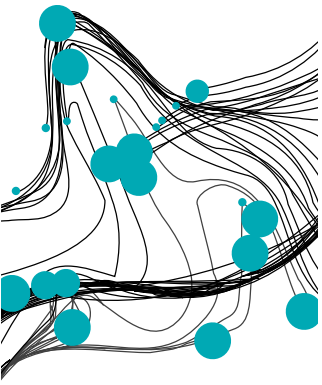
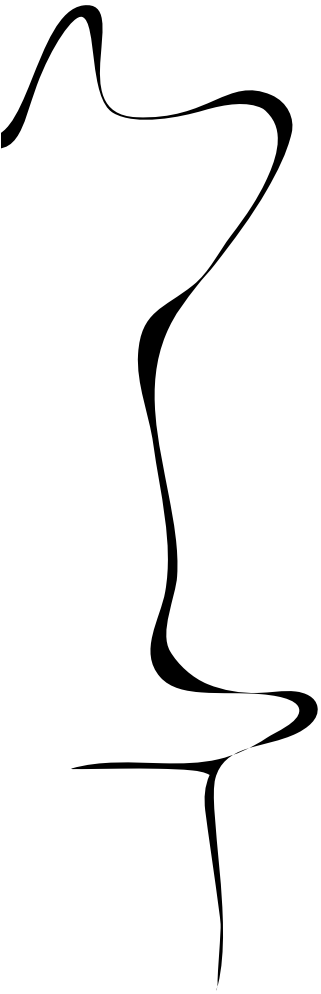


DMB

DATABASE MANAGEMENT
AND
BIOMETRICS

.86214



PERSPECTIVE INTERACTIONS: DETECTING MULTIMODAL SOCIAL INTERACTIONS FROM AN EGOCENTRIC VIEW

Aditya Nadar

MASTER'S ASSIGNMENT

Committee:
Dr. M.Sc. Estefanía Talavera Martínez
Dr. Ir. Luuk Spreeuwers
Dr. Faizan Ahmed

February, 2024

2024DMB0001
Data Management and Biometrics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Perspective Interactions: Detecting Multimodal Social interactions from an Egocentric View

Aditya Nadar
M.Sc. Embedded Systems
University of Twente
Enschede, The Netherlands

Dr. Estefanía Talavera Martínez
Faculty EEMCS
University of Twente
Enschede, The Netherlands

***Abstract*—The idea of integrating data from multiple modalities is instinctively attractive as it can enhance the efficacy of Machine learning models. The system proposed here utilizes multiple modalities in the form of video and audio to develop a multimodal deep learning system capable of classifying *Talking to me* based social interactions from an egocentric point of view. This study extends the baseline work of Ego4D social interactions by devising a methodology to employ different multimodal fusion techniques, namely Early and Late fusion, and later realizing the optimal alternative to fuse the modalities. To employ these fusion techniques and implement optimizations at different stages of a multimodal model, the system explores a multimodal framework called Multibench. The dataset used for this study is Ego4D, which consists of 3,670 hours of egocentric videos, the subset pertaining to social benchmark has been used. By employing Multibench and its offered optimizations, our approach shows a mAP performance improvement of 3.67% (for Early fusion) and 5.52% (for Late fusion) compared to the baseline. The study also establishes a performance comparison between Early and Late fusion to identify the superior alternative of multimodal fusion with the dataset in hand. This study concludes by discussing the shortcomings of the system and guidelines for future improvements.**

***Index Terms*—Multimodal models, Multimodal fusion, Talking to me, Social interactions, Multibench**

I. Introduction

A. Motivation

Multimodal learning is a subset of machine learning which aims to train AI models while taking multiple modalities into account. This technique thus manages to adequately consider the cues present in multiple modalities, thereby giving robust results.

There has been significant research in the field of unimodal models, but they have their limitations when understanding complex human behavioural pattern. Thus, multimodal models present a robust solution by taking multiple modalities into account [1][2][3]. The current research work proposes use of multimodal models in social settings with the aim of accurately analyzing social interactions and classifying complex scenarios of *Talking to me* from an egocentric point of view.

The motivation of our proposed solution in this study stems from following application scenarios:

- Improvements in the performance of egocentric-based models contribute immensely to the advancements in the field of social robots. These models enable social robots and Human-computer Interaction (HCI) based systems to indulge in more natural and human-like interactions with users.
- Efficient Talking to me models can assist robots to detect when the conversation is directed towards them in an improved manner. Egocentric models also play a crucial role in the field of affective computing by facilitating socially and emotionally aware interactions.
- Understanding Talking to me based interactions from an egocentric perspective play a monumental role in applications in the field of assistive robotics and virtual assistants. By better understanding these complex social interactions, these systems can provide more effective support and assistance, ultimately enhancing user's quality of life.

Our proposed study impart innovations in developing efficient Talking to me (egocentric) based models, and thus has a potential application in development of smarter social robots, smarter virtual assistants and egocentric robots, with a broader goal to develop AI solutions for the betterment of mankind. Some major contributions of our proposed study are as follows:

- Exploring the applicability of Multibench framework and two multimodal fusion techniques, Early and Late fusion.
- Devising a novel methodology of developing a multimodal system using Ego4D with Multibench multimodal framework.
- Analysing the performance comparison across different set of experiments, identifying optimal fusion alternative and establishing new set of results on Ego4D dataset.

B. Research Questions

The system proposed here builds on top of the baseline work of Ego4D *Talking to me* based social interactions [1]. To develop such a system a multimodal framework called Multibench is employed [4]. The work aims at exploring different multimodal fusion architectures and optimization techniques present in Multibench and its effect on the performance of multimodal model. A performance comparison has also been drawn to highlight the improvements when developing Multibench based multimodal model over the baseline work. To adequately explore this following research question and sub research questions are proposed.

RQ: How does multiple modalities affect the prediction of *Talking to me* based social interactions from egocentric point of view?

The sub research questions indispensable in solving the above research question are as follows:

SRQ1: How does the implementation of the *Multibench* framework contribute to performance enhancements, if any compared to the *Ego4D* baseline in classifying Talking to me based social interactions?

SRQ2: Which fusion technique provides superior performance when integrating modalities?

The organization of the paper is as follows, Section II discuss the scientific background and work already done in the past in the field of multimodal models and fusion techniques. Section III introduces the methodology and techniques implemented to answer the research questions. It also includes the details about the baseline model, the framework implemented , the fusion techniques considered with their conceptual aspects. Section IV discuss the experimental setup to perform the experiments. It includes thorough description of the dataset, performance evaluation metrics, and implementation details enlisting information about the hyperparameters used across different set of experiments. Section V discuss the results and performance comparison across different experiments, along with an in-depth discussion about the results. Section VI includes the discussion regarding how the proposed work answers the research questions enlisted before. It also discuss the limitations and possible guidelines for the future work. Section VII concludes the study and give some final statements regarding the relevance of work and what contribution it presents to the existing technologies.

II. Scientific background

A. Technical background

1) Talking to me Baseline model: As a starting point of our research, baseline implementation of social interactions benchmark of Ego4D dataset was considered [1]. Out of 3,670 hours of video in Ego4D, approximately 764 hours of data containing conversational content was pertinent to the Audio-Visual diarization and Social benchmark tasks. The data of 572 clips was organized as follows for baseline model training : 389 clips for training, comprising 32.4 hours in total. Further, 50 clips (4.2 hours) and 133 clips (11.1 hours) were held out for the validation and testing sets, respectively.

The Talking to me (TTM) task gives a frame level classification label y which is a binary label with $y = 1$ indicating that the target is talking to the camera bearer and $y = 0$ indicating otherwise. For performance evaluation the benchmark uses mean average precision (mAP) and Top-1 accuracy to quantify the classification performance for the TTM task. The precision parameters are measured for each frame of the videos. The baseline model gave a mAP score of 55.06 while testing for TTM task. This clearly has a significant scope of improvement. The work proposed in this paper utilizes the baseline model as the backbone and explores significantly in investigating the most optimal fusion alternative with the primary aim of improving the performance of the system.

Baseline Model : The baseline model framework is depicted in Figure 1, where it utilizes a backbone of Resnet-18 and Bi-LSTM for video encoder, while MFCC filtering technique in conjunction with Resnet-18 (essentially called ResSE) for audio encoder. These techniques play an important role from the perspective of feature extraction and model framework, part of which is employed in our research as well.

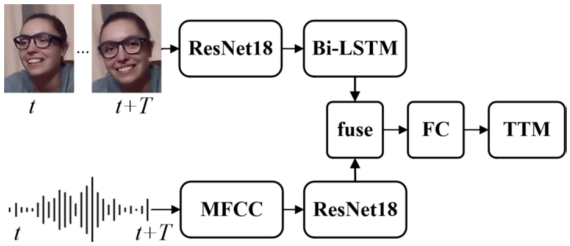


Fig. 1: Talking to me Model Baseline approach [1].

Annotations: The baseline work provides annotations pertaining to face tracking with bounding boxes, active speaker detection and voice activity annotations. The tracked annotations comprises of face bounding boxes with ids of participants labelled across frames [1]. Active speaker annotation identify if the faces present in the frames are actually speaking. A ground truth information providing binary labels for each face in the frames are also provided in the baseline implementation. Furthermore, the TTM task as implemented in baseline does the job of identifying the time segments when the speech is directed at the camera bearer. Figure 2 gives a visualization of various annotations provided by the baseline work of Ego4D social interactions. It depicts how the TTM annotations are employed to the tracked information of faces and audio in the clips.

2) MultiBench framework: Multibench is a Multimodal framework which aims at developing an end to end machine learning pipeline capable of simplifying data loading, experimental setup and model evaluation [4]. It offers an in depth methodology to assess (1) generalization, (2) time and space complexity, and (3) modality robustness. It introduces several multimodal methods to adequately analyze the models based on these criteria. Thus it provides a single platform for an end to end evaluation of multimodal models. These methods are adapted from individual research works, and thus the framework facilitates provision of plethora of mulitmodal methods under a single umbrella. This is achieved by using a comprehensive toolkit, 'Multizoo' which provides a starter code for multimodal algorithms implementing 20 methods spanning different methodological innovations in (1) data preprocessing, (2) fusion paradigms, (3) optimization objectives, and (4) training procedures. A detailed implementation of Multibench with Ego4D dataset is provided in Methodology section of this report.

The framework has been implemented on 15 datasets (also comprising of multimodal datasets) of varying fields of appli-

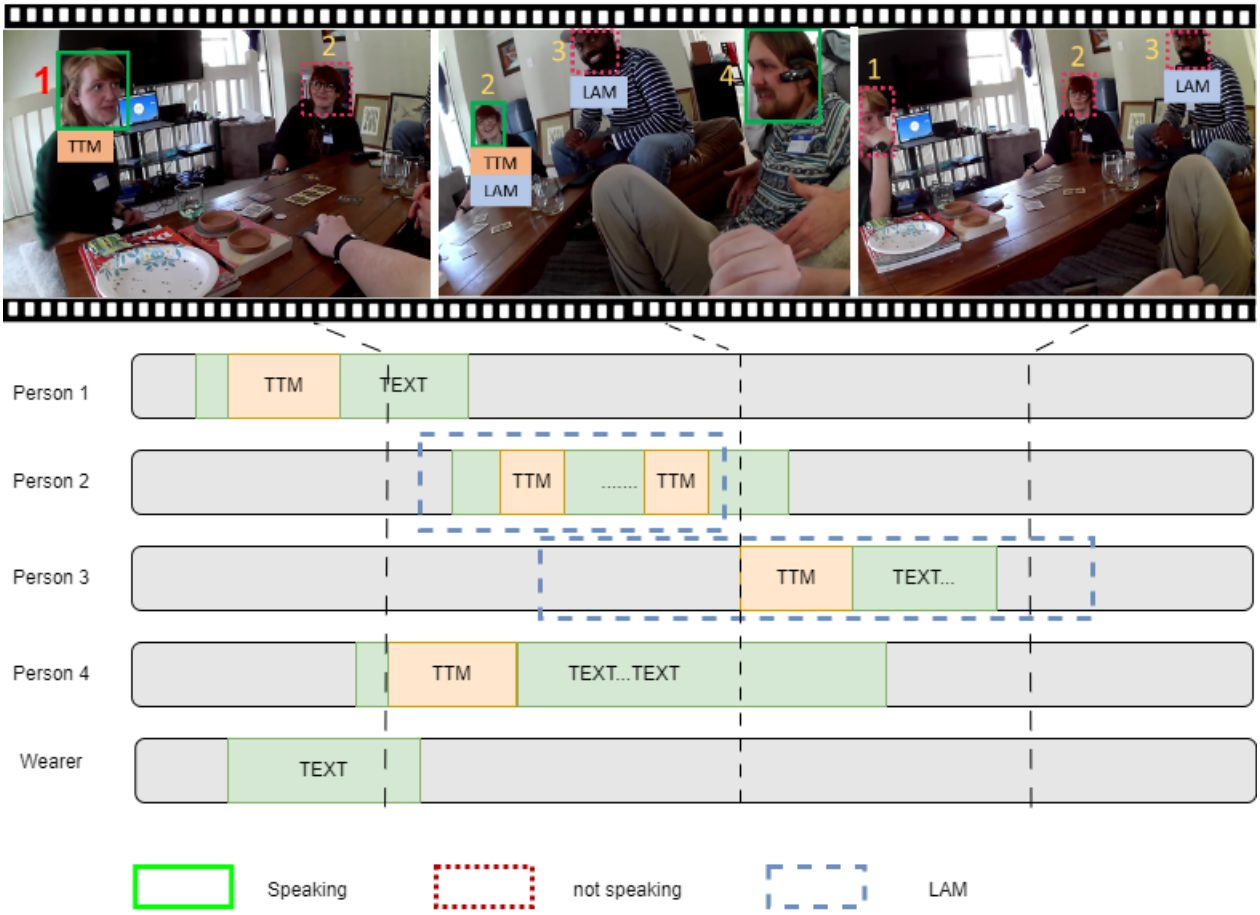


Fig. 2: Social benchmark annotations Ego4D depicting Talking to me(TTM) instances, LAM denotes looking at me scenarios[1]. Green boxes indicate frame stamps of Talking to me instances, red dashed boxes indicate instances of speaker not speaking. Blue dashed box indicate frame stamps of person Looking at me.

cation, and giving state of the art results in 9/15 datasets. But its applicability is yet to be realised with Ego4D dataset and has never been explored before.

3) Multimodal fusion: Multimodal fusion involves combining data from numerous modalities, like text, images, audio, video, and sensor readings, to enrich the comprehension or analysis of a specific phenomenon or problem. Essentially, it integrates information from diverse sources to achieve a more holistic and precise depiction of the underlying data. This study analyzes two fusion techniques namely Early and Late fusion. The two fusion techniques are implemented using *Multizoo* toolkit encompassed under Multibench framework. Mathematical background for the two fusion techniques is as follows:

- **Early fusion :** In early fusion, features extracted from separate raw data modalities are integrated into a unified representation which is later subjected to classification methodologies [4]. Mathematically it is represented by

$$z_{mm} = [x_1, x_2] \tag{1}$$

where x_1, x_2 refers to input features from individual modalities and z_{mm} for multimodal representations.

- **Late fusion :** In late fusion modality wise classification results are combined to give the output [5]. Mathematically it is represented by

$$z_{mm} = [z_1, z_2] \tag{2}$$

where z_{mm} hold the similar representations as mentioned in early fusion, and z_1, z_2 represent unimodal representations.

A detailed methodology for the implementation of the two fusion techniques along with the architectural diagrams is discussed in III-D.

B. Related works

Multimodal fusion is an indispensable task when developing multimodal deep learning models [6]. Effective fusion of multiple modalities is highly challenging and still forms a substantial area of research in multimodal learning.

Another work by the authors of Multibench [7], where a High modality multimodal transformer model was implemented to enable multitask and transfer learning. It also devised *modality heterogeneity* metrics to measure how much information can be transferred from one modality to other

and *interaction heterogeneity* to study how similarity exists between different modalities when they interact. Moreover, the work thoroughly utilized multibench framework to develop the system and thus proved crucial while developing the multimodal system proposed in our approach as well. The work is published in 2023 and discuss several improvement techniques in the field of multimodal learning, while also showcasing the applicability of Multibench framework.

1) Early fusion and Late fusion: A performance comparison between unimodal and multimodal models for modalities in the form of RGB video, optical flow and skeleton data on NTU RGB+D dataset was established in the work proposed by *Gadzicki et al.*[2] Multimodal fusion techniques such as early and late fusion were also employed and the performance comparison between the two fusion techniques was established thereby giving insights about the pros and cons of each technique with the considered dataset. The work showed some significant performance improvements of using multimodal approaches over conventional unimodal approaches. It also gave a reasoning behind observed finding of early fusion performing better than late fusion and the idea of cross correlation of features between modalities(a potential in early fusion) as the reason behind it. The work acknowledges some future improvements such as identifying the applicability of half way fusion and hybrid approaches of fusion. It also fails to consider the robustness of individual modalities and the application of weighted fusion approaches.

To analyze the shortcomings of early, intermediate and late fusion techniques, a performance comparison across three techniques was implemented on two datasets including NTU RGB-D dataset [8]. The modalities considered were RGB, Depth and Skeleton modalities. For early fusion the work proposes a two step recognition approach by applying depth mask to RGB images. For intermediate fusion, the work proposed utilizing a deep learning technique to combine features by employing a qualitative feature selection method. This method aimed to identify and select the most unique features. For late fusion, an end to end pipeline based on deep neural network was proposed. It relied on three pre-trained architectures to generate score vectors from each modality individually. After the preprocessing step, feature vectors were utilized for training purposes. The work acknowledges future improvements such as search strategies for acquiring best features of the intermediate approach, also considering better fusion techniques.

A multimodal classification system with image and text modalities was developed by *Gallo et al.* [9]. The work also performs a careful trade off between two main basic fusion techniques early and late fusion with an add-on of stacking techniques. The choice of dataset for the experiment was UPMC Food-101 which is a noisy multimodal dataset. The paper identified that early fusion performed better than late fusion and gave state of the art results on the dataset, as claimed by the authors.

To perform transcription of music from audio and image modalities a multimodal approach was proposed [10]. Several late fusion strategies were analyzed and a performance comparison with unimodal counterparts was also established. The work concludes by identifying significant performance improvements over unimodal transcription by some of the fusion methods.

2) Other fusion techniques: A novel architecture for multimodal fusion named conditional attention fusion for conditional dimensional emotion prediction based on LSTM-RNN was proposed by *Chen et al.*[11]. The technique employs LSTMs to pay attention to different modalities by taking into account current features and history information. The experiments were done on AVEC 2015 dataset and the fusion technique claims to outperform conventional fusion technique such as early, model-level and late fusion. Future improvements include implementing more features from different modalities.

Gradient blend fusion technique addresses the bottlenecks of conventional multimodal fusion techniques [12][13][14]. This fusion technique computes an optimal blend of modalities based on their overfitting behaviour. It does so by assigning dynamic weights to the modalities and then employing fusion techniques. An implementation of this technique is also available in multibench framework. It certainly holds a great research potential and as future work it can prove to be a good alternative to conventional fusion techniques like Late and Early fusion.

An optimal fusion neural architecture design for the task of image classification was introduced by *Zhou et al.* [15]. The paper focused on devising a neural architectural search space for uni modality (image specifically) and introduced a surrogate function which forms the backbone of implementing an efficient progressive neural architecture search. The dataset used in this paper was CIFAR-10. However, the work only utilizes unimodality and thus of limited use for the problem discussed in this paper. Nevertheless, the technique of surrogate function forms the basis of neural architectural search which can further be extended to multimodal models as well.

Extending the research of neural architectural search, *Xu, Dai et al.*[16] identify multimodal fusion architectures to address the problem of applying deep learning in Electronic health records (EHR). The modalities include codes (structured) and free-text (unstructured). The work extends state of the art Neural architectural search (NAS) by proposing Multimodal fusion architectural search (MUFASA). Furthermore, it also draws the comparison between unimodal NAS and MUFASA on public EHR. As claimed in the paper the fusion technique devised outperforms the established NAS. MUFASA does so by customizing each data modality and finding effective fusion strategies. As discussed in the paper, future work involves investigating the applicability of MUFASA to other types of modalities such as medical imaging. Thus this fusion technique is yet to be tested with modalities like Audio and image data.

Another multimodal fusion method called Low rank tensor fusion (LRTF) employs fusion using low rank tensors [17][18]. The study analyzes the computational complexity issues of

the multimodal methods and was evaluated on three tasks namely sentiment analysis, speaker trait detection and emotion recognition. The model achieved significantly good results on all the tasks and immensely reduced the computational complexity issues. An implementation of this same work is also provided as one of the multimodal method in Multibench framework, and thus can be explored as a potential future work alternative.

3) Multimodal deep learning: To analyze the advantages of multimodal models and its advantages over unimodal models *Ngiam, A.Ng et al.* [3] demonstrate cross modality learning , where better features of a modality are learned using multiple cues of data. The study considers modalities in the form of audio and video inputs on several datasets. The work adequately depicts the importance of multimodal learning but does not give much insights about optimal fusion techniques and their effect on the performance of the system, also Ego4D as dataset was not considered in this study.

Multimodal transformer models are highly prone to increased computational complexities[19] [20]. Thus to reduce these computational costs an efficient multimodal fusion technique called Prompt-based Multimodal Fusion (PMF) was proposed [21], which reduced the trainable parameters and training memory usage by 3% and 66% respectively.

Feature extraction plays a monumental role while developing machine learning solutions. Better feature extraction pipelines in individual modality signify improved accuracy of multimodal models [22][23]. The study [22] considered modalities in form of image and text and aimed at overcoming the bottlenecks of unimodal methods. For image features GoogLeNet deep convolutional neural network (DCNN) was employed and for text features word2vec methods were used. The study explores numerous feature extraction alternatives and thus can find its potential application as future improvements to our approach as well.

Ego4D video task translation by *Z. Xue et al.* [24] was the winner of Ego4D Challenge 2022. The study developed a multimodal system to perform Talking to me (TTM) predictions by utilizing a two stage training approach. In the proposed task translator, three tasks from Ego4D baseline were taken into account, with primary task being TTM, auxiliary tasks being Looking at me (LAM) and active speaker detection (ASD). In first stage of training, learned features were obtained from individual task models.In second stage of training these learned features were passed through a transformer encoder decoder model where the system learns to interpret into TTM predictions. The work thoroughly utilized techniques of different tasks present in the baseline work to develop an efficient solution for Talking to me predictions.

A multimodal model utilizing physiological signals such as heart rate to enhance the understanding of egocentric videos was presented in the study by *Nakamura et al.* [25]. The model aimed at developing a multitask prediction system to jointly predict energy expenditures and activity prediction. The study also introduced a custom made dataset comprising of 31 hours of egocentric video augmented with heart rate and acceleration signals.

To analyze the performance comparison between unimodal and multimodal models, discuss A multimodal price prediction model was proposed by *Zehtab et al.*[26]. A performance across several model variants namely, unimodal model ,Inception based feature extraction model, and CNN based feature extraction model was performed in the study. The results conclude that multimodal models performed better than the unimodal approach.

4) Social interactions using audio signals: The study proposed by *F. Vossebeld* [27] explores an approach for classification of social interactions using audio features on Ego4D dataset. Audio features were extracted from audio signals which were later fed to a model. A subset of Ego4D dataset was used in this paper. Numerous testing algorithms were considered in the paper and were further implemented on the training dataset. A performance trade off between considered algorithms was carried out thereby indicating the best performing alternative. The work presents an in depth implementation of ML classification algorithms for unimodal data in form of audio. However it does not consider the impact of other modalities and thus have a limited application in the current proposed research work.

Graph Convolutional Network (GCNs) have proven to be an effective method for social interaction recognition in egocentric videos [28]. However, this model alternative has not been explored with Multimodal methods and also with Ego4D dataset. Also due to non availability of model framework in Multibench, this work is of limited relevance in our proposed approach. However, it can be an alternative to explore in the future.

5) Data preprocessing techniques: Mel frequency cepstral coefficients (MFCC) is a robust mechanism for voice activity recognition and speaker detection in audio data [29] [30]. The study proposed by *Martinez at al.*[29] provide an in depth explanation to extract speech features, apply Mel filtering operations and vector quantization matching process. They also present tests and results on a database of 20 speakers thus identifying the right conditions and choice of filtering parameters to obtain accurate results. The baseline work also implements this method for audio processing and hence our proposed approach too.

A comparative study to evaluate the performance of several fusion techniques in context of image classification is discussed in the work [31]. Both binary and multi class classifications were considered in the study to identify the optimal feature extraction strategy for image classification.

Due to the emergence of Deep Learning, there are now sophisticated models capable of extracting valuable characteristics from audio signals. Tools like Wav2Vec (1.0 or 2.0) offer beneficial vector descriptions of audio files [32].

Analyzing the effects of different activation functions plays a very significant role when developing a deep learning system. The study proposed by *Wang et al.* [33] presents a CNN model for facial expression recognition. It points out the imminent shortcomings of activation techniques including

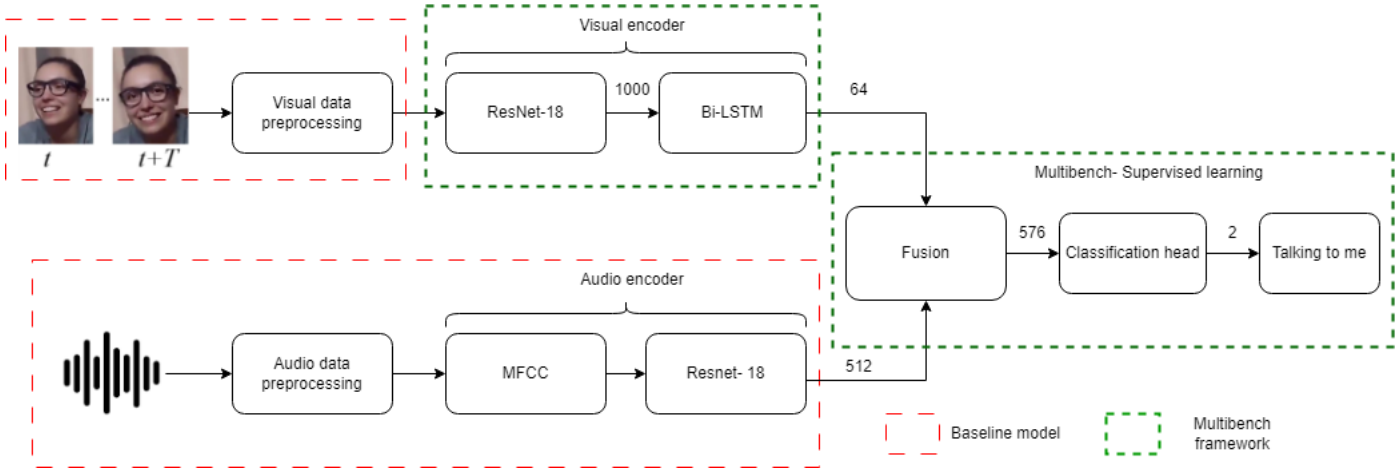


Fig. 3: Proposed Talking to me system architecture. The green dashed box demarcates the contribution of Multibench framework, while the red box indicates the techniques/ methodology derived from the baseline work of Ego4D social interactions.

Relu which is the most common one in use. It also highlights the importance of activation function in a model's ability to learn. Furthermore, it does a thorough trade off between numerous activation functions and also proposes a new activation function. The datasets used for the experiments were JAFFE and FER2013.

III. Methodology

Figure 3 demarcates the approach derived from the baseline work (in red dashed box) and the implementation of Multibench framework to employ supervised learning methods (in green dashed box) to the modalities, see https://github.com/pliang279/MultiBench.git. The framework introduces plug-in modules namely fusion modules, classification head and optimization objectives to improve model's performance. Further subsections discuss these steps in detail.

A. Input representations

In this study we take input representations in the form of visual and audio modalities. This section highlights the kind of input that is fed to the model and steps utilized to prepare the data. The steps discussed in this section are derived from the baseline implementation of Ego4D, see https://github.com/EGO4D/social-interactions.git. A detailed methodology of data preparation as done in the baseline work is given in *Appendix* D.

1) Visual representation: To enhance the Region of interest (ROI) the frames contained in visual modality are cropped according to width and height of the bounding box information provided in the tracked JSON segments. Face crops thus obtained are later resized to 224x224, thereby preparing visual information ready to be fed to the model for feature extraction process. A visual representation of this process is depicted in *Appendix* B. To account for the instances where the speaker leaves the field of vision and invisibility due to rapid motion, a padding of blank images is also employed to face sequences [1].

2) Audio representation: To prepare Talking to me segments from the baseline annotations, audio segments corresponding to the associated face crops are extracted. The varying length of audio segments is adjusted to limit the maximum duration to 1.5s. Segments shorter than 0.15s are skipped in training stage.

B. Porting Ego4D to Multibench framework

An important step to successfully develop a Talking to me based multimodal system using Multibench framework, and to employ the fusion techniques and optimization parameters, was to port the baseline work of Ego4D to Multibench.

Following steps were implemented to achieve this, also depicted in Figure 4:

- Developing dataloaders from the prepared dataset to synchronize data loading formats between the baseline work and Multibench. This essentially required developing loaders in sequence batch formats for training and validation sets.
- Finding the right model architecture along with necessary regularization techniques for the encoders. In current work the model Resnet + Bi-LSTM with 32 hidden layers and dropout functionality was explored as the video encoder. For audio encoder, the ResSE model as presented in baseline work was chosen (further details in next subsection).
- This step involved identifying the desired multimodal fusion techniques and developing in a manner such that fusion techniques could be synchronized with the prepared data. Out of numerous fusion techniques encompassed under *Fusion paradigms* of *Multibench*, two techniques namely early and late fusion are explored in this work.
- At this stage a classification head was chosen which performed the job of binary classification of talking to me based social interactions. For this purpose, a linear layer with *Xavier initialization* parameter was chosen.

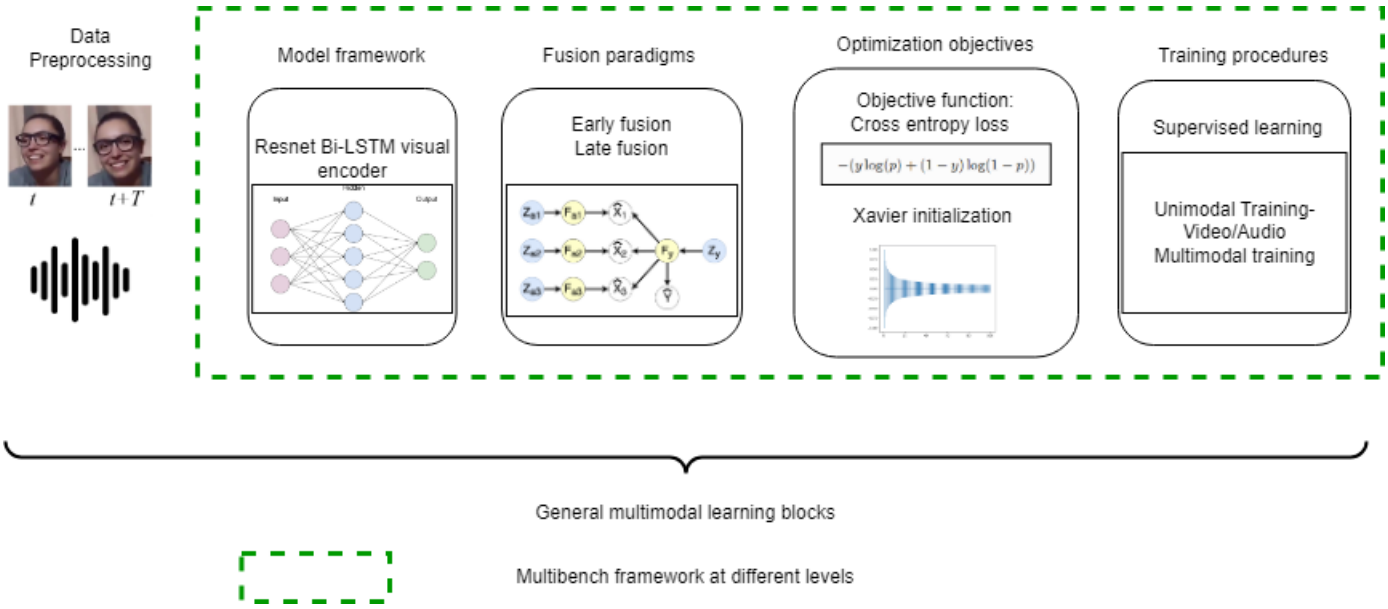


Fig. 4: Block diagram depicting general multimodal learning steps. The green highlighted box represent the blocks to which Multibench provide methods and optimizations while developing Talking to me based system in the proposed work.

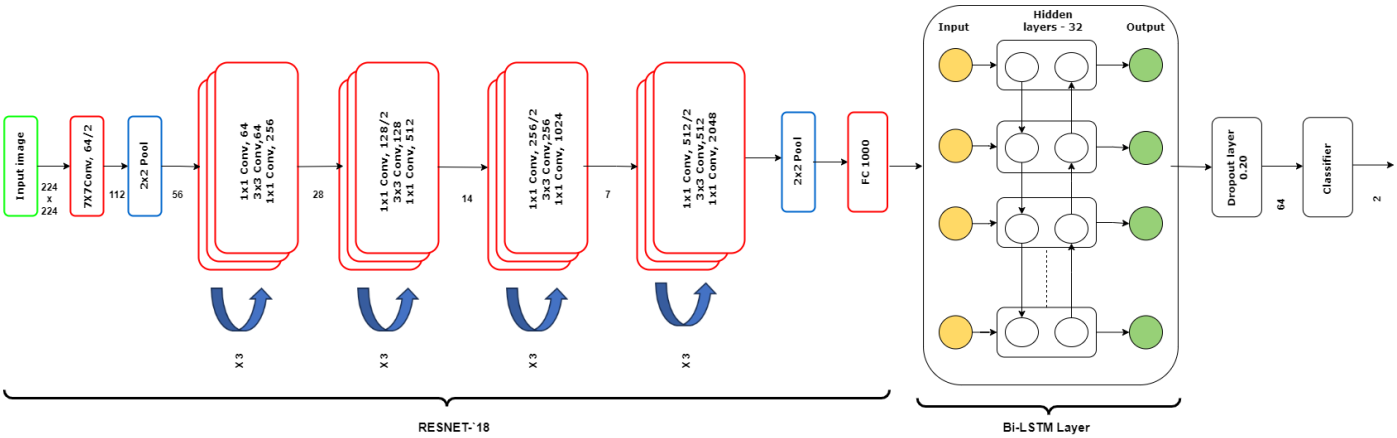


Fig. 5: Our proposed visual encoder model architecture - implemented using Multibench framework.

- For training procedure a supervised learning implementation from *Multibench* was employed. For training, hyper-parameters such as class weights, optimizer, loss function and learning rate were carefully chosen and tuned. Two training categories namely unimodal and multimodal supervised learning were utilized in current approach.

C. Model framework

This section describes the visual and audio encoder pipelines implemented in our study. It also highlights the feature extraction process for each modality and how the learned features are obtained from the input representations. For visual modality, *ResnetLSTM* based encoder was implemented using *Multibench* framework. The frames from videos are fed to Resnet-18 model which extracts the features. These features were later fed to Bi-LSTM model which encoded features into one embedding(further details in Video encoder subsection). The inspiration of Audio encoder was derived from the baseline Ego4D work. The encoder extracted MFCC frequency map of the audio segments which essentially served the purpose of audio feature extraction. These features were further fed to a Resnet-18 network.

1) Visual Encoder: The visual encoder is developed by employing Resnet LSTM model as provided by multibench framework, see Figure 5.

Visual feature extraction : To extract high level spatial features, Resnet-18 pre-trained network was employed. For this purpose the final classification softmax layer is removed and the features are obtained from the fully connected layer of 1000 feature dimension which is later fed to a Bi-LSTM network. These features represent the visual content of each frame and spatial relationships present in the image.

Capturing temporal information : To effectively capture tem-

poral information from the video data, Bi-LSTM model with 32 hidden layers was used. By considering the sequence of visual features from Resnet-18, the Bi-LSTM can significantly learn temporal pattern and dependencies in the video data. Also, being capable of processing sequences bidirectionally the network can capture both past and future frames effectively, thereby enhancing model's learning.

To counter the possibilities of overfitting during training, to the model a dropout based regularization technique was employed[34][35]. A dropout factor of 0.20 was utilized for proposed system which significantly facilitated in countering the overfitting issues.

2) Audio Encoder: Audio encoder pipeline consists of sequential model of MFCC feature extraction block and Resnet-18 pre-trained model.

Audio feature extraction : To extract relevant features from the prepared audio segments MFCC feature map is extracted every 10ms with 25ms window length [1] [30]. These features are later fed to a ResNet-18 network. A detailed discussion regarding MFCC feature extraction process is discussed in *Appendix* C.

ResNet-18 for Audio Processing : When MFCCs are used as inputs to ResNet-18, the network's convolutional layers learn to process these coefficients as spatial features. Thus, it sufficed the role of extracting meaningful high-level representations of the audio features. ResNet also facilitated in effectively capturing both local and global patterns present in the MFCCs.

D. Fusion techniques

A detailed background knowledge for the fusion techniques is discussed in *Technical background* II-A. As discussed earlier, the fusion techniques were implemented as per the approach developed in Multibench. Figure 6 represents the methodology for the implementation of Early fusion technique. The features from individual pipelines were fused in a common shared feature representation. A binary classification was later employed to determine the classification of instances of talking to me or not.

Figure 7 represents the methodology for the implementation of Late fusion technique. It is implemented by recovering scores from the softmax layers of the unimodal networks. These scores are later fused and finally a softmax based classification is employed to obtain desired classification results.

To fuse the classifier scores from individual modalities a naive-product based approach was chosen (as developed in Multibench framework). In essence, it is the multiplication of the probabilities assigned by each classifier for a particular class to obtain the combined probability for that class. Mathematically it is represented as :

$$F_{Prod} = \frac{\Pi_{i=1}^{nc} L_i}{\Pi_{i=1}^{nc} L_i + \Pi_{i=1}^{nc} (1 - L_i)} \tag{3}$$

Where L_i is probability scores obtained by the deep models $(i = 1, 2, ..., nc)$ for our approach $nc = 2$, corresponding classifier scores for 2 modalities.

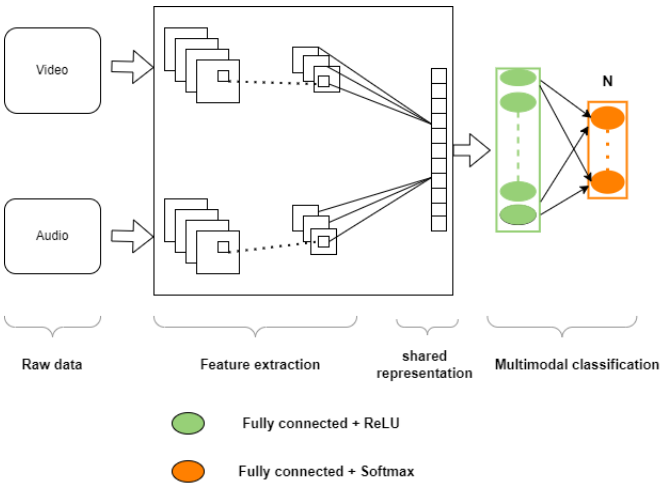


Fig. 6: Early fusion model architecture, here N denotes number of classes, for current case N = 2.

This combined probability distribution is later utilized to make decisions, essentially done by selecting the class with highest combined probability.

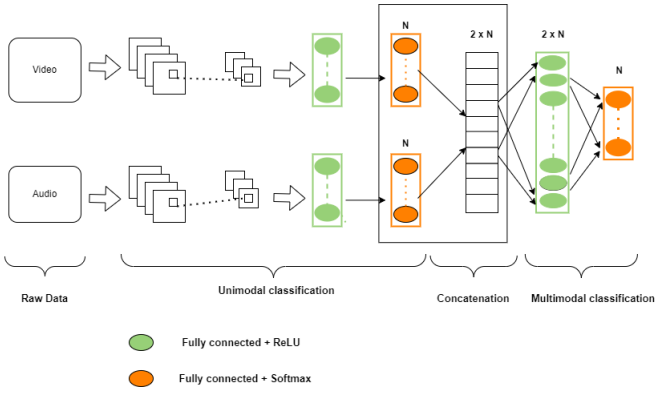


Fig. 7: Late fusion model architecture, here N denotes number of classes, for current case N = 2.

IV. Experimental Setup

A. Dataset

The dataset used in our study is Ego4D dataset, out of 3,670 hours of video in Ego4D, approximately 764 hours of data containing conversational content was pertinent to the Audio-Visual diarization and Social benchmark tasks. The system proposed here utilizes data of 572 social interaction videos. Out of which 389 clips were split for training, comprising 32.4 hours in total, 50 clips (4.2 hours) and 133 clips (11.1 hours) were dedicated to validation and testing sets, respectively. Figure 2 highlights the annotations provided by the baseline work. It also gives a visualization of how annotated data was developed by taking into account the tracked face sequences, active speaker detection and recognizing voice activities in the video frames. This annotated data marks the initial steps of

dataset preparation which is thoroughly discussed in *Appendix* D

B. Comparison with other baselines

1) Ego4D Baseline - Talking to me (TTM): For starting point of experimentation, the results of the baseline work of Ego4D Talking to me were reproduced. This also served as a benchmark to compare the results of experiments which is outlined in further sections. The process of developing TTM segments from baseline annotations is discussed in depth in *Appendix* D. These segments govern the further data preparation process and eventually visual, audio and targets are prepared from the pre processed data.

2) Multibench framework: As mentioned in methodology section, system proposed in this work is implemented using Multibench framework. The prepared visual and audio data is fed to respective encoders. To achieve performance improvements compared to baseline work, a different model architecture for visual encoder is explored. Furthermore, two techniques Late and Early fusion are implemented. Training is done using a supervised learning implementation present in the framework, and optimization of parameters at different stages of multimodal learning has also been implemented.

C. Performance evaluation metrics

1) Quantitative metrics: To evaluate the performance of the model, and also to compare with the baseline work mean average precision (mAP) and Top-1 accuracy (accuracy) were used as validation metrics. Accuracy is calculated using Equation 4. Mean average precision (mAP) is calculated using Equation 5, where Average Precision is calculated as the weighted mean of precisions at each threshold; the weight is the increase in recall from the prior threshold.

Mean Average Precision is the average of AP of each class.

$$Accuracy = \frac{Correct\ predictions}{Total\ predictions} \tag{4}$$

$$mean\ average\ precision\ (mAP) = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{5}$$

Where AP = average precision at different recall levels
N = number of classes

Precision and Recall as discussed above are calculated using Equation 6 and 7 respectively.

$$Precision = \frac{True\ Positive}{True\ positive + False\ Positive} \tag{6}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{7}$$

Additionally, AUPRC value for best performing epoch is also provided. To evaluate the missclassification, confusion matrix plot has been plotted to highlight the samples classified as True negative, False negative, True positive and False positive.

2) Qualitative metrics: To assess the quality of extracted features t-SNE (t-distributed stochastic neighbor embedding) dimensionality reduction technique was implemented [36]. It is used for visualizing high-dimensional data in lower-dimensional space. Following aspects were considered while performing t-SNE evaluation

- *Feature separability* : If inter-class features are well distinguishable, this suggests that the features are able to capture discriminate information between classes.
- *Outlier identification*: Outliers or data points indicate features not conforming to general patterns. Most of the times it also indicates the presence of noise or anomalies in the data.

D. Implementation details

For development PyTorch machine learning framework was used throughout. All the experiments were conducted on High performance cluster (HPC/Slurm) hardware on the server with availability of 4x NVIDIA A40/48G GPUs. The hyperparameters involved with the experiments are as follows: Early stopping with patience score of 7, such that if accuracy failed to improve in 7 epochs, the training stopped. Learning rate was chosen to be 0.00005, with the use of Adam as the optimizer. Considering classification problem at hand cross entropy loss was chosen as the loss function with class weights as [0.266, 0.734] similar to the baseline approach. Table I indicate the hyperparamaters used at different levels of system development.

Component	Model	Parameters	Value
Visual Encoder	Resnet-18+ Bi-LSTM	ResNet version LSTM layers LSTM hidden layers Dropout Video encoder output dim	18-layer 2 32 0.20 64
Audio Encoder	Resnet-18+ MFCC	ResNet version MFCC num filters MFCC output dim	18-layer [32, 64, 128, 256] 512
Classification Head	Linear	Xavier initialization	True
Fusion	Late fusion Early fusion	- output dim	- 576
Training	Unimodal LF Supervised EF Supervised	Loss Num epochs Optimizer optimizer weight decay Learning rate class weights	Cross entropy 8/10/15 Adam 0.01 5e-4 [0.266, 0.734]

TABLE I: Table of hyperparameters for training on Ego4D dataset.

V. Results

This section presents the results and performance comparison of the experiments. Two experiments corresponding to two explored fusion techniques (early and late fusion) were performed. Eventually a performance comparison is

established analyzing the better alternative out of the two. Furthermore, a comparison of the proposed techniques with baseline work is established. A separate set of experiments taking unimodal models for video and audio were also performed to give a broader perspective on the the advantages of multiple modalities over unimodal counterparts.

A. Qualitative analysis of the results

To qualitatively access model's performance it was required to analyze the quality of extracted features using different model architecture. For this purpose, a t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization of features was employed. To analyze how extracted features have improved a comparison between baseline and multibench implemented model was performed.

Figure 8 depicts feature evaluation across three model versions : Baseline visual encoder model, Multibench visual encoder model, Baseline audio encoder model. In subplot (a), the features in their respective classes 0 and 1 are considerably overlapped. While in (b) the extracted feature quality seem to have improved as features in their respective classes are more distinguishable. Another important thing to notice is that there are more outliers present in (a) compared to (b). These outliers may mostly comprise of noise or miss-classified features. Taking these observations into account, model architecture provided by Multibench framework has seemingly performed better than the baseline model. However, the features extracted by audio encoder model were significantly better and lot more distinguishable in their respective classes than any of the video encoder alternative. This highlights the robustness of feature extraction process in audio modality thus giving better and more distinguishable features. Another possible reason behind these observations could be better data preprocessing methods employed for audio modality in comparison to video which in turn points out to the differences in robustness between two modalities. An issue in robustness of features could most possibly be related to incompetency in the tracking algorithms from which the data was actually prepared. The occurrences of occlusion in visual modality was fairly evident when visualizing the data, which can potentially hamper the quality of extracted features. Thus, better tracking methods could in turn improve the quality of features being extracted.

B. Quantitative analysis of the results

Table II depicts the results of the performed experiments. For unimodal experiment with visual modality, *ResnetLSTM* based encoder was employed. This model generated a mean accuracy score of 52.67% and a mAP score of 53.56%, which was better than the random guess model.

For audio modality, model architecture presented in Baseline was utilized as mentioned in model framework section of this report. This model generated a mean accuracy score of 55.42% and mAP 55.20%. It should be noted that the audio modality performed better than its video counterpart, which is synchronous with the finding observed in qualitative analysis subsection. This is mostly due to better features extracted from audio modality in comparison with the video modality. This points out the robustness issues of video modality and thus necessitates improving the quality of feature extraction.

The experiments for Early and Late fusion, also implemented using Multibench, generated a mAP score of 60.17% and 62.02% respectively, on the validation dataset. It should be noted how multimodal training performed better than its unimodal constituents and thus presented some significant performance improvements.

- An intuitive reason behind this observation is that multimodal models by combining different modalities gain a more comprehensive understanding of the dataset, capturing dependencies and patterns which unimodal model at times are incapable of.
- The results also highlight how multimodal models were more robust than audio and video unimodal models. This is due to the fact that multimodal models are more robust to noise or missing information in individual modalities
- By learning representations from multiple modalities, the model can capture more abstract and generalizable features. This can enhance the model's ability to generalize to unseen data or tasks.

1) Early vs late fusion: Early fusion generated Top-1 mean accuracy of 58.27% and a mAP score of 60.17%. On the other other hand, Late fusion generated Top-1 mean accuracy of 60.75% and a mAP score of 62.02%. Clearly late fusion performed better, following reasoning explains this observation:

- Late fusion performs better in scenarios where modalities have a component of asynchronicity between them. Due to several occurrences of malformed data, both video and audio modalities are asynchronous at times. Late fusion counters this by allowing the model to learn form each modality independently before combining them.
- Late fusion tends to be more resilient to noise or variability within individual modalities, since it does not directly combine features from different modalities.

To account for the missclassification by the model, confusion matrix was also plotted for the two multimodal experiments, depicted in Figure 9 and Figure 10. High number of True positives account for higher accuracy values of Late fusion over Early fusion.

2) Comparison with the Baseline: Top-1 accuracy and mAP results are obtained for segment based batches, this is similar across both Baseline implementation and our proposed experiments. Both fusion techniques achieved a mAP score significantly better than the baseline highlighting the performance improvements of our proposed work over baseline.

Figure 11 depicts a plot of accuracy values with epochs across Late and Early fusion experiments. As it can be seen, Late fusion performed better than Early fusion, thereby achieving higher accuracy values. A separate set of experiments considering accuracy values for all 4 experiments is depicted in Figure 18 .It should be noted that accuracy values for unimodal visual experiment is co-terminus with the observed qualitative and quantitative analysis done in previous sections. Since

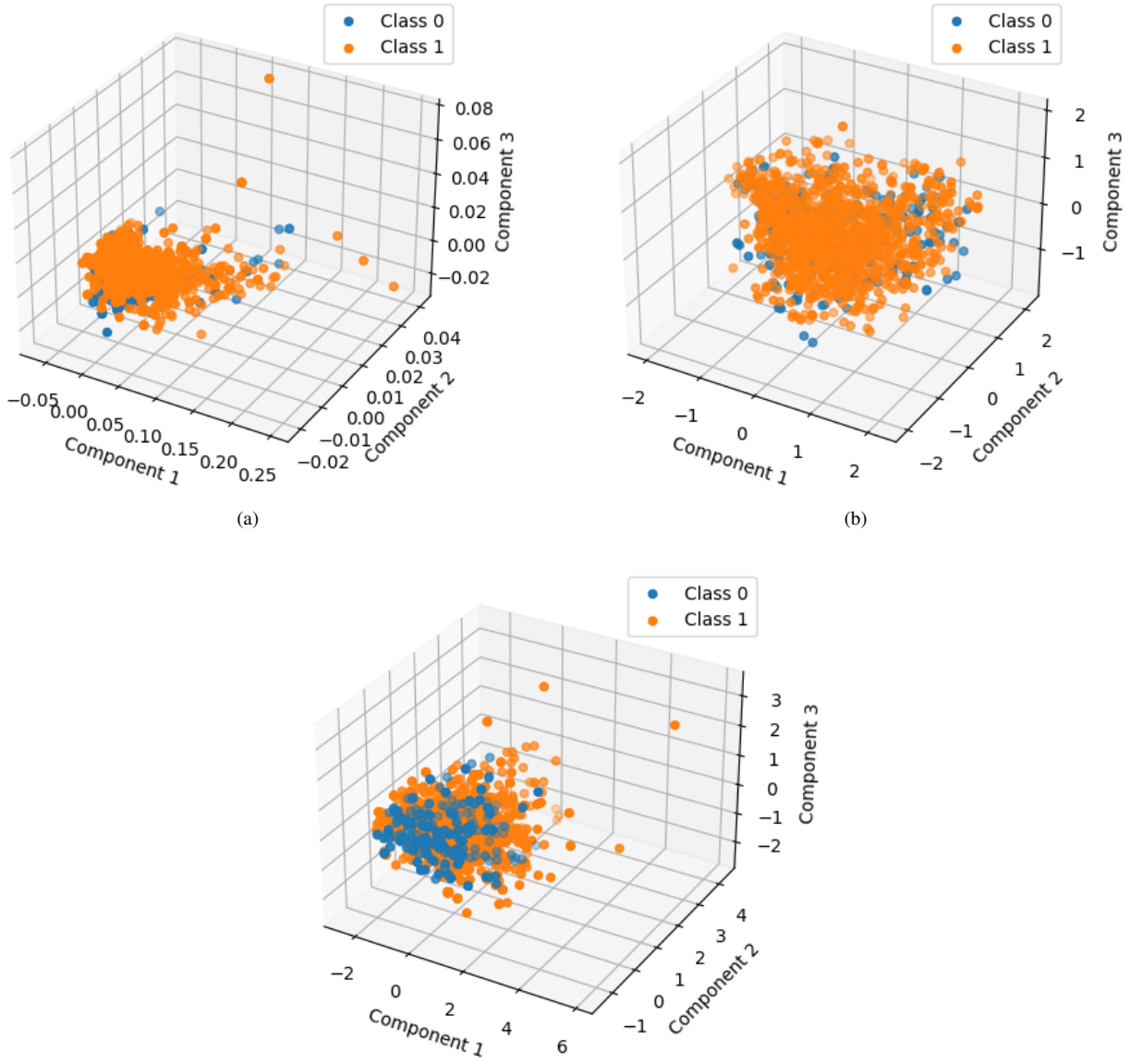


Fig. 8: Plot indicates t-SNE visualization of extracted features in respective classes (0 and 1), subplot (a) using *Baseline* visual encoder model architecture, subplot (b) using *ResnetLSTM* based visual encoder proposed in *Multibench*, subplot (c) using audio encoder model provided in *Baseline* work and used in current system as well.

the video modality suffers from robustness issues (in feature extraction and data preparation), the accuracy values show a slow improvement with iterations compared to multimodal and unimodal audio experiments. It should also be noted that multimodal methods outperformed its unimodal counterparts, highlighting the model's ability to have a more comprehensive understanding of data features, and capturing more patterns by mitigating robustness issues in individual modalities.

A visualization of the results is provided in *Appendix* H. The visualization is performed for the experiment of Late fusion for the frame sequences obtained from the validation subset. The prediction scores for the frame sequences are generated when the output from the model (softmax score) is fed to the post processor (derived from the baseline work, see *Appendix* E). Ground truth values for the corresponding frame sequences are also generated in the post processor. Model's classification and missclassification upon correlating prediction scores with the ground truth can be observed in the Figure 17.

It should also be noted that the experiments have only been performed on validation dataset. The baseline work of Ego4D TTM is specified under the challenge of Talking to me social interactions given by EvalAI community. The ground truth

Experiments	Top-1 Accuracy	mAP	AUPRC for best epoch
Unimodal -Visual	0.5267 ± 0.016	0.5356	0.5522
Unimodal -Audio	0.5542 ± 0.014	0.5520	0.5631
Multimodal-Early fusion	0.5827 ± 0.030	**0.6017**	0.6189
Multimodal-Late fusion	0.6075 ± 0.046	**0.6202**	0.6415
Baseline TTM Ego4D [1]	0.6431	0.5650	–
Random Guess	0.4989	0.50	–

TABLE II: Performance evaluation across different experiments.

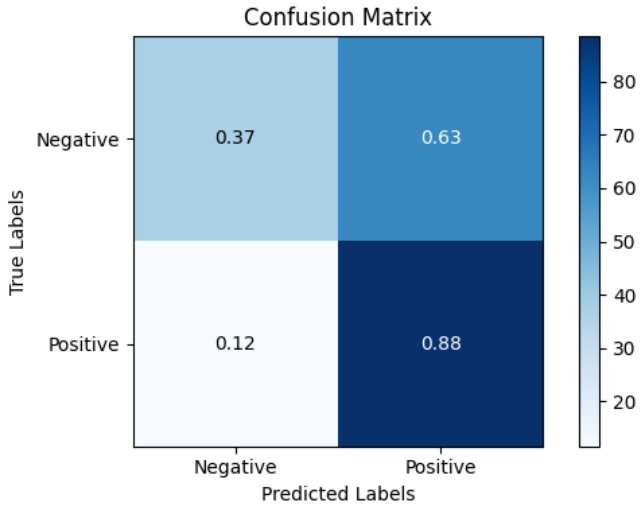


Fig. 9: Confusion matrix plot for Late fusion.

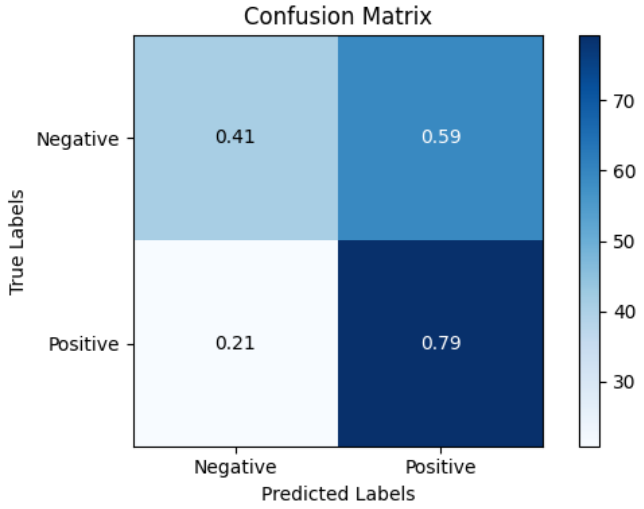


Fig. 10: Confusion matrix plot for Early fusion.

labels have not been provided for the testing dataset and the results from the baseline work were sent for evaluation on

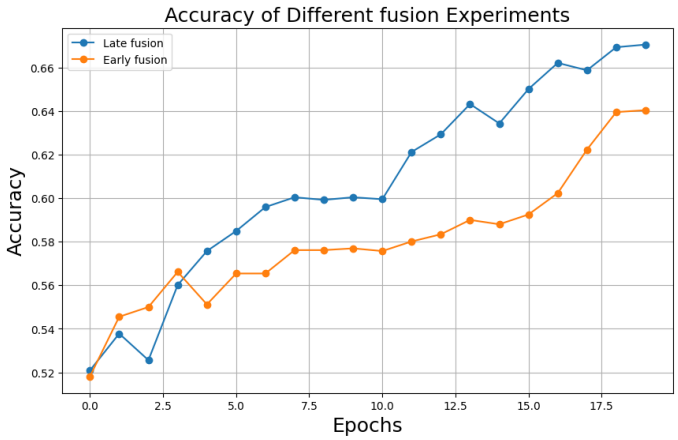


Fig. 11: Accuracy plot for Early and Late fusion experiments depicting accuracy values with epochs.

the testing dataset. Since the dataset is substantially large for social interactions, running experiments on validation dataset also give near comparable results and should be quite valid as well.

VI. Discussion

A. Answer to research questions

The research questions have been adequately answered in following manner:

RQ: How does multiple modalities affect the prediction of *Talking to me* based social interactions from egocentric point of view?

Multiple modalities, given the dataset Ego4D in hand, presented significant performance improvements when classifying *Talking to me* based social interactions as highlighted below.

SRQ1: How does the implementation of the *Multibench* framework contribute to performance enhancements, if any compared to the *Ego4D* baseline in classifying Talking to me based social interactions?

Developing multimodal models using *Multibench* and employing optimization techniques provided by the framework at numerous levels, namely fusion techniques, model alternatives, optimization and regularization techniques, facilitated in achieving significant performance improvements compared to the baseline work of Ego4D *Talking to me* social interactions. Observed performance improvements were mAP score of **62.02%** for Late fusion and **60.17%** for Early fusion developed using *Multibench* over **56.50%** for baseline Ego4D.

SRQ2: Which fusion technique provides superior performance when integrating modalities?

Among the fusion techniques considered, *Late fusion* performed better than *Early fusion* by a mAP score of **62.02%** for Late fusion over **60.17%** for Early fusion. This highlights performance improvements offered by Late fusion given the dataset (Ego4D) and modalities at hand, while also considering the robustness differences present in the modalities.

B. Countering overfitting during training

During earlier stage of experimentation overfitting was observed as the loss values depicted significant fluctuations on validation dataset. As a counter measure regularization techniques were implemented in model frameworks. For video encoder, a Resnet bi-LSTM network with 32 hidden layers was employed using multibench framework. To apply regularization technique of dropout with rate 0.20 was utilized. Thereby simplifying the model and facilitating the model to make proper generalizations about the input data. By essentially dropping neurons from model layers, dropout forced model to not depend on just one neuron or limited set of features thereby diversifying the approach and taking relevant information from more set of features into consideration.

C. Challenges faced during implementation

As mentioned, the dataset consisted of videos which made the training process extremely slow. Due to extremely large nature of the dataset the experiments were performed only for limited number of epochs (maximum 20 epochs). This significantly hindered training for more number of epochs.

Another important thing to highlight is the dynamic nature of the dataset. The data as mentioned in previous section is prepared taking segments of tracked information and ground truth into account. Due to the manner in which data was prepared, it was not feasible to develop self made batches, thereby making the training process really slow.

D. Limitations of the proposed system

Taking the baseline work into account the accuracy values in general have a significant room to improve. Much of it could be due to the manner in which data is developed in the first place. Better tracking methods could in turn mean better data quality to apply machine learning algorithms to. The segments developed from such efficient tracking methods could improve the results.

The cross validation approaches like K nearest neighbour(KNN) cross validation were not implemented in this study due to following reasons, firstly, the test data split was not considered due to non availability of the ground truth data from the baseline work of Ego4D for the testing dataset. Furthermore, the extremely dynamic nature of dataset which varied throughout, made it really tough to implement the KNN algorithm itself. Also, due to immensely large running duration, it was not feasible to perform cross validation and running several fold iterations. Considering these factors, it was decided to perform other validation metrics and to skip KNN cross validation.

The results discussed in previous section helps us identify the issues of robustness in individual modalities, While implementing malformed data occurrences were encountered at few occasions, which points out to the fact that there were problems in the way data was prepared from the baseline work.

E. Future work and possibilities

While the fusion techniques discussed do show promising results when comparing to baseline work, but even better techniques do exist for multimodal fusion. Within multibench framework the fusion techniques of Gradient blend and MFAS(multimodal fusion architectural search) exist which have proven to overcome the shortcomings of common fusion techniques. In gradient blend the concept of weighted modalities is taken into account which can facilitate in fusing modalities giving right weightage to the more robust modality [13][12]. MFAS method searches the architectural space and specifies the layers in the model when multimodal fusion should be employed [15][37][14]. Thus this technique presents a robust manner to fuse modalities, thereby proving to be an efficient multimodal fusion method for several datasets. To make the best use of the framework these fusion techniques can be explored with Ego4D. Although that would still be from investigative point of view, but the underlying concepts of these methods address the shortcomings of general fusion techniques.

Several feature extraction techniques as discussed in scientific background section indicate employing better feature extraction techniques [22][31][23]. By better feature extraction methods for image and audio modality, performance of unimodal models can significantly improve and consequently for the multimodal approaches as well. A possible future work can also entail developing a trade off between the feature extraction techniques and analyzing the optimal alternative.

Yet another possibility that exists is to speed up the process of training .To achieve this, dynamic nature of data needs to be dealt with. This might be developed by performing data cleaning operations and that model could be trained using self made batches.

VII. Conclusion

This study devised a deep learning approach to classify the occurrences of *Talking to me* based social interactions from egocentric point of view. Inputs in the form of video and audio modality were considered for developing such a system. The system explored the applicability of a Multimodal framework called Multibench with Ego4D dataset. By implementing the optimization methods provided by the framework at numerous levels ranging from model architecture, fusion techniques and optimization objectives, a performance comparison with the baseline was established. Out of two fusion techniques explored in this paper, Late fusion performed better than Early fusion with a mAP score of 62.02% over 60.17%. Additionally, both these techniques reported a mAP score better than the baseline, highlighting the performance improvements over baseline Ego4D. The paper also discuss the bottlenecks of the current data preparation and feature extraction strategies, thereby giving suitable guidelines about the improvements that can be employed in future.

Acknowledgement

I would like to express my sincere thanks to Data Management and Biometrics (DMB) research chair and University of Twente to give me an opportunity to carry out my research work on this topic.

I would like to thank my Thesis supervisor Dr. Estefanía Talavera Martínez for her valuable guidance throughout the course of the research work, which helped me immensely to put my plans into action. I also wish to express my gratitude to my supervisors and committee members, Dr.Ir. Luuk Spreeuwers and Dr. Faizan Ahmed for their valuable feedback and supervision which helped me immensely to complete my work in the right manner. I also would like to extend my gratitude to Prof. Willams de Lima,(UFPE),Brazil and Prof. Federico González Brizzio, Spain, for their valuable pointers and technical guidance which helped me to develop effective solutions.

References

[1] K. Grauman, A. Westbury, E. Byrne, *et al.*, *Ego4d: Around the world in 3,000 hours of egocentric video*, 2022. arXiv: 2110.07058 [cs.CV].

[2] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–6. DOI: 10.23919/FUSION45008.2020.9190246.

[3] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," Jan. 2011, pp. 689–696.

[4] P. P. Liang, Y. Lyu, X. Fan, *et al.*, "Multibench: Multiscale benchmarks for multimodal representation learning," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[5] G. Melotti, C. Premebida, N. M. M. d. S. Goncalves, U. J. C. Nunes, and D. R. Faria, "Multimodal cnn pedestrian classification: A study on combining lidar and camera data," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3138–3143. DOI: 10.1109/ITSC.2018.8569666.

[6] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, *Multi-modal sensor fusion for auto driving perception: A survey*, 2022. arXiv: 2202.02703 [cs.CV].

[7] P. P. Liang, Y. Lyu, X. Fan, *et al.*, *High-modality multimodal transformer: Quantifying modality interaction heterogeneity for high-modality representation learning*, 2023. arXiv: 2203.01311 [cs.LG].

[8] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications*, vol. 32, no. 6, p. 121, Sep. 2021, ISSN: 1432-1769. DOI: 10.1007/s00138-021-01249-8. [Online]. Available: https://doi.org/10.1007/s00138-021-01249-8.

[9] I. Gallo, G. Ria, N. Landro, and R. L. Grassa, "Image and text fusion for upmc food-101 using bert and cnns," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2020, pp. 1–6. DOI: 10.1109/IVCNZ51579.2020.9290622.

[10] M. Alfaro-Contreras, J. J. Valero-Mas, J. M. Iñesta, and J. Calvo-Zaragoza, "Late multimodal fusion for image and audio music transcription," *Expert Systems with Applications*, vol. 216, p. 119 491, 2023, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.119491. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422025106.

[11] S. Chen and Q. Jin, *Multi-modal conditional attention fusion for dimensional emotion prediction*, 2017. arXiv: 1709.02251 [cs.CV].

[12] W. Wang, D. Tran, and M. Feiszli, *What makes training multi-modal classification networks hard?* 2020. arXiv: 1905.12681 [cs.CV].

[13] S. Chen and B. Li, "Towards optimal multi-modal federated learning on non-iid data with hierarchical gradient blending," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, London, United Kingdom: IEEE Press, 2022, pp. 1469–1478. DOI: 10.1109/INFOCOM48880.2022.9796724. [Online]. Available: https://doi.org/10.1109/INFOCOM48880.2022.9796724.

[14] Y. Yao and R. Mihalcea, "Modality-specific learning rates for effective multimodal additive late-fusion," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1824–1834. DOI: 10.18653/v1/2022.findings-acl.143. [Online]. Available: https://aclanthology.org/2022.findings-acl.143.

[15] Y. Zhou, P. Wang, S. Arik, *et al.*, *Epnas: Efficient progressive neural architecture search*, 2019. arXiv: 1907.04648 [cs.LG].

[16] Z. Xu, D. R. So, and A. M. Dai, *Mufasa: Multimodal fusion architecture search for electronic health records*, 2021. arXiv: 2102.02340 [cs.LG].

[17] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, *Efficient low-rank multimodal fusion with modality-specific factors*, 2018. arXiv: 1806.00064 [cs.AI].

[18] X. Miao, X. Zhang, and H. Zhang, "Low-rank tensor fusion and self-supervised multi-task multimodal sentiment analysis," *Multimedia Tools and Applications*, Jan. 2024, ISSN: 1573-7721. DOI: 10.1007/s11042-023-18032-8. [Online]. Available: https://doi.org/10.1007/s11042-023-18032-8.

[19] S. Park and E. Choi, *Multimodal transformer with a low-computational-cost guarantee*, 2024. arXiv: 2402.15096 [cs.LG].

[20] H. Akbari, L. Yuan, R. Qian, *et al.*, *Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text*, 2021. arXiv: 2104.11178 [cs.CV].

[21] Y. Li, R. Quan, L. Zhu, and Y. Yang, *Efficient multimodal fusion via interactive prompting*, 2023. arXiv: 2304.06306 [cs.CV].

[22] S. Ishikawa and J. Laaksonen, "Comparing and combining unimodal methods for multimodal recognition," in *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2016, pp. 1–6. DOI: 10.1109/CBMI.2016.7500253.

[23] F. T. Ito, H. de Medeiros Caseli, and J. Moreira, "The effects of unimodal representation choices on multimodal learning," in *International Conference on Language Resources and Evaluation*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:21719536.

[24] Z. Xue, Y. Song, K. Grauman, and L. Torresani, *Egocentric video task translation @ ego4d challenge 2022*, 2023. arXiv: 2302.01891 [cs.CV].

[25] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, "Jointly learning energy expenditures and activities using egocentric multimodal signals," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6817–6826. DOI: 10.1109/CVPR.2017.721.

[26] A. Zehtab-Salmasi, A.-R. Feizi-Derakhshi, N. Nikzad-Khasmakhi, M. Asgari-Chenaghlu, and S. Nabipour, "Multimodal price prediction," *Annals of Data Science*, vol. 10, no. 3, pp. 619–635, Apr. 2021, ISSN: 2198-5812. DOI: 10.1007/s40745-021-00326-z. [Online]. Available: http://dx.doi.org/10.1007/s40745-021-00326-z.

[27] F. Vossebeld, *Towards understanding social interactions through audio signals*, Jul. 2022. [Online]. Available: http://essay.utwente.nl/91976/.

[28] S. Felicioni and M. Dimiccoli, *Interaction-gcn: A graph convolutional network based framework for social interaction recognition in egocentric videos*, Apr. 2021.

[29] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, "Speaker recognition using mel frequency cepstral coefficients (mfcc) and vector quantization (vq) techniques," in *CONIELECOMP 2012, 22nd International Conference on Electrical Communications and Computers*, 2012, pp. 248–251. DOI: 10.1109/CONIELECOMP.2012.6189918.

[30] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022. DOI: 10.1109/ACCESS.2022.3223444.

[31] S. A. Medjahed, "A comparative study of feature extraction methods in images classification," *International Journal of Image, Graphics and Signal Processing*, vol. 7, pp. 16–23, Feb. 2015. DOI: 10.5815/ijigsp.2015.03.03.

[32] S. Schneider, A. Baevski, R. Collobert, and M. Auli, *Wav2vec: Unsupervised pre-training for speech recognition*, Apr. 2019.

[33] Y. Wang, Y. Li, Y. Song, and X. Rong, "The influence of the activation function in a convolution neural network model of facial expression recognition," *Applied Sciences*, vol. 10, no. 5, 2020, ISSN: 2076-3417. DOI: 10.3390/app10051897. [Online]. Available: https://www.mdpi.com/2076-3417/10/5/1897.

[34] R. vij, *Combating overfitting with dropout regularization*, Mar. 2023. [Online]. Available: https://towardsdatascience.com/combating-overfitting-with-dropout-regularization-f721e8712fbe.

[35] P. Baheti, *What is overfitting in deep learning [+10 ways to avoid it]*, Dec. 2021. [Online]. Available: https://www.v7labs.com/blog/overfitting.

[36] L. (Shin, *Ch 5. t-sne plots as a human-ai translator*, Sep. 2021. [Online]. Available: https://medium.com/@lucrece.shin/chapter-4-using-t-sne-plots-as-human-ai-translator-c5ef9c2f2fa4.

[37] X. Ding, T. Han, Y. Fang, and E. Larson, "An approach for combining multimodal fusion and neural architecture search applied to knowledge tracing," *Applied Intelligence*, vol. 53, no. 9, pp. 11 092–11 103, May 2023, ISSN: 1573-7497. DOI: 10.1007/s10489-022-04095-x. [Online]. Available: https://doi.org/10.1007/s10489-022-04095-x.

APPENDIX A
MULTIBENCH FRAMEWORK IMPLEMENTATION PSEUDO CODE

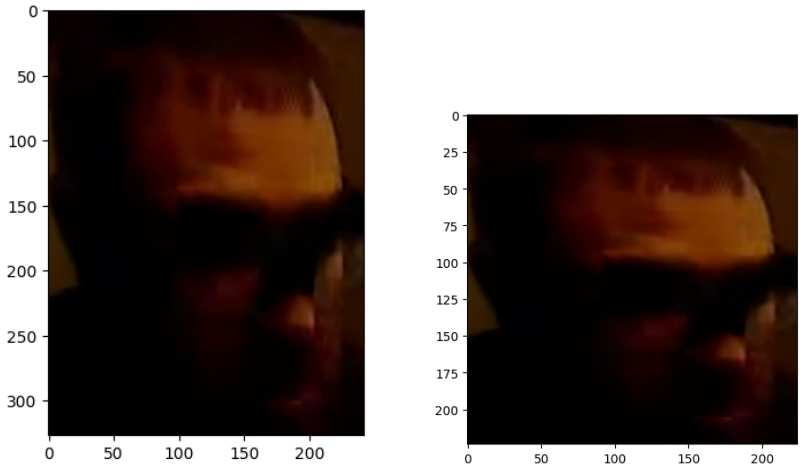
Given below is a PyTorch implementation pseudo code using Multibench Figure 12, indicating the plug in multimodal methods at numerous stages of the system ranging from data preparation, Unimodal/multimodal models, Classification head ,fusion paradigms, Optimization objectives, Training structures and performance evaluation.

```
from Multibench.datasets.get_data import get_dataloader
from ego4d.models import video_encoder, audio_encoder
from Multibench.fusions.common_fusions import Early_fusion,
Late_fusion
from Multibench.model.common_models import Linear
from Multibench.training_structures.Supervised_learning import train,
test
# loading Ego4D dataset
traindata, validdata, testdata = get_dataloader('ego4d_social_
interaction')
out_channels = 512
# defining ResNet and Transformer unimodal encoders
encoders = [video_encoder(in_channels=1, out_channels,
LSTM_layers=2),
audio_encoder(in_channels=1, out_channels)]
# defining Late_fusion fusion layer
fusion = Late_fusion([out_channels*2])
# defining a classification head
classifier = Linear(out_channels*2, 2, 'xavier_init' = True)
# training using Supervised learning method
model = train(encoders, fusion, classifier, traindata, validdata,
epochs=15, optimtype=torch.optim.Adam, lr=5e-4, weight_decay=0.01)
# testing
performance, complexity, robustness = test(model, testdata)
```

Fig. 12: PyTorch pseudo code using Multibench framework

APPENDIX B
BOUNDING BOX INPUT REPRESENTATION

Figure shows the region of interest extracted from the video frames by performing bounding box based cropping. Furthermore, it also shows the resized frame dimension 224x224 to align the visual modality with the model's input.



(a) (b)

Fig. 13: On left (a) indicates frame in video with bounding box based cropping and focusing on the face ,region of interest (ROI). On right (b) indicates the resized frames to dimension 224x224.

APPENDIX C
AUDIO FEATURE EXTRACTION STEPS

MFCC audio features were extracted from audio data by employing following steps (derived from the baseline work):

- **Step1: Pre-emphasis** : To balance the frequency spectrum, a pre-emphasis filter is applied to the signal. It boosts the higher frequencies and reduces lower ones, implemented using a first-order FIR filter.

- **Step2: Windowing** : Windowing operation performs the job of slicing audio wave forms into sliding frames. Mathematically, it is represented as

$$x[n] = w[n] * s[n] \tag{8}$$

where, $x[n], s[n]$ represents sliced frame and original audio segment respectively. For current system, hamming window was employed and is represented by $w[n]$.
Equation for Hamming window is as follows:

$$w[n] = (1 - \alpha) - \alpha * cos(2\pi n/L - 1) \tag{9}$$

for Hamming $\alpha = 0.46164$, L = Window length
Through numerous studies [29][30] it has been identified Hamming window performs better than a rectangular window and thus it was implemented in current system as well.
- **Step3: Fourier transform**: This operation converts time domain information to frequency domain.
- **Step4: Mel filter bank**: The power spectrum obtained from the FFT is passed through a Mel filter-bank. Here, Mel scale is used to divide the frequency range into equally spaced intervals. The following equation converts frequency in Hz to Mel:

$$F(mel) = 1127ln(1 + f/700) \tag{10}$$

- **Step5: Discrete Cosine Transform (DCT)** : After obtaining logarithm of filter-bank outputs, Discrete cosine transform is applied to de-correlate the filter-bank energies. This step performs the function of converting Mel spectrum to time domain using cosine transform. Typically, the lower-frequency coefficients are retained as they contain most of the information. Consequently a sequence of acoustic vectors is obtained corresponding to each input expression.
- **Step6: Delta Energy** : Delta coefficients capture rate of change of MFCCs over time. Delta coefficients are computed by taking the differences between consecutive MFCC vectors. They provide information about dynamic changes in audio signals.

APPENDIX D
DATA PREPARATION METHODOLOGY

Figure 14 indicates the data preparation methodology which is derived from the baseline work and reused in our approach as well. The baseline annotations are utilized to develop JSON segments, this is essentially the tracked segments of each person in the videos. These tracked segments are correlated with the ground truth annotations thereby developing Talking to me based segments. While preparing the dataset, label 0 corresponds to the instances when Target values (Tags) are none. On the other hand, label 1 corresponds to Target values not None present in the ground truth.

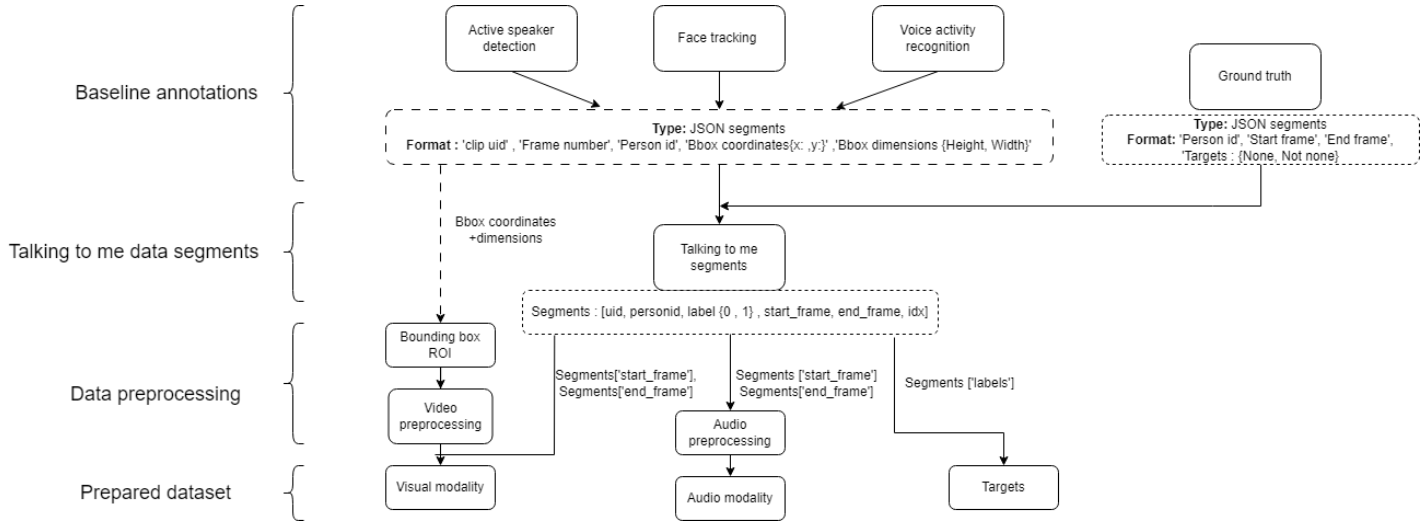


Fig. 14: The flow diagram depicts the manner in which data is prepared (for training and validation split), derived from the baseline work of Ego4D, reused in our approach as well. The figure gives a detailed overview of data preparation steps ranging from annotations provided by the baseline to prepared video , audio and targets ready to be fed to the model. Dashed boxes indicate the type and format of data at a particular stage.

APPENDIX E
VALIDATION SET POST PROCESSOR IMPLEMENTATION

Baseline work of Ego4D employs a post processor for the validation dataset. This post processor generates softmax prediction scores csv file corresponding to segment batches, simultaneously it also generate ground truth csv files obtained using target values. This post processor implementation is used for our experiments as well to visualize sequence batch outputs, see Figure 15.

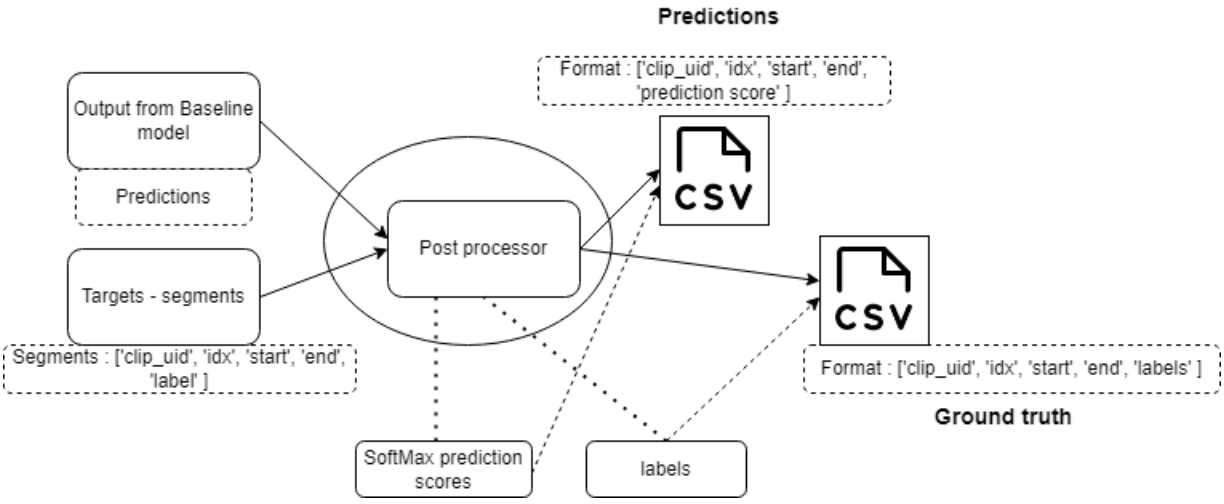


Fig. 15: Post processor implementation. Dashed boxes indicate the format of data present at a particular stage.

APPENDIX F
TRAINING AND VALIDATION LOSS CURVES ACROSS DIFFERENT EXPERIMENTS

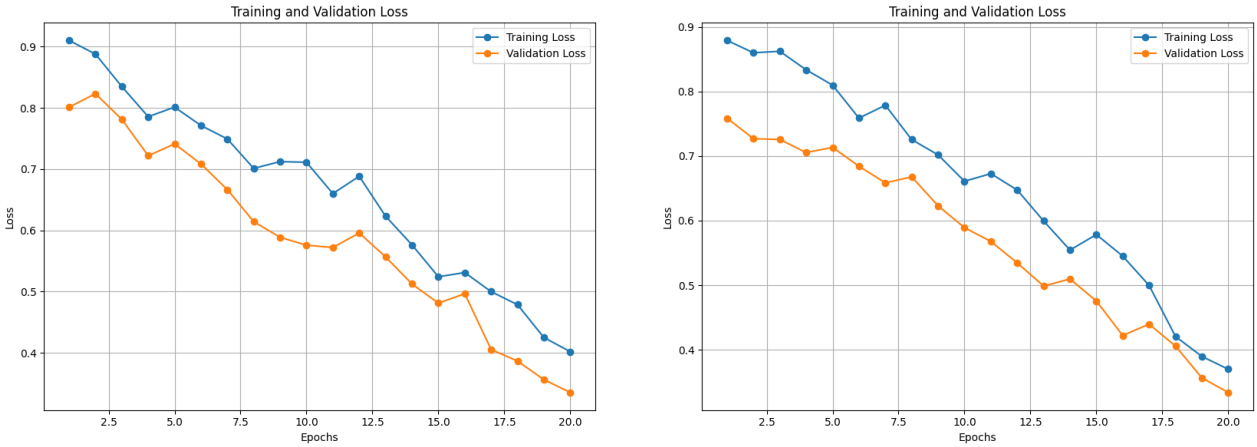
Figure 16 indicates a plot of loss values across different experiments, on analyzing loss value plots for training and validation it can be observed that the loss values show a general decreasing trend. It should also be noted that the loss values for validation closely resemble its training counterpart which clearly indicates that the model is performing significantly well on the new dataset. The validation loss values consistently achieved lower loss values than training loss suggesting that the model performed considerably well on new unseen dataset and further validating model's learning ability.

APPENDIX G
SEARCH ENGINES

Scientific search engines such as Google scholar,Papers with code and Research Rabbit were used to find papers using keywords such as 'Multimodal learning', 'Multimodal learning for egocentric videos', 'Multimodal fusion techniques', 'Unimodal vs Multimodal learning', 'Multibench framework', Classification on Image and audio modalities', 'Audio processing in deep learning' and variations of the aforementioned terms. Furthermore, several relevant literature review/overviews were found as references, which were used to find useful articles on specific topics. For each article the publication year was also carefully considered, to determine the information is contemporary and still relevant.

APPENDIX H
VISUALIZATION OF RESULTS

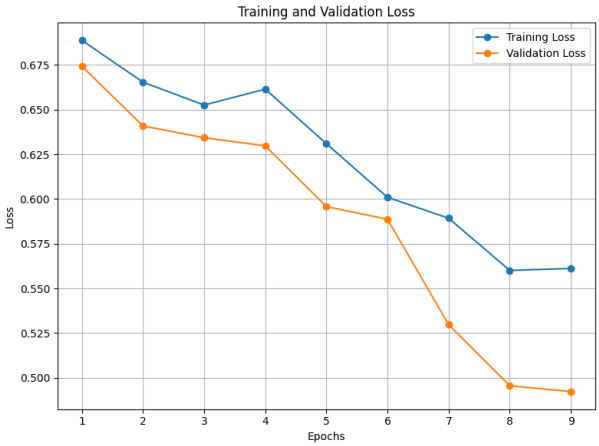
When model's output from batch segments of validation dataset are passed through the post processor (methodology described in *Appendix* E), csv files for predictions and ground truth are obtained. Figure 17 depicts the visualization of video frames for the frame segments obtained by correlating predictions with the ground truth for the best performing experiment that is Late fusion. Classification instances of False positive, False negative, True positive and True negative are depicted in the figure.



(a)

(b)

(c)



(d)

Fig. 16: Training and validation losses incurred during (a) *Early fusion experiment*, (b) *Late fusion experiment*, (c) *Unimodal audio experiment* (d) *Unimodal video experiment*.

Fig. 17: Visualization of results on validation subset. The figure depicts the visualization of model's output in terms of softmax probability scores correlated with ground truth labels for the Late fusion experiment. The results are corresponding to frame sequences(segments), obtained when model's output is passed through the post processor.

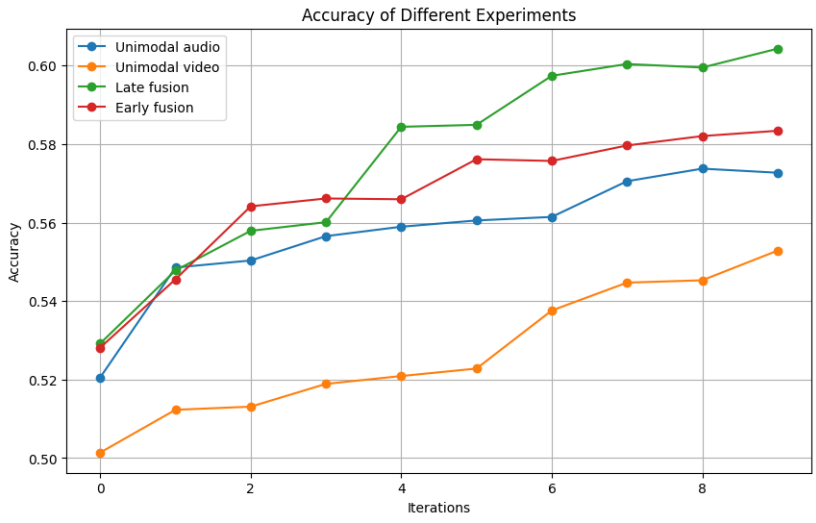


Fig. 18: Accuracy plot for different experiments depicting accuracy values with epochs.