic M. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nature Machine Intelligence. 2021;3:42-50. Available from: doi:10.1038/s42256-020-00280-0.

[2] World Health Organization. The WHO special initiative for mental health (2019-2023): universal health coverage for mental health. World Health Organization; 2019.

[3] Santomauro DF, Mantilla Herrera AM, Shadid J, Zheng P, Ashbaugh C, Pigott DM, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. The Lancet. 2021;398(10312):1700-12. Available from: doi:https://doi.org/10.1016/S0140-6736(21)02143-7.

[4] Cullen W, Gulati G, Kelly BD. Mental health in the COVID-19 pandemic. QJM: An International Journal of Medicine. 2020 03;113(5):311-2. Available from: doi:10.1093/qjmed/hcaa110.

[5] Edalat A. Self attachment: A holistic approach to computational psychiatry. In: Peter E, Bhattacharya B, Cochran A. (eds.) Computational Neurology and Psychiatry. vol. 6 of Springer Series in Bio-/Neuroinformatics. Springer, Cham; 2017. p. 273-314. Available from: doi:10.1007/978-3-319-49959-8_10.

[6] Polydorou N, Edalat A. An interactive VR platform with emotion recognition for self-attachment intervention. EAI Endorsed Transactions on Pervasive Health and Technology. 2021 9;7(29). Available from: doi:10.4108/eai.14-9-2021.170951.

[7] Keltner D, Sauter D, Tracy J, Cowen A. Emotional Expression: Advances in Basic Emotion Theory. Journal of nonverbal behavior. 2019;43(2):133-60. Available from: doi:10.1007/s10919-019-00293-3.

[8] Wu CH, Lin JC, Wei WL. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA Transactions on Signal and Information Processing. 2014;3:e12. Available from: doi:10.1017/ATSIP.2014.11.

[9] Bowlby J. Attachment and loss / Vol.1, Attachment. 2nd ed. London: Pimlico; 1997.

[10] Lyons-Ruth K. Attachment Relationships Among Children With Aggressive Behavior Problems: The Role of Disorganized Early Attachment Patterns. Journal of consulting and clinical psychology. 1996;64(1):64-73. Available from: doi:10.1037/0022-006X.64.1.64.

[11] Freeman D, Haselton P, Freeman J, Spanlang B, Kishore S, Albery E, et al. Auto-

mated psychological therapy using immersive virtual reality for treatment of fear of heights: a single-blind, parallel-group, randomised controlled trial. The Lancet Psychiatry. 2018;5(8):625-32. Available from: doi:10.1016/S2215-0366(18)30226-8.

[12] Park MJ, Kim DJ, Lee U, Na EJ, Jeon HJ. A Literature Overview of Virtual Reality (VR) in Treatment of Psychiatric Disorders: Recent Advances and Limitations. Frontiers in Psychiatry. 2019;10. Available from: doi:10.3389/fpsyt.2019.00505.

[13] Jiang T, Gradus JL, Rosellini AJ. Supervised Machine Learning: A Brief Primer. Behavior Therapy. 2020 09;51(5):675-87. Available from: doi:10.1016/j.beth.2020.05.002.

[14] Baloglu O, Latifi SQ, Nazha A. What is machine learning? Archives of disease in childhood Education and practice edition. 2021:edpract-2020319415.

[15] Wang SC. Artificial neural network. In: Interdisciplinary Computing in Java Programming. vol. 743 of The Springer International Series in Engineering and Computer Science. Boston, MA: Springer; 2003. p. 81-100. Available from: doi:10.1007/978-1-4615-0377-4_5.

[16] Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET); 2017. p. 1-6. Available from: doi:10.1109/ICEngTechnol.2017.8308186.

[17] IBM Cloud Education. Recurrent Neural Networks. IBM; 2020. Available from: https://www.ibm.com/cloud/learn/recurrent-neural-networks. [Accessed 2nd December 2021].

[18] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997 11;9(8):1735-80. Available from: doi:10.1162/neco.1997.9.8.1735.

[19] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv [Preprint]. 2014. arXiv:1406.1078. Available from: https://arxiv.org/pdf/1406.1078.pdf. [Accessed 2nd December 2021].

[20] Phi M. Illustrated Guide to LSTM's and GRU's: A step by step explanation. Medium; 2018. Available from: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21. [Accessed 2nd December 2021].

[21] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv [Preprint]. 2016. arXiv:1409.0473. Available from: https://arxiv.org/pdf/1409.0473.pdf. [Accessed 4th December 2021].

[22] Galassi A, Lippi M, Torroni P. Attention in Natural Language Processing. IEEE Transactions on Neural Networks and Learning Systems. 2021;32(10):4291-308. Available from: doi:10.1109/TNNLS.2020.3019893.

[23] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al.. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. p. 5998-6008. Available from: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. [Accessed 11th December 2021].

[24] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv [Preprint]. 2019. arXiv:1810.04805. Available from: https://arxiv.org/pdf/1810.04805.pdf. [Accessed 27th December 2021].

[25] Khan S. BERT Technology introduced in 3-minutes. Medium; 2019. Available from: https://towardsdatascience.com/bert-technology-introduced-in-3-minutes-2c2f9968268c. [Accessed 28th December 2021].

[26] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc.; 2019. Available from: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf. [Accessed 27th December 2021].

[27] Khan S. BERT, RoBERTa, DistilBERT, XLNet — which one to use?. Medium; 2019. Available from: https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8. [Accessed 28th December 2021].

[28] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv [Preprint]. 2019. arXiv:1907.11692. Available from: https://arxiv.org/pdf/1907.11692.pdf. [Accessed 27th December 2021].

[29] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv [Preprint]. 2020. arXiv:1910.10683. Available from: https://arxiv.org/pdf/1910.10683.pdf. [Accessed 27th December 2021].

[30] Roberts A, Raffel C. Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer. Google AI Blog; 2020. Available from: https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html. [Accessed 28th December 2021].

[31] Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower Provost E, Kim S, et al. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation. 2008 12;42(4):335-59. Available from: doi:10.1007/s10579-008-9076-6.

[32] Bagher Zadeh A, Liang PP, Poria S, Cambria E, Morency LP. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 2236-46. Available from: doi:10.18653/v1/P18-1208.

[33] Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 527-36. Available from: doi:10.18653/v1/P19-1050.

[34] Chen SY, Hsu CC, Kuo CC, Ting-Hao, Huang, Ku LW. EmotionLines: An Emotion Corpus of Multi-Party Conversations. arXiv [Preprint]. 2018. arXiv:1802.08379. Avail-

able from: https://arxiv.org/pdf/1802.08379.pdf. [Accessed 19th November 2021].

[35] Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE. 2018 05;13(5):1-35. Available from: doi:10.1371/journal.pone.0196391.

[36] Park CY, Cha N, Kang S, Kim A, Khandoker A, Hadjileontiadis L, et al. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Scientific Data. 2020 09;7:293. Available from: doi:10.1038/s41597-020-00630-y.

[37] Tzirakis P, Trigeorgis G, Nicolaou MA, Schuller BW, Zafeiriou S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. IEEE Journal of Selected Topics in Signal Processing. 2017 12;11(8):1301-9. Available from: doi:10.1109/JSTSP.2017.2764438.

[38] Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S. Multimodal Sentiment Analysis using Hierarchical Fusion with Context Modeling. Knowledge-Based Systems. 2018;161:124-33. Available from: doi:10.1016/j.knosys.2018.07.041.

[39] Chen G, Zeng X. Multi-Modal Emotion Recognition by Fusing Correlation Features of Speech-Visual. IEEE Signal Processing Letters. 2021;28:533-7. Available from: doi:10.1109/LSP.2021.3055755.

[40] Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Computation. 2004 12;16(12):2639-64. Available from: doi:10.1162/0899766042321814.

[41] Shan C, Gong S, McOwan PW. Beyond Facial Expressions: Learning Human Emotion from Body Gestures. In: Rajpoot NM, Bhalerao AH. (eds.) Proceedings of the British Machine Vision Conference. BMVA Press; 2007. p. 43.1-43.10. Available from: doi:10.5244/C.21.43.

[42] Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D. M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. Proceedings of the AAAI Conference on Artificial Intelligence. 2020 04;34(02):1359-67. Available from: doi:10.1609/aaai.v34i02.5492.

[43] Liu K, Li Y, Xu N, Natarajan P. Learn to Combine Modalities in Multimodal Deep Learning. arXiv [Preprint]. 2018. arXiv:1805.11730. Available from: https://arxiv.org/pdf/1805.11730.pdf. [Accessed 26th November 2021].

[44] Shenoy A, Sardana A. Multilogue-Net: A Context-Aware RNN for Multimodal Emotion Detection and Sentiment Analysis in Conversation. In: Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML). Seattle, USA: Association for Computational Linguistics; 2020. p. 19-28. Available from: doi:10.18653/v1/2020.challengehml-1.3.

[45] Delbrouck JB, Tits N, Brousmiche M, Dupont S. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. In: Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML). Seattle, USA: Association for Computational Linguistics; 2020. p. 1-7. Available from: doi:10.18653/v1/2020.challengehml-1.1.

[46] Xie B, Sidulova M, Park CH. Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion. Sensors. 2021;21(14). Available from: doi:10.3390/s21144913.

[47] Siriwardhana S, Kaluarachchi T, Billinghurst M, Nanayakkara S. Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion. IEEE Access. 2020;8:176274-85. Available from: doi:10.1109/ACCESS.2020.3026823.

[48] Choi JH, Lee JS. EmbraceNet: A robust deep learning architecture for multimodal classification. Information Fusion. 2019;51:259-70. Available from: doi:10.1016/j.inffus.2019.02.010.

[49] Bulat A, Tzimiropoulos G. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In: International Conference on Computer Vision (ICCV). Los Alamitos, CA, USA: IEEE Computer Society; 2017. p. 1021-30. Available from: doi:10.1109/ICCV.2017.116.

[50] Simkanin L. Multi-emotion Recognition and Dialogue Manager for VR-based Self-attachment Therapy. MSc Project. Department of Computing, Imperial College London; 2020. Available from: http://humandevelopment.doc.ic.ac.uk/papers/ls-thesis-2020.pdf.

[51] Tavernor J. Cross-corpus Speech and Textual Emotion Learning for Psychotherapy. MEng Project. Department of Computing, Imperial College London; 2020. Available from: https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1920-ug-projects/distinguished-projects/Cross-corpus-Speech-and-Textual-Emotion-Learning-for-Psychotherapy.pdf.

[52] Rizos G, Hemker K, Schuller B. Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19). CIKM '19. New York, NY, USA: Association for Computing Machinery; 2019. p. 991–1000. Available from: doi:10.1145/3357384.3358040.

[53] Gotsman T, Polydorou N, Edalat A. Valence/Arousal Estimation of Occluded Faces from VR Headsets. In: 2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI); 2021. p. 96-105. Available from: doi:10.1109/CogMI52975.2021.00021.

[54] Mollahosseini A, Hasani B, Mahoor MH. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing. 2019;10(1):18-31. Available from: doi:10.1109/TAFFC.2017.2740923.

[55] Tong E, Zadeh A, Jones C, Morency LP. Combating Human Trafficking with Multimodal Deep Models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics; 2017. p. 1547-56. Available from: doi:10.18653/v1/P17-1142.

[56] Paul Ekman Group. Enjoyment. Paul Ekman Group LLC; 2019. Available from: https://www.paulekman.com/universal-emotions/what-is-enjoyment/. [Accessed 27th April 2022].

[57] Paul Ekman Group. Sadness. Paul Ekman Group LLC; 2019. Available from: https:

//www.paulekman.com/universal-emotions/what-is-sadness/. [Accessed 27th April 2022].

[58] Paul Ekman Group. Anger. Paul Ekman Group LLC; 2019. Available from: https://www.paulekman.com/universal-emotions/what-is-anger/. [Accessed 27th April 2022].

[59] Paul Ekman Group. Surprise. Paul Ekman Group LLC; 2019. Available from: https://www.paulekman.com/universal-emotions/what-is-surprise/. [Accessed 27th April 2022].

[60] Paul Ekman Group. Disgust. Paul Ekman Group LLC; 2019. Available from: https://www.paulekman.com/universal-emotions/what-is-disgust/. [Accessed 27th April 2022].

[61] Paul Ekman Group. Fear. Paul Ekman Group LLC; 2019. Available from: https://www.paulekman.com/universal-emotions/what-is-fear/. [Accessed 27th April 2022].

[62] Gabert-Quillen CA, Bartolini EE, Abravanel BT, Sanislow CA. Ratings for emotion film clips. Behavior Research Methods. 2015 Sep;47(3):773-87. Available from: doi:10.3758/s13428-014-0500-0.

[63] Chen H, Chin KL, Tan CBY. Selection and validation of emotional videos: Dataset of professional and amateur videos that elicit basic emotions. Data in Brief. 2021;34:106662. Available from: doi:https://doi.org/10.1016/j.dib.2020.106662.

[64] Rottenberg J, Ray RD, Gross JJ. Emotion elicitation using films. In: Coan JA, Allen JJB. (eds.) Handbook of Emotion Elicitation and Assessment. Oxford University Press; 2007. p. 9-28.

[65] Hewig J, Hagemann D, Seifert J, Gollwitzer M, Naumann E, Bartussek D. A revised film set for the induction of basic emotions. Cognition and Emotion. 2005;19(7):1095-109. Available from: doi:10.1080/02699930541000084.

[66] Zupan B, Eskritt M. Eliciting emotion ratings for a set of film clips: A preliminary archive for research in emotion. The Journal of Social Psychology. 2020;160(6):768-89. Available from: doi:10.1080/00224545.2020.1758016.

[67] Cabral JCC, Tavares PdS, Weydmann GJ, das Neves VT, de Almeida RMM. Eliciting Negative Affects Using Film Clips and Real-Life Methods. Psychol Rep. 2017 Sep;121(3):527-47.

Appendix A

Ethics Checklist and Ethics Approval

	Yes	No
Section 1: HUMANS		
Does your project involve human participants?	x	
Section 2: PROTECTION OF PERSONAL DATA		
Does your project involve personal data collection and/or processing?	x	
Does it involve the collection and/or processing of sensitive personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?	x	
Does it involve processing of genetic information?		x
Does it involve tracking or observation of participants? It should be noted that this issue is not limited to surveillance or localization data. It also applies to Wan data such as IP address, MACs, cookies etc.	x	
Does your project involve further processing of previously collected personal data (secondary use)? For example Does your project involve merging existing data sets?		x
Section 3: ANIMALS		
Does your project involve animals?		x
Section 4: DEVELOPING COUNTRIES		
Does your project involve developing countries?		x
If your project involves low and/or lower-middle income countries, are any benefit-sharing actions planned?		x
Could the situation in the country put the individuals taking part in the project at risk?		x
Section 5: ENVIRONMENTAL PROTECTION AND SAFETY		
Does your project involve the use of elements that may cause harm to the environment, animals or plants?		x
Does your project involve the use of elements that may cause harm to humans, including project staff?		x
Section 6: DUAL USE		
Does your project have the potential for military applications?		x
Does your project have an exclusive civilian application focus?		x
Will your project use or produce goods or information that will require export licenses in accordance with legislation on dual use items?		x
Does your project affect current standards in military ethics – e.g., global ban on weapons of mass destruction, issues of proportionality, discrimination of combatants and accountability in drone and autonomous robotics developments, incendiary or laser weapons?		x

Section 7: MISUSE		
Does your project have the potential for malevolent/criminal/terrorist abuse?		x
Does your project involve information on/or the use of biological-, chemical-, nuclear/radiological-security sensitive materials and explosives, and means of their delivery?		x
Does your project involve the development of technologies or the creation of information that could have severe negative impacts on human rights standards (e.g. privacy, stigmatization, discrimination), if misapplied?		x
Does your project have the potential for terrorist or criminal abuse e.g. infrastructural vulnerability studies, cybersecurity related project?		x
SECTION 8: LEGAL ISSUES		
Will your project use or produce software for which there are copyright licensing implications?		x
Will your project use or produce goods or information for which there are data protection, or other legal implications?		x
SECTION 9: OTHER ETHICS ISSUES		
Are there any other ethics issues that should be taken into consideration?		x

Imperial College London

Imperial College Research Ethics Committee and Science, Engineering and Technology Research Ethics Committee.
Imperial College London
Room 221, Medical School Building
St Marys Campus
London, W2 1PG
Tel: +44 (0)207 594 1862

researchethicscommittee@imperial.ac.uk

24 March 2022

Dear Prof. Abbas Edalat

Study Title: Emotion Recognition using a Multi-modal approach (Algorithmic Human Development)

This study is deemed low by the Research Governance and Integrity Team (RGIT) on 24/03/22.

Based on the information provided; our decision is that the project can be reviewed for ethical consideration by the project supervisor.

Documents
The documents reviewed were:

- Undergraduate study proposal ethics checklist

Yours sincerely,

Ruth Nicholson,
Head of Research Governance and Integrity,
Imperial College London

Appendix B

Selected Film Clips for Emotion Elicitation

Table B.1 details the film clips selected for the human trial. These clips are chosen because it has been demonstrated by researchers that they are effective in eliciting the target emotion. In addition, their effectiveness is validated by checking if the video matches one of the common triggers for each emotion. Some example common triggers are as follows [56–61]:

- **Happy:** witnessing something humorous or amusing, witnessing act of kindness or compassion
- **Sad:** sickness or death of a loved one, endings and goodbyes, rejections
- **Angry:** interference, injustice, someone trying to hurt loved ones
- **Surprised:** mostly unexpected events such as unexpected movements, loud sounds
- **Disgusted:** expelled bodily products, something rotting, sometimes may be bloody
- **Fearful:** darkness, death and dying, normally from a horror film

Emotion	Film Title	Trigger [56–61]	Description
Happy	Wall-E [62]	Witnessing something humorous or amusing; witnessing act of compassion	Two robots dancing together and falling in love in outerspace. Available from: https://www.youtube.com/watch?v=kZnzPdN_R7A
Happy	Playing with Fire [63]	Witnessing or participating in acts of human goodness, kindness, and compassion; witnessing something humorous	Three man taking care over children who are creating a lot of mess. Available from: https://www.youtube.com/watch?v=2altG6wElVE&list=PLZbXA4lyCtqrOGOEA_0xXrVD875NUj4i7&index=4
Sad	The Champ [64]	Sickness or death of a loved one; endings and goodbyes	A boy is crying and finding it hard to accept the death of his father. Available from: https://www.youtube.com/watch?v=SfuewyrDSZc

Emotion	Film Title	Trigger [56–61]	Description
Sad	My Girl [62]	Sickness or death of a loved one; endings and goodbyes	A girl not understanding the death of her friend and tries to give him his glasses in the funeral. Available from: https://www.youtube.com/watch?v=woLbaFLoJI8
Angry	My Bodyguard [64]	Interference; injustice	Two guys get picked on by two other men and getting beaten up. Available from: https://www.youtube.com/watch?v=NzRDoDV8Sbo
Angry	Witness [65]	Interference; injustice	A group of Amish travels to the town centre and gets stopped and harassed by some teenagers. Available from: https://www.youtube.com/watch?v=6DkDGCI9fkc
Surprised	One Day [66]	Unexpected movements	A woman rides on a bicycle to meet someone and gets hit by a truck. Available from: https://www.youtube.com/watch?v=YDN6Ch5PhS4
Surprised	The Call [66]	Unexpected movements	A woman on the phone talking to someone and gets kidnapped. Available from: https://www.youtube.com/watch?v=oZ0tK7wVAIk (22:16 22:45)
Disgusted	The Fly [62]	A person, animal or thing one considers physically ugly	A creature (half man half fly) vomits digestive enzymes onto the man. Available from: https://www.youtube.com/watch?v=yuWXMMvnVcY
Disgusted	Crash [66]	Perceived perversions or actions of other people	An interracial couple gets stopped by the police when they did not drink. The police sexually harasses the woman until the man apologises. Available from: https://www.youtube.com/watch?v=cYwGh8XZb3U
Fearful	Lights Out [67]	Darkness; social interaction	A woman turns the lights on and off in her house and a creature appears whenever the lights are off. At the end the creature appears next to her bed. Available from: https://www.youtube.com/watch?v=FUQhNGEu2KA
Fearful	The Ring [62]	Darkness; social interaction	A man's TV turns on and off itself, and a woman crawls out of the TV. Available from: https://www.youtube.com/watch?v=hpb2-ZOzc_o

Table B.1: Selected film clips that elicit the six emotions in CMU-MOSEI

Appendix C

Parameter Configuration for Hyperparameter Tuning

Table C.1 details the hyperparameter values in various model trained as part of hyperparameter tuning. Model A is the final model configuration.

Model	Learning Rate	Batch Size	Hidden Size	Embracement Size
A	0.0001	32	1024	256
B	0.0001	16	1024	256
C	0.0001	64	1024	256
D	0.0001	32	256	256
E	0.0001	32	512	256
F	0.0001	32	1024	512

Table C.1: Model parameter configuration for hyperparameter tuning