Multimodal Humour Recognition

(Final Presentation)

Presented by: Shivangana Rawat(cs20mtech12001)

Pranoy Panda(cs20mtech12002)

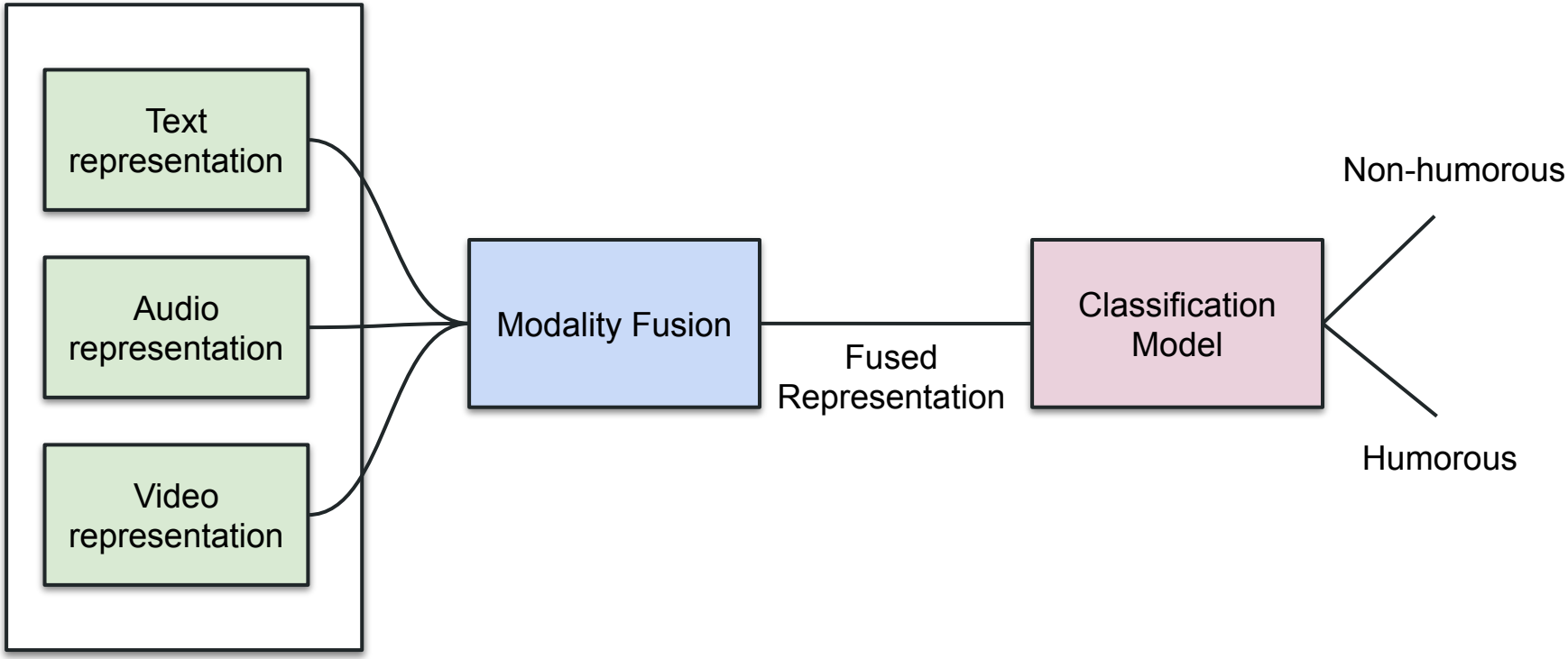
Outline (30th Nov 2021)

1. Problem Statement
2. Pipeline
3. Previous results
4. Attempt at remedying class imbalance
5. Flowchart of the project
6. Context Modelling
7. Modality Fusion
8. Speaker Modelling
9. Results Summary
10. Challenges faced and Conclusions

Problem Statement

- Let the utterance at time t be represented as u_t (Utterance is a part of the dialogue where one party speaks without taking any breathes/pauses)
- In a multimodal scenario, we have three signals for describing an utterance(audio, video, text/language). That is $u_t = [u^a_t , u^v_t , u^l_t]$
- ***Task: Given the utterances $[u_1, u_2, ..., u_t]$, predict whether the utterance at time t was humorous or not.***

Pipeline*



* We are also exploiting temporal structure in the conversation between the speakers.

Previous Results(F1-Precision-Recall values)

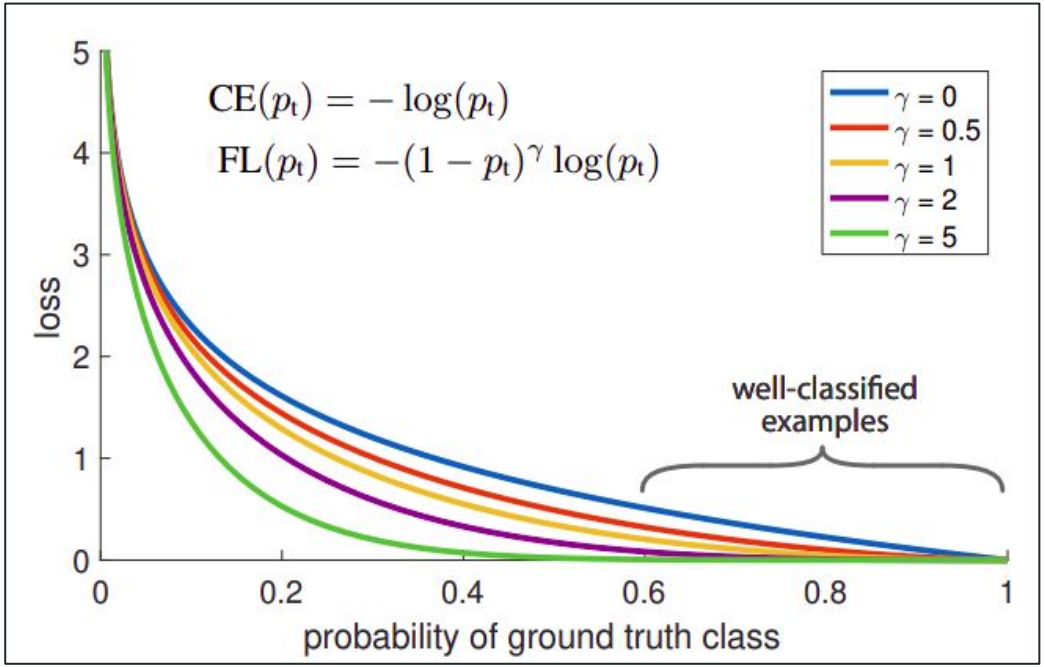
Method	F1 Score	Precision	Recall
IndicBERT	0.59	0.60	0.65
sentence-BERT(SBERT)	0.59	0.62	0.66
FastText	0.53	0.55	0.66
ResNext101 3D CNN	0.55	0.57	0.66
SBERT+ResNext101 3D CNN	**0.60**	0.60	0.64

Note: The above provided scores are weighted average scores(average over classes and weighted by support of different classes)

Class Imbalance

Class Imbalance Problem

- Down-weight the contribution of easy examples during training and rapidly focus the model on hard examples via Focal Loss[1]
- Higher gamma implies lesser weightage to easy examples



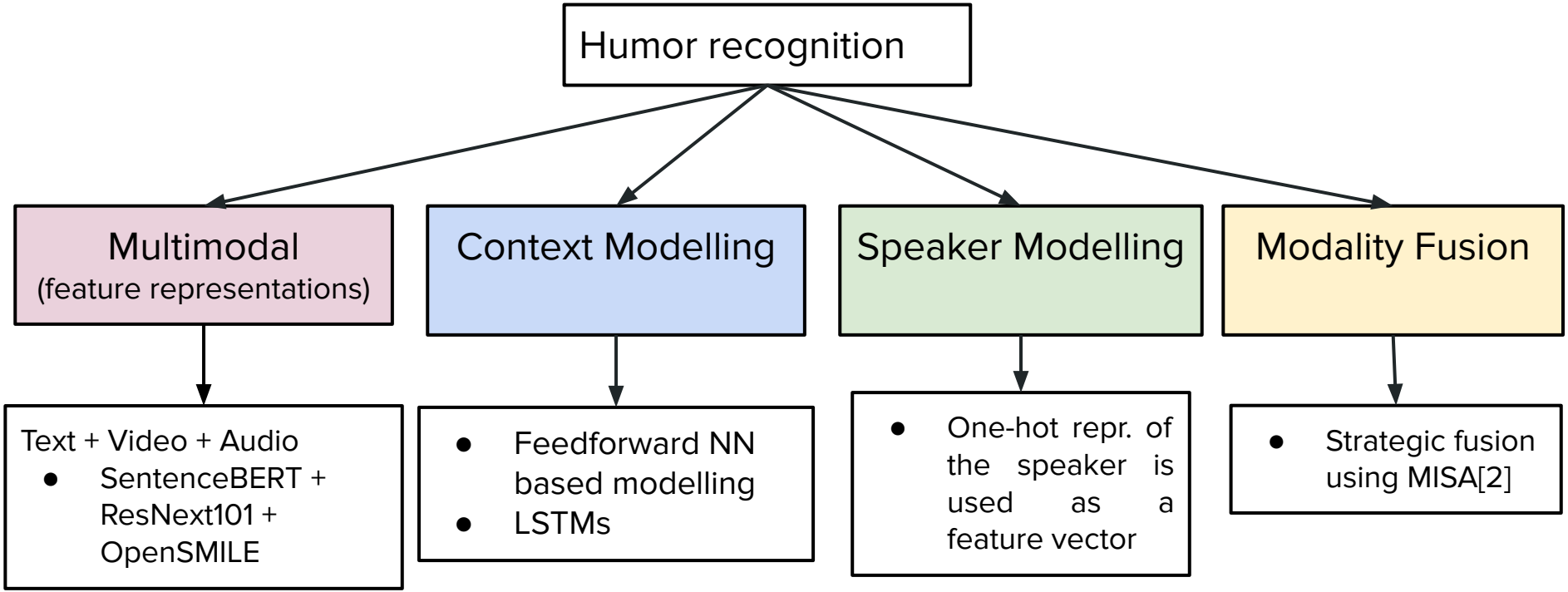
[1] Focal Loss for Dense Object Detection, CVPR 2017

Focal Loss

Method	F1 Score	Precision	Recall
SBERT	0.59	0.62	0.66
SBERT + 3D CNN	0.60	0.60	0.64
SBERT + 3D CNN + Focal Loss	0.58	0.60	0.60

Focal Loss gives similar performance to cross entropy loss, so we stick to the standard cross entropy loss for all future experiments.

Flowchart for the project



[2] MISA: Modality-Invariant and-Specific Representations for Multimodal Sentiment Analysis, ACM MM 2020

Multimodal

OpenSmile for Acoustic features

- OpenSmile[3] Extracts low level features from the audio signal
- Features include frequency, bandwidth, loudness, amplitude of signal, etc.
- We use a 65 dimensional representation for audio associated with each utterance.

```
['Loudness_sma3',
 'alphaRatio_sma3',
 'hammarbergIndex_sma3',
 'slope0-500_sma3',
 'slope500-1500_sma3',
 'spectralFlux_sma3',
 'mfcc1_sma3',
 'mfcc2_sma3',
 'mfcc3_sma3',
 'mfcc4_sma3',
 'F0semitoneFrom27.5Hz_sma3nz',
 'jitterLocal_sma3nz',
 'shimmerLocaldB_sma3nz',
 'HNRdBACF_sma3nz',
 'logRelF0-H1-H2_sma3nz',
 'logRelF0-H1-A3_sma3nz',
 'F1frequency_sma3nz',
 'F1bandwidth_sma3nz',
 'F1amplitudeLogRelF0_sma3nz',
 'F2frequency_sma3nz',
 'F2bandwidth_sma3nz',
 'F2amplitudeLogRelF0_sma3nz',
 'F3frequency_sma3nz',
 'F3bandwidth_sma3nz',
 'F3amplitudeLogRelF0_sma3nz']
```

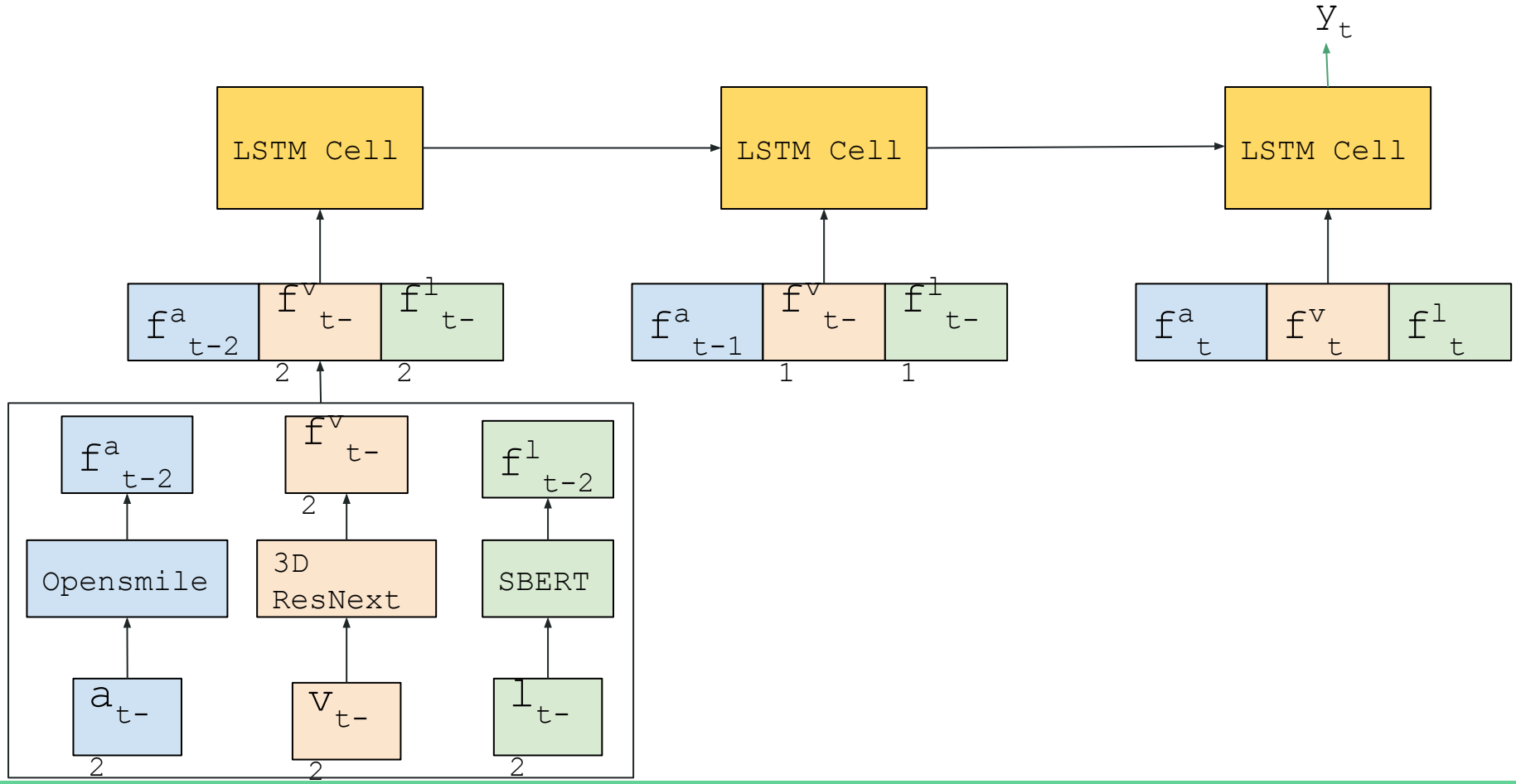
[3] https://github.com/audeering/opensmile

From BiModal to MultiModal

Method	F1 Score	Precision	Recall
openSmile	0.53	0.44	0.66
SBERT + openSmile	0.55	0.61	0.66
3D CNN + openSmile	0.53	0.55	0.66
SBERT + 3D CNN	0.60	0.60	0.64
SBERT + 3D CNN + openSmile	**0.58**	**0.62**	**0.67**

Context Modelling

Context modelling block diagram

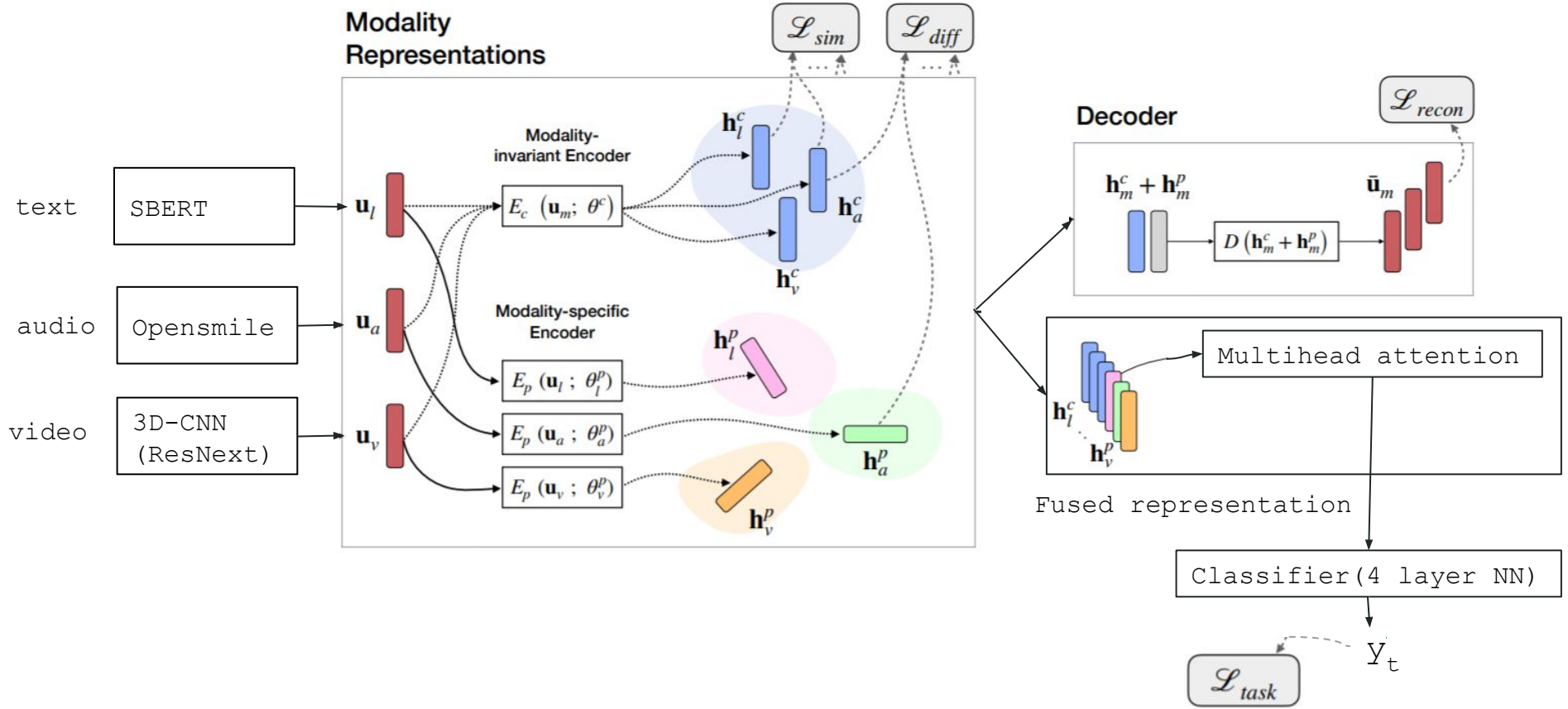


Context Modelling (Text only)

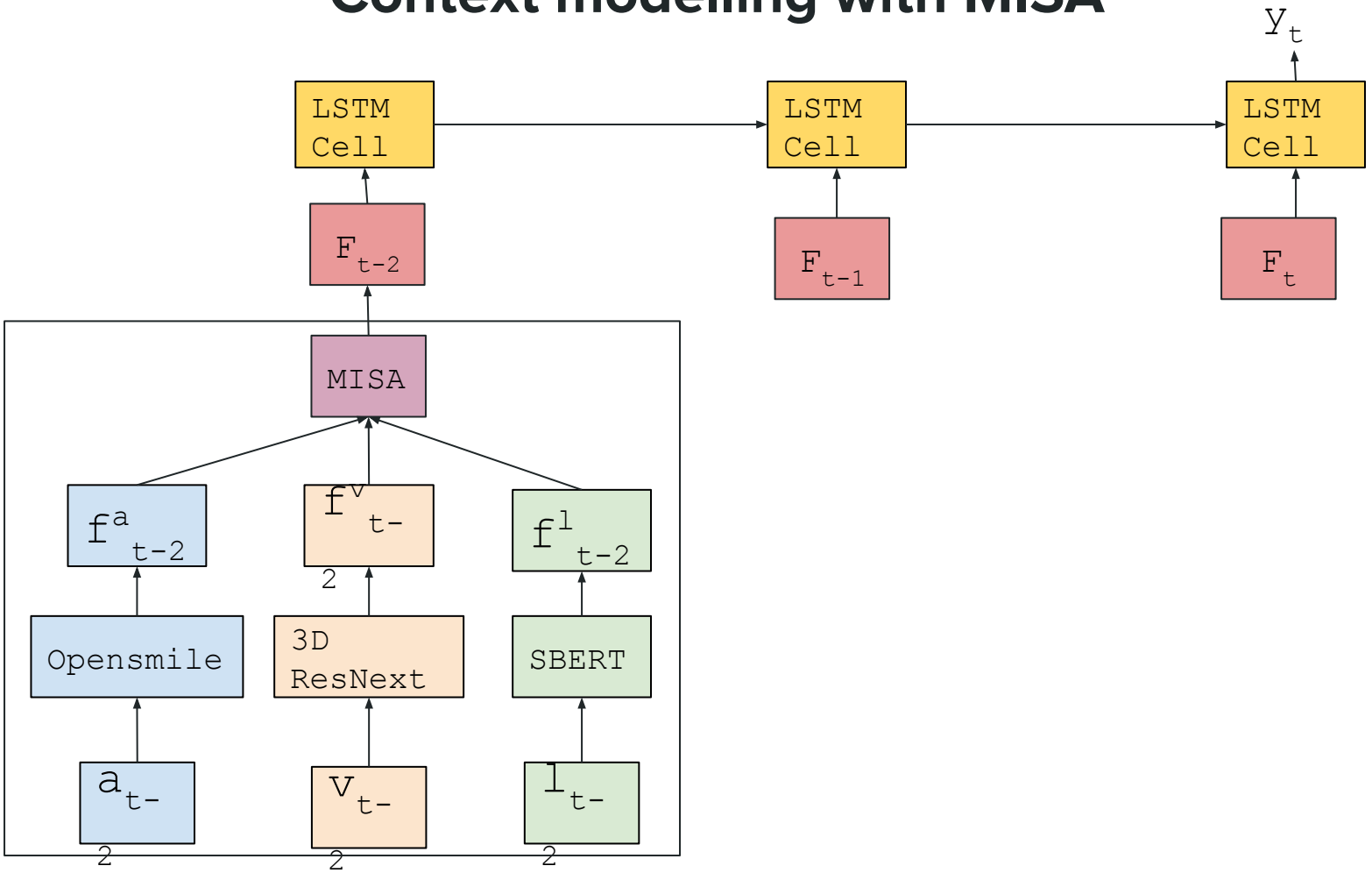
Method	F1 Score	Precision	Recall
SBERT	0.59	0.62	0.65
SBERT + RNN	0.59	0.60	0.59
SBERT + LSTM	**0.60**	0.62	0.65
SBERT + Bidirectional LSTM	0.60	0.60	0.63

Modality Fusion

Modality Fusion: MISA



Context modelling with MISA

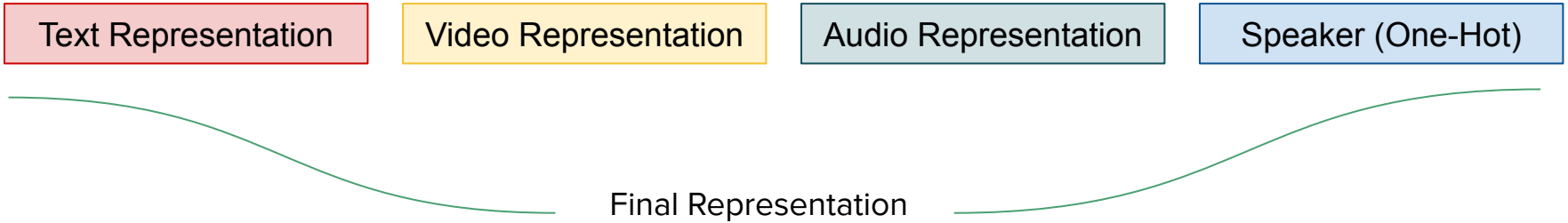


Modality Fusion

Method	F1 Score	Precision	Recall
MISA	0.62	0.62	0.65
MISA with NN based context modelling	0.62	0.65	0.62
MISA with LSTM based context modelling	**0.63**	**0.62**	**0.66**

Speaker Modelling

Speaker Modelling



Method	**F1 Score**	**Precision**	**Recall**
SBERT + 3D CNN + openSmile	0.58	0.62	0.67
SBERT+3D CNN+openSmile+Speaker	**0.60**	0.60	0.65

Comparisons

- Although it would be unfair to compare with the numbers in the M2H2 paper, we still provide them for the sake of completeness.
- For comparison on our test set, we use the naive concatenation of multimodal features with a linear classifier as our baseline.

Labels	*MISA+DialogueRNN*			*MISA+bcLSTM*		
	P	**R**	**F1**	**P**	**R**	**F1**
T	67.11	68.21	67.65	66.27	66.51	66.38
A	57.91	59.10	58.52	57.51	58.92	58.21
V	55.16	57.32	56.74	53.13	55.21	54.19
T+V	70.03	70.61	70.31	67.74	68.89	68.31
T+A	69.70	69.90	69.63	67.41	68.12	67.75
A+V	61.70	63.21	62.44	59.61	61.23	60.40
T+V+A	**71.21**	**72.11**	**71.67**	**69.04**	**69.83**	**69.43**

Result Summary

Method	F1 Score	Precision	Recall
SBERT + 3D CNN + openSmile	0.58	0.62	**0.67**
SBERT+3D CNN+openSmile+Speaker	0.60	0.60	0.65
MISA + LSTM based context modelling + Speaker	**0.63**	**0.63**	0.66

Challenges Faced (Data and Compute)

- No available test set for M2H2 dataset
  - Train-test split not provided
  - We use our own split for training and testing.
  - The training set contains 104 scenes while the test set contains 32 scenes.

- Large models
  - BERT and 3D CNN models used are extremely large to fit onto a single GPU.
  - Due to this it is computationally infeasible to fine tune all the models together.
  - So we instead use pretrained models for all our tasks. (Although we do show one of the result where we fine tune a BERT model on the text modality only.)

Conclusions

- Text is an important modality as compared to video and audio for the M2H2 dataset.
- Multimodal performs better than bimodal .
- Modality fusion techniques gives a performance boost as compared to conatinating individual modalities.
- Simple context modelling using LSTM gives marginal improvements. Sophisticated techniques like DialogueRNN might give better results.