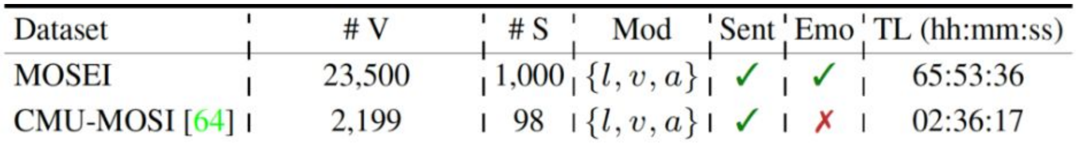
STM: SenTiMetalchemy
Multi-modal approach to Sentiment Analysis

Team Member: Tom Liu, Kexin Gao, Lexie Wang, Yutong Li

Presentation Outline

- Tasks Description
- System overview
- Model design
  - Multi-modal attention-based model
- Results
  - Primary task results
  - Adaptation task results
- Issues and Successes
- Related Readings

Dataset and Tasks



Dataset	# V	# S	Mod	Sent	Emo	TL (hh:mm:ss)
MOSEI	23,500	1,000	$\{l, v, a\}$	✓	✓	65:53:36
CMU-MOSI [64]	2,199	98	$\{l, v, a\}$	✓	✗	02:36:17

D2 - D3

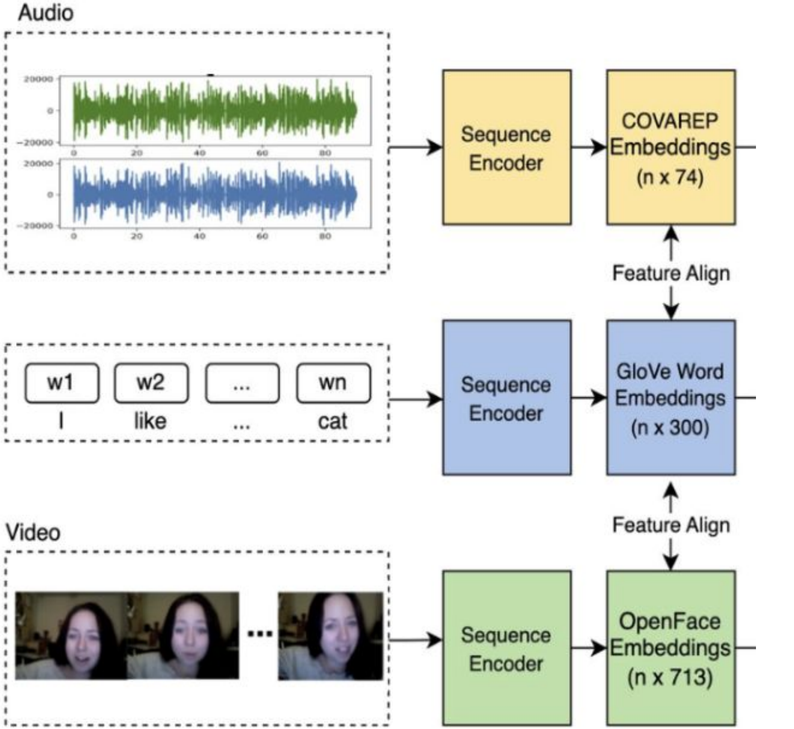
- Dataset: CMU-MOSI
  - **2199** opinion video clips annotated with **sentiment in the range [-3,3]**
- Model
  - Baseline: SVR (text-only)
  - Fully-connected neural network (text-only)
  - LSTM neural network (text-only)
  - LSTM-based **Multimodal** neural network
- Main Task - Sentiment analysis
  - Input: **Text** (Unimodal), Audio, Video (Multimodal)
  - Output: Sentiment intensity score within [-3, 3]

D4 - Larger dataset, enhanced model, adapted task

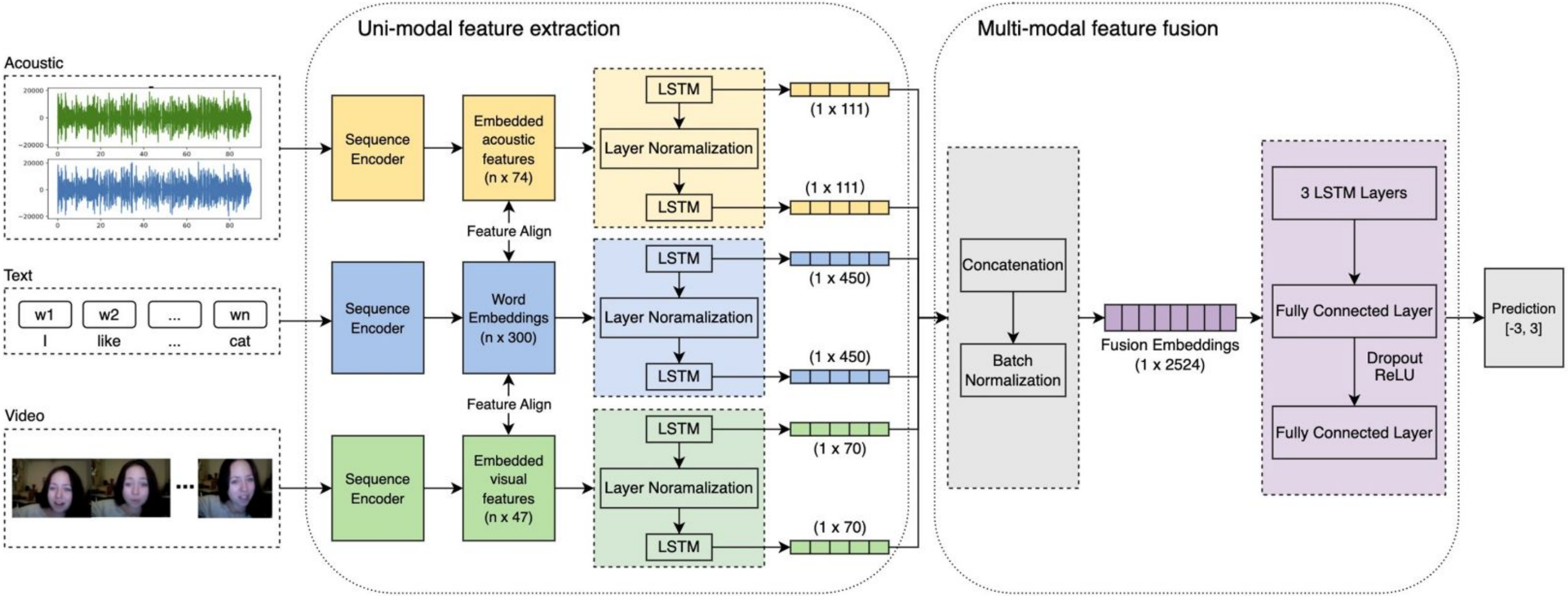
- Dataset: CMU-MOSEI
  - **23,500** sentence video clips, from 1000 online YouTube speakers
  - Annotated with **sentiment intensity in the range [-3,3]**and **6 emotion labels**
- Model
  - LSTM and attention-based multi-modal neural network
- Main Task - Sentiment analysis
  - Input: Text, Audio, Video
  - Output: sentiment intensity score within [-3, 3]
- Adaptation Task - Emotion detection
  - Input: Text, Audio, Video
  - Output: emotion label (happy, sad, angry, disgust, surprise, fear)

System Overview

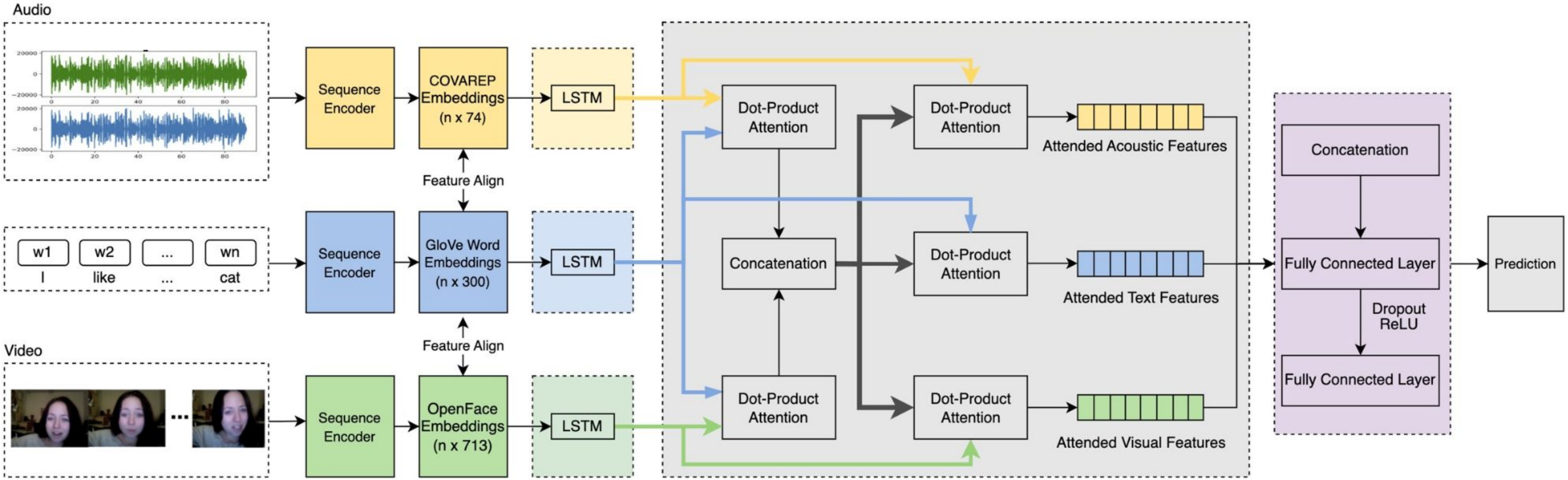
- Data Loading
  - CMU-Multimodal SDK - get the embeddings of the three modalities of each instance
- Data Processing
  - Feature Extraction:
    - Text: GloVe Embeddings
    - Acoustic: COVAREP Embeddings
    - Visual: OpenFace Embeddings
  - Feature Alignment:
    - Align acoustic and visual features with textual features
- Dataset Split
  - dataset is split into train (58%), test (10%), and development (32%)
- Model training
  - One multimodal model (details in next part)
- Evaluation Metrics:
  - Sentiment Analysis Task: ACC-7, ACC-2, F1 score, MSE, R^2
  - Emotion Recognition Task: ACC



Multi-Modal Model (D3)



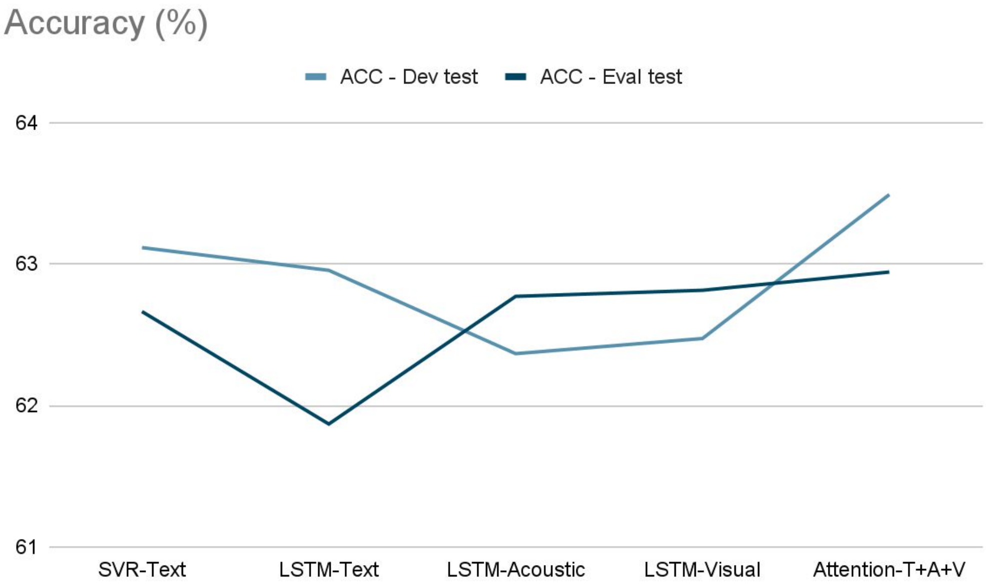
Multi-Modal Attention-based Model (D4)



Results of Main Task (Sentiment Analysis)

	Model	Modality	F1	MAE	ACC-2	ACC-7	R^2
Eval	SVR	Text	77.4	0.695	62.5	40.1	0.349
Eval	Attention	T+A+V	**78.9**	**0.679**	**62.8**	**44.5**	**0.354**
Dev	SVR	Text	75.7	0.672	59.3	42.5	0.289
Dev	Attention	T+A+V	**77.9**	**0.652**	**60.1**	**46.6**	**0.318**
	UniMSE (SOTA) (Sentiment)	*T+A+V*	*85.79*	*0.523*	*85.8*	*54.4*	*0.773*

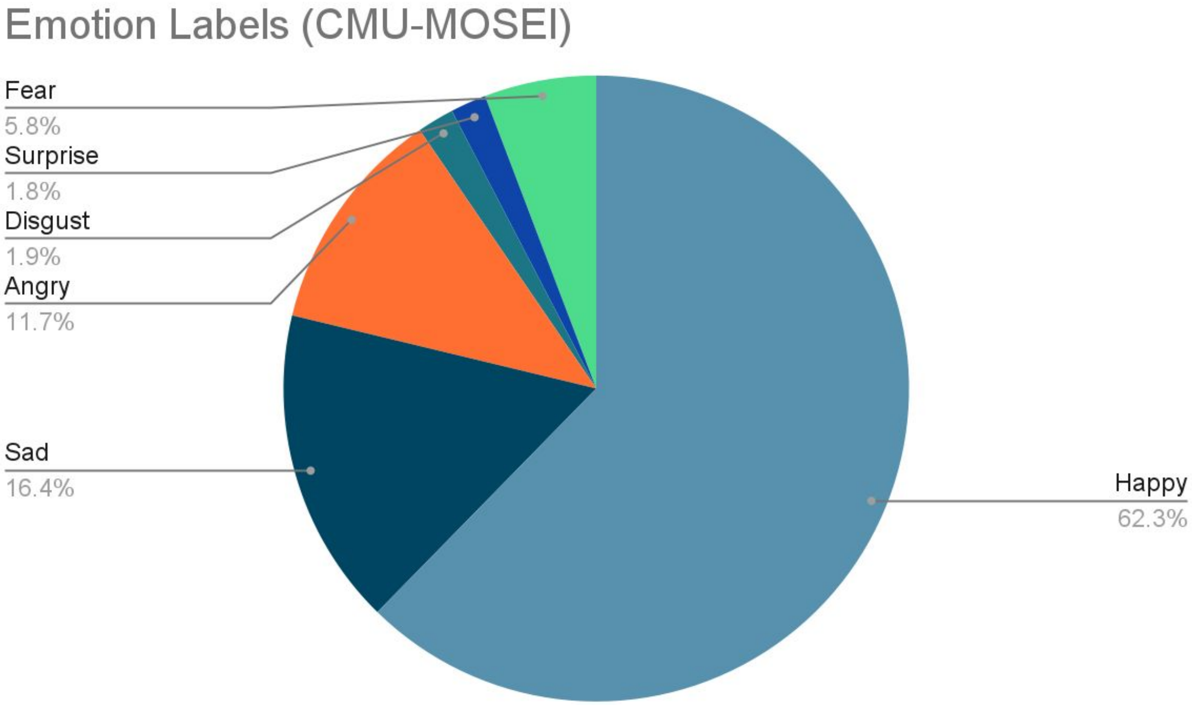
Results of Adaptation Task (Emotion Recognition)



Model	Modality	ACC - Dev test	ACC - Eval test
SVR	Text	63.1	62.6
LSTM	Text	62.9	61.8
LSTM	Acoustic	62.3	62.7
LSTM	Visual	62.4	62.8
Attention	T+A+V	**63.4**	**62.9**

- Multimodal model outperformed Uni-modal models
- Issue: Lack of complementary information

Issue: Imbalanced dataset



Happy	10172
Sad	2679
Angry	1910
Disgust	311
Surprise	301
Fear	945

Imbalanced dataset causes problems like …

- Biased model predictions
- Poor generalization

Issue: Prediction

Emotion Label (in test set)	Number of GOLD	Number of True Predictions	%
Happy	2925	2760	94.3
Sad	601	159	26.4
Angry	487	4	0.8
Disgust	145	0	0
Surprise	245	0	0
Fear	99	0	0

Sentiment Range (in test set)	Number of GOLD	Number of True Predictions	%
Positive [1, 3]	1172	140	11.9
Neutral [-1,1]	2673	2604	97.4
Negative [-3,-1]	810	117	14.4

Successes

- Learned TensorFlow to reimplement a multimodal model
- Experimented on 10% of the dataset, then trained on the entire dataset
- Designed optimized training approach for 30GB+ embeddings
- Attempted shrinking process to help imbalanced dataset problem (despite it did not help)
- Applied attention mechanism helped the bottleneck problem in D3
- Improvement in results (except ACC-2)

Reference

Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph (Bagher Zadeh et al., 2018)

MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos (Zadeh et. al, 2016)

Attention is All You Need (Vaswani et. al., 2017)

Long Short-term Memory (Hochreiter and Schmidhuber, 1997)

Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis (Han et. al., 2021)

Contextual Inter-modal Attention for Multi-modal Sentiment Analysis (Ghosal et al., 2018)

UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition (Hu et. al., 2022)