Multimodal Sentiment Analysis : Case studies and Paradigms

Team ID 6:

Mohammad Ali Rehan	180050061
Neel Aryan Gupta	180050067
Shreya Pathak	180050100

Introduction

Multimodal Sentiment Analysis includes taking input as feature vectors of audio, video and text of a speaking person and figuring out the emotion he/she is expressing.
We present models that give good results by systematic integration of information across time and modalities. The models differ in the type of data they need, based on their alignment.

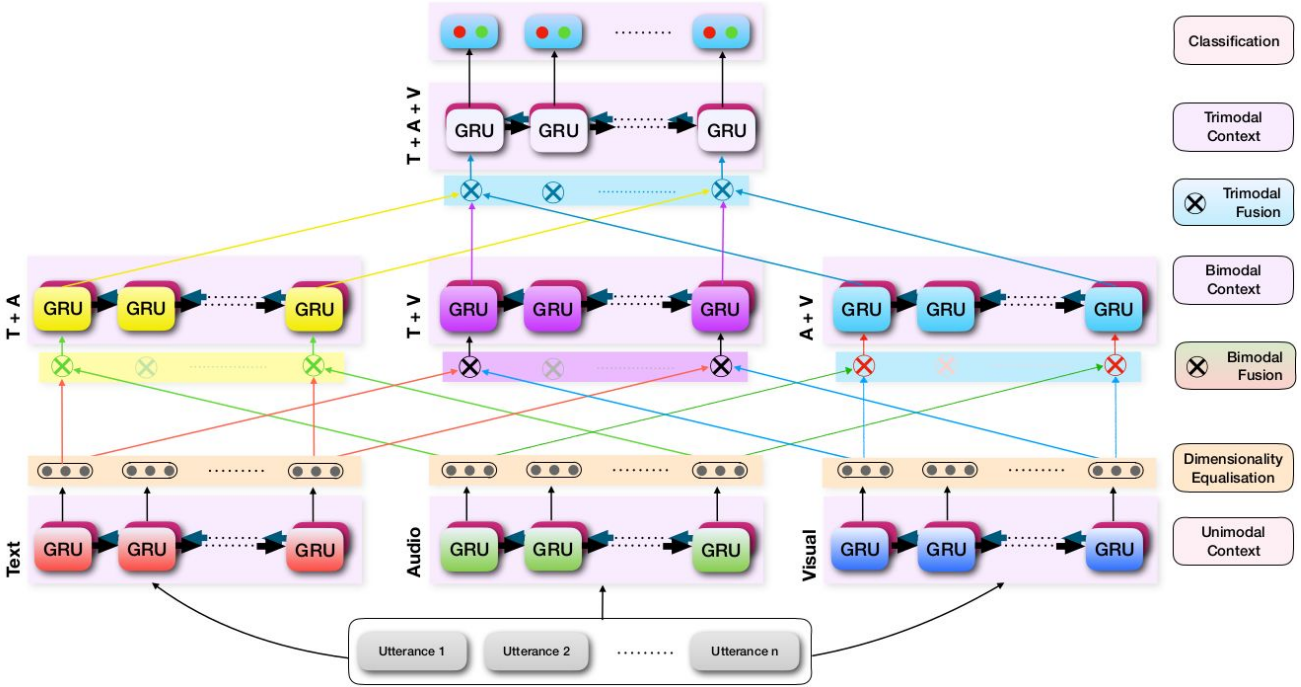
Alignment

Information in benchmark datasets are collected by sampling the various modalities and sentiments at different frequencies. We choose a pivot modality and use its timestamps to split the other modalities into bins. Since one bin may have more than one value of the other modality, we apply a coalescing function (averaging) to them. This way we can get data that is word aligned or label aligned and both require different model architectures.

Hierarchical Fusion for word aligned data

The model first fuses the modalities in pairs by passing them through a GRU and taking their weighted sum followed by *tanh*. This process is again repeated at the next fusion layer.

Finally we apply a GRU on the trimodal fused vectors and get our final representation of the utterances.



Utterance Aligned Dataset Details

IEMOCAP Dataset

	#Annotations	Seq. Length	Text	Audio	Video	Labels
Train	120	110	100	100	100	4
Test	30	110	100	100	100	4

MOSEI Dataset

	#Annotations	Seq. Length	Text	Audio	Video	Labels
Train	2150	98	300	74	35	2
Test	100	98	300	74	35	2

Results

IEMOCAP Dataset

Modality→ Metrics↓	Unimodal	Bimodal	Trimodal
Accuracy	0.73	0.78	0.80
Macro F1	0.72	0.77	0.80

MOSEI Dataset

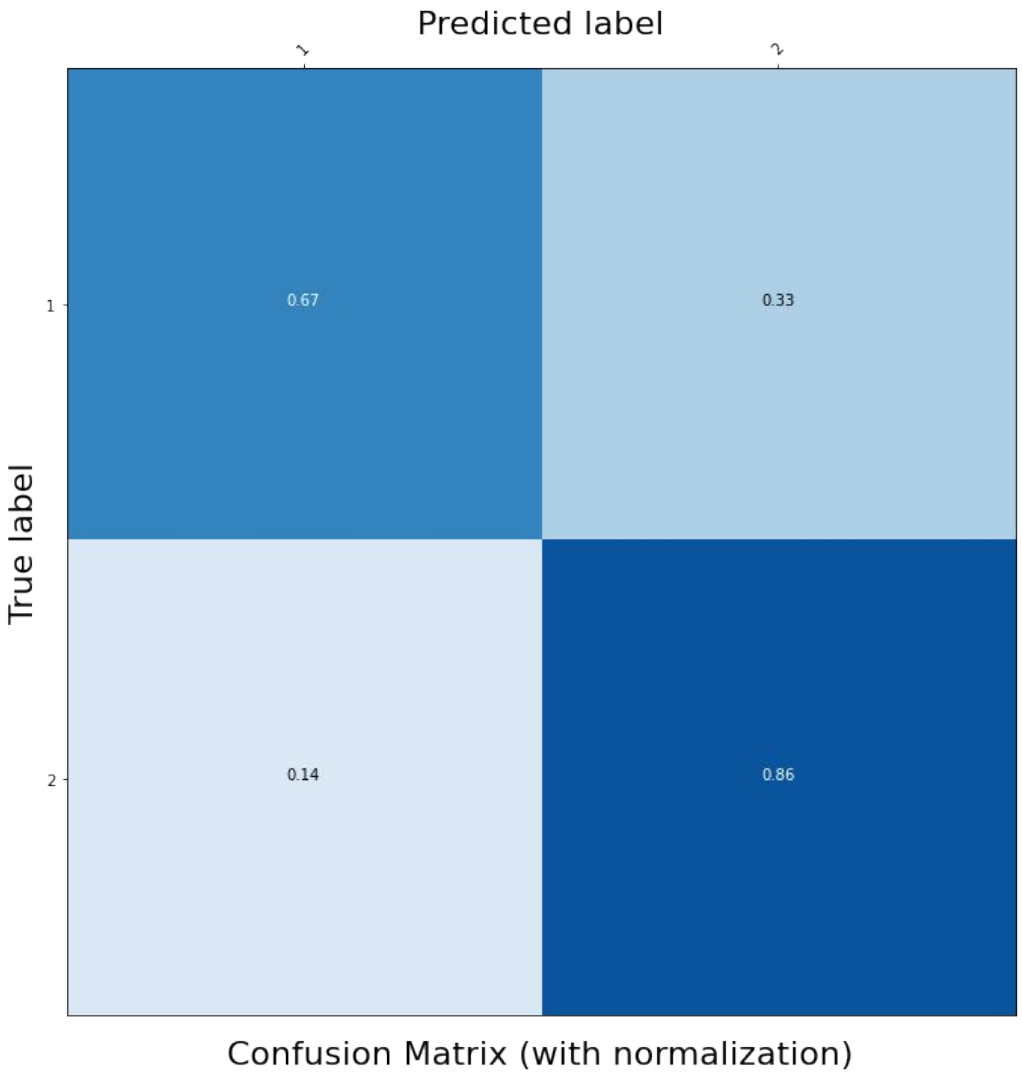
Modality→ Metrics↓	Unimodal	Bimodal	Trimodal
Accuracy	0.71	0.75	0.80
Macro F1	0.63	0.72	0.77

In the trimodal case we searched over the memory size of last GRU. The results are shown below. As we can see the model goes from underfitting to overfitting achieving an optimal value as shown

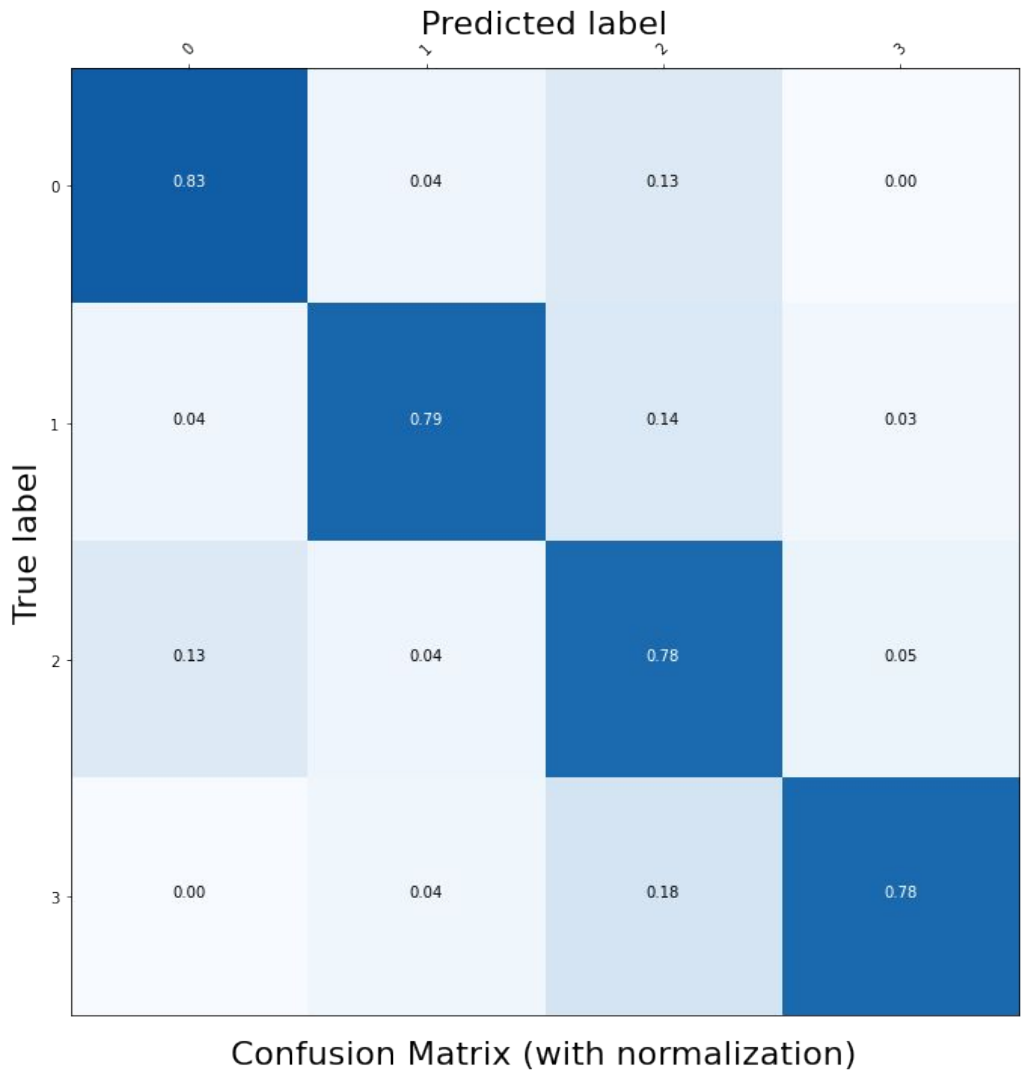
Parameters	400	450	500	600
Accuracy	78.5	**79.8**	78.89	79.1

Parameters	400	450	500	600
Accuracy	78.7	78.9	**80.1**	79.7

A similar parameter search was performed for models for each of the 3 modalities



Confusion Matrix for Trimodal Model trained and tested on MOSEI dataset



Confusion Matrix for Trimodal Model trained and tested on IEMOCAP dataset

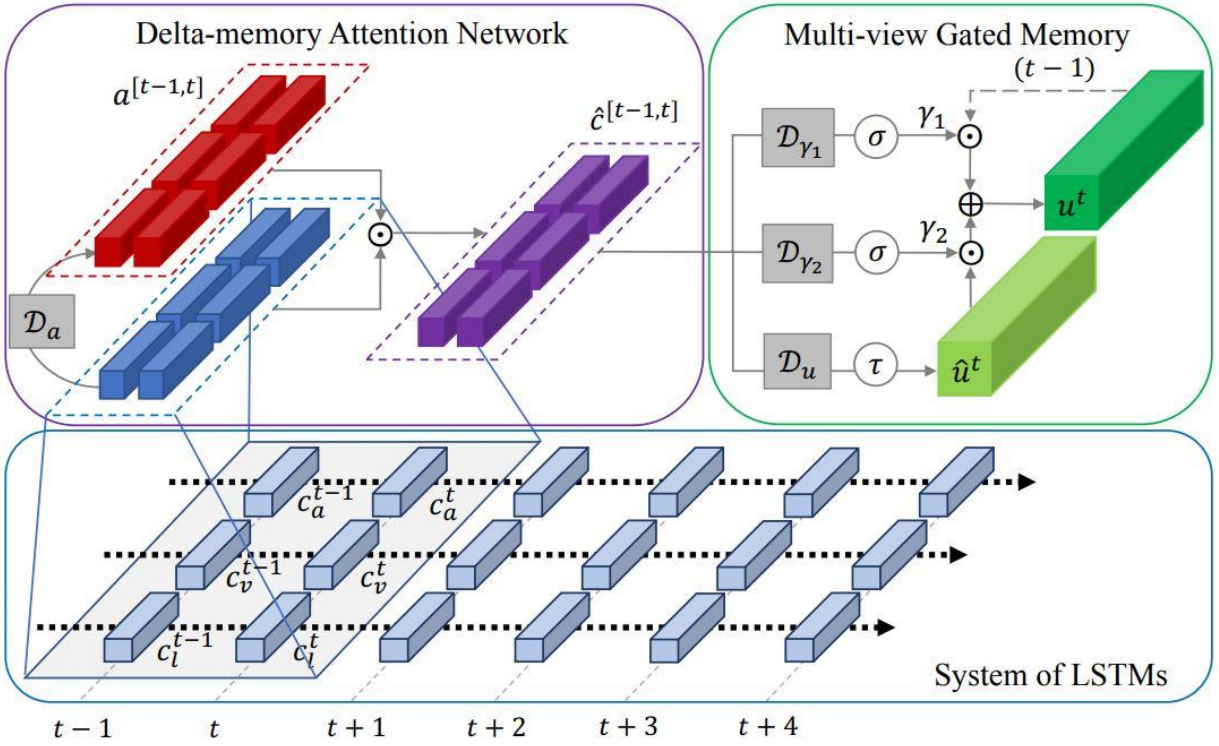
Memory Fusion Network for label aligned data

The system of LSTMs helps in generating memories by exploiting temporal relationship between inputs.

DMAN learns how modalities affect each other. It studies change in elements of memory vector in successive timesteps and suppresses those that remain constant.

MGM is like an LSTM but here the gates are neural networks. It learns the time dependent relation between the cross modal memories (ĉ).

We concatenate the final output of LSTM and MGM and use this to predict the sentiment.



Label Aligned Dataset Details

Dataset	#Annotations	Seq. Length	Text	Audio	Video	#Labels
IEMOCAP	7878 / 1970	21	300	74	35	4
MOSEI	15913 / 3979	20	300	74	35	3
MOSI	1682 / 421	20 / 50	300	74	35	2
MMMO	248 / 63	21	300	74	35	2
MOUD	308 / 78	21	300	74	35	3
YouTube	215 / 54	21	300	74	35	3

Annotations column shows the ‘train / test’ number of data examples.
For MOSI dataset, 20 and 50 sequence length datasets both were used.

IEMOCAP Results

The values inside the cell in the table below are the accuracy scores and the macro-averaged F1 scores delimited by |. We observe slight overfitting at 512 when modality(hence data) is lesser

Mem size→ Metrics↓	128	256	512
T	0.53 | 0.37	0.54 | 0.33	0.53 | 0.43
T+A	0.54 | 0.42	0.56 | 0.47	0.56 | 0.50
T+A+V	0.55 | 0.48	0.57 | 0.52	0.57 | 0.53

Modality→ Metrics↓	T	A	V	T+A	T+V	A+V	T+A+V
Accuracy	0.53	0.53	0.50	0.56	0.54	0.53	0.57
Macro F1	0.43	0.32	0.23	0.50	0.45	0.36	0.53

Text intrinsically carries much more info then Audio as is clear from unimodal F1 scores

MOSI - 20 Seq. length

Modality→ Metrics↓	T	A	V	T+A	T+V	A+V	T+A+V
Accuracy	0.71	0.47	0.51	0.73	0.72	0.51	0.74
Macro F1	0.71	0.39	0.50	0.73	0.72	0.47	0.74

MOSI - 50 Seq. Length

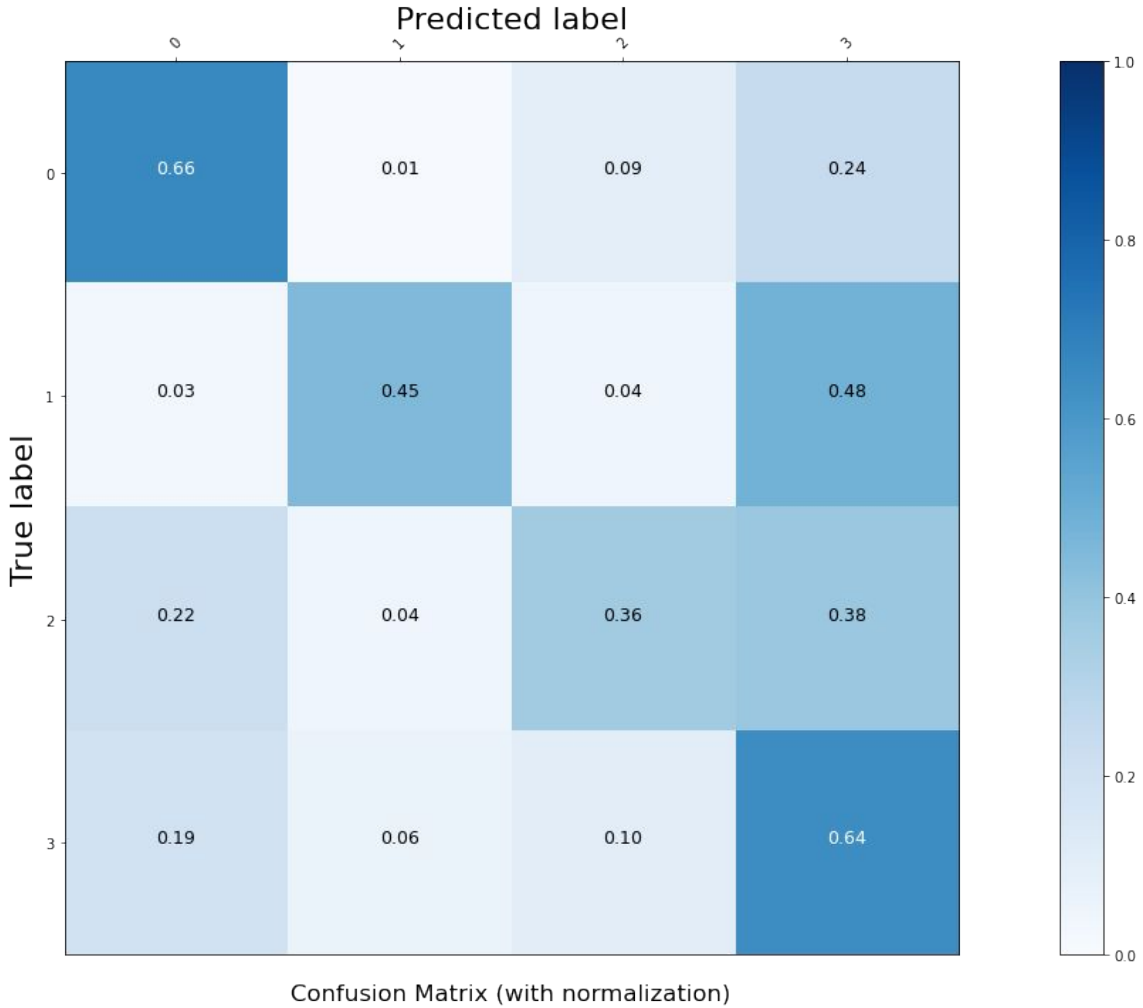
Modality→ Metrics↓	T	A	V	T+A	T+V	A+V	T+A+V
Accuracy	0.70	0.49	0.48	0.73	0.73	0.51	0.77
Macro F1	0.70	0.35	0.43	0.73	0.73	0.49	0.76

MOSEI

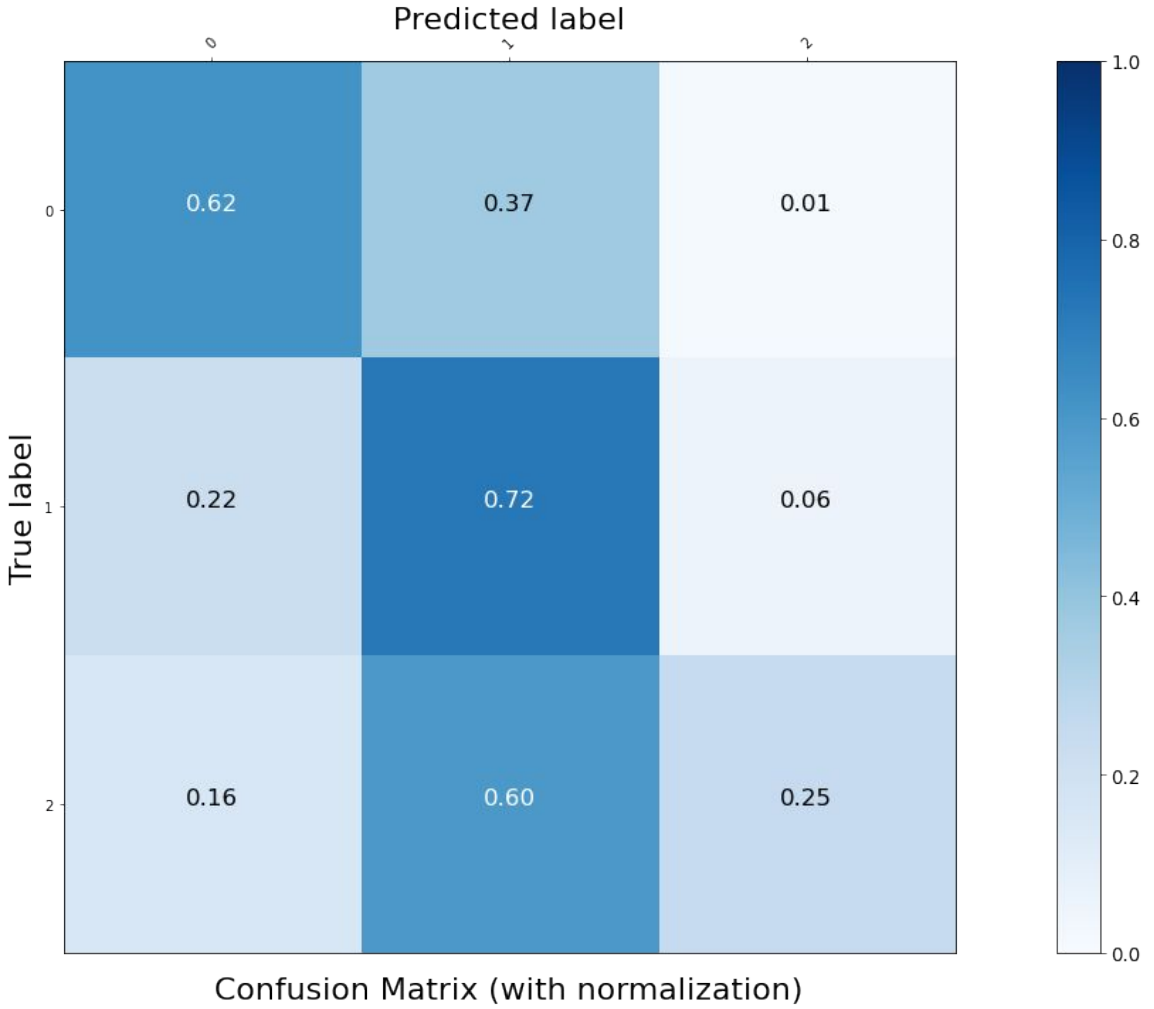
Modality→ Metrics↓	T	A	V	T+A	T+V	A+V	T+A+V
Accuracy	0.60	0.58	0.57	0.60	0.62	0.58	0.62
Macro F1	0.51	0.41	0.44	0.53	0.53	0.45	0.54

5-Fold Cross Validated datasets

Modality→ Dataset↓	T	A	V	T+A	T+V	A+V	T+A+V
MMMO	0.68 | 0.64	0.73 | 0.69	0.65 | 0.62	0.78 | 0.74	0.76 | 0.73	0.74 | 0.69	0.80 | 0.75
MOUD	0.59 | 0.49	0.53 | 0.42	0.56 | 0.53	0.60 | 0.55	0.63 | 0.61	0.59 | 0.57	0.64 | 0.63
YouTube	0.46 | 0.44	0.38 | 0.24	0.40 | 0.24	0.47 | 0.44	0.47 | 0.42	0.42 | 0.28	0.48 | 0.44



Confusion Matrix for MFN model for IEMOCAP dataset



Confusion Matrix for MFN model for MOSEI dataset

Hypothesis Testing (χ^2 Test)

$$\sum_{ij} \frac{(E_{ij} - O_{ij})^2}{E_{ij}} \sim \chi_c^2$$

We can use the above table to compare cases like IEMOCAP Bimodal and Trimodal wherein the change in accuracy is not large but the test statistic report an increase in performance.
It can be applied to MOSEI Unimodal and Bimodal systems also.

Dataset	Model	Unimodal	Bimodal	Trimodal
IEMOCAP	HFusion	1554.53 (73)	1794.94 (78)	1982.28 (80)
IEMOCAP	MFN	396.79 (53)	846.41 (56)	930.64 (57)
MOSEI	MFN	704.62 (60)	797.61 (60)	949.385 (62)

Drawbacks and Future

- Label aligned datasets have much more availability than utterance-aligned datasets.
- Preprocessing datasets take a lot of time, hence only pre-processed datasets (mostly made available by CMU) have been used.
- The accuracy scores of both HFN and MFN cannot be directly compared due to the difference in their inputs, which are aligned on utterances and labels respectively.
- The 300D word embedding contain much more information than can be gained by just analysing the dataset. This sometime causes the models to overpower the other modalities. It can be rectified in future works. We can also use other audio/video feature extractors for this
- Small datasets like Youtube and MMMO can give rise to results with high variance as the models are quite large. Hence it takes a lot of tuning to ensure that overfitting is not taking place. It can be improved by kind of a lite or sparser version of our models.
- The models require highly specialised features based on eye/mouth etc tracking of subjects. In the future, one can work on robust models that do well with conventional video footage.

Methodology & Teamwork

We enlisted small subtasks that a group of two people could work upon with one of them being a primary lead in it. This lead to good teamwork while allowing everyone to have a task for which they were responsible. The work was divided as follows:

- Dataset Research and Formatting : Ali & Neel
- HFusion Paper Summarisation : Ali
- HFusion Implementation : Ali
- HFusion Inference and Error Analysis : Shreya
- Hierarchical Bimodal Model Implementation : Shreya
- Hierarchical Unimodal Model Implementation : Shreya
- MFN Paper Summarisation : Ali
- MFN Model Implementation : Neel
- MFN Inference and Error Analysis : Neel
- MFN k-Fold Cross Validation : Neel
- Significance Testing : Ali & Shreya

Thank you