Computational Modeling of Human Multimodal Language: The MOSEI Dataset and Interpretable Dynamic Fusion

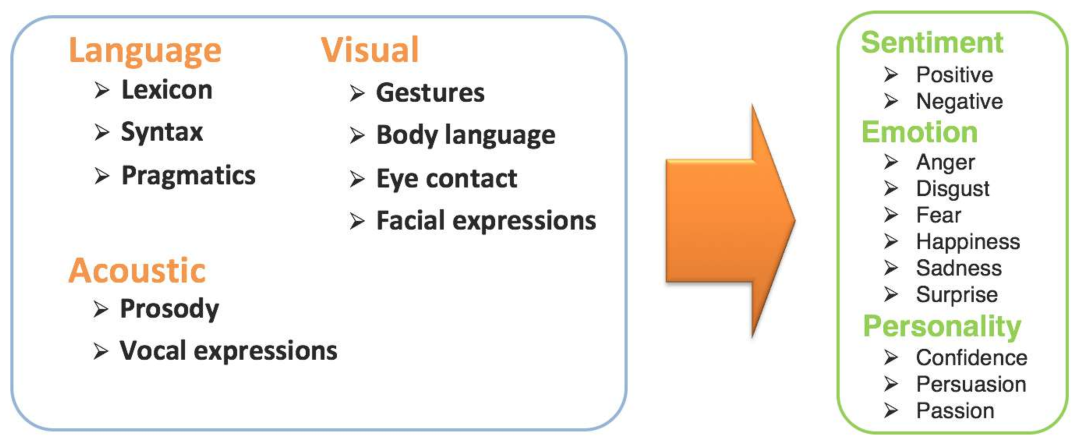
Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency

Machine Learning Department, School of Computer Science, Carnegie Mellon University

{pliang, rsalakhu, morency}@cs.cmu.edu

Introduction

Computational modeling of human multimodal language is an emerging research area in natural language processing spanning the **language**, **visual** and **acoustic** modalities.



Require resources with diversity in: **training samples**, **topics**, **speakers**, **annotations** and **modalities**. This will allow us to build models that **generalize** across speakers, gender, topics and modalities.

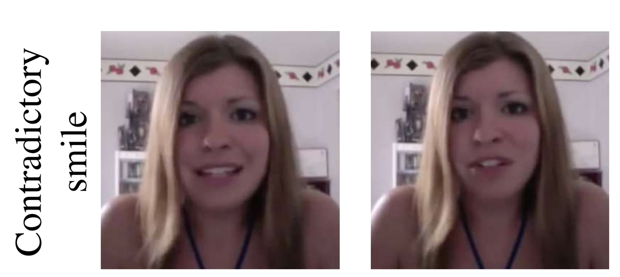
MOSEI Dataset

We leverage social multimedia to acquire large quantities of data.

- The MOSEI dataset contains **23,453 video clips** from **1,000 speakers** and spans **250 topics**.
- Features extracted include the **3 modalities of language, visual and acoustic**.
- MOSEI is **annotated for sentiment and emotions**.

Language: *All I can say is he's a pretty average guy.*

Vision:

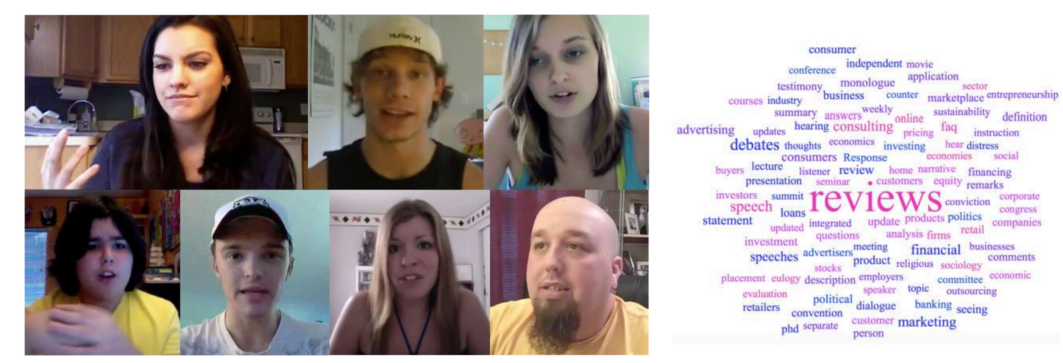


Acoustic: (disappointed voice)

Dataset	# S	# Sp	Mod	Sent	Emo	TL (hh:mm:ss)
MOSEI	**23,453**	**1,000**	$\{l, v, a\}$	✓	✓	**65:53:36**
CMU-MOSI	2,199	98	$\{l, v, a\}$	✓	✗	02:36:17
ICT-MMMO	340	200	$\{l, v, a\}$	✓	✗	13:58:29
YouTube	300	50	$\{l, v, a\}$	✓	✗	00:29:41
MOUD	400	101	$\{l, v, a\}$	✓	✗	00:59:00
SST	11,855	–	$\{l\}$	✓	✗	–
Cornell	2,000	–	$\{l\}$	✓	✗	–
Large Movie	25,000	–	$\{l\}$	✓	✗	–
STS	5,513	–	$\{l\}$	✓	✗	–
IEMOCAP	10,000	10	$\{l, v, a\}$	✗	✓	11:28:12
SAL	23	4	$\{v, a\}$	✗	✓	11:00:00
VAM	499	20	$\{v, a\}$	✗	✓	12:00:00
VAM-faces	1,867	20	$\{v\}$	✗	✓	–
HUMAINE	50	4	$\{v, a\}$	✗	✓	04:11:00
RECOLA	46	46	$\{v, a\}$	✗	✓	03:50:00
SEWA	538	408	$\{v, a\}$	✗	✓	04:39:00
SEMAINE	80	20	$\{v, a\}$	✗	✓	06:30:00
AFEW	1,645	330	$\{v, a\}$	✗	✓	02:28:03
AM-FED	242	242	$\{v\}$	✗	✓	03:20:25
Mimicry	48	48	$\{v, a\}$	✗	✓	11:00:00
AFEW-VA	600	240	$\{v, a\}$	✗	✓	00:40:00

Comparison between the MOSEI dataset and standard sentiment analysis and emotion recognition datasets e.g. [1-3]. #S: number of annotated data points. #Sp: number of distinct speakers. Mod: subset of modalities present from $\{(l)anguage, (v)ision, (a)coustic\}$. Sent and Emo columns indicate presence of sentiment and emotion labels. TL: total number of video hours.

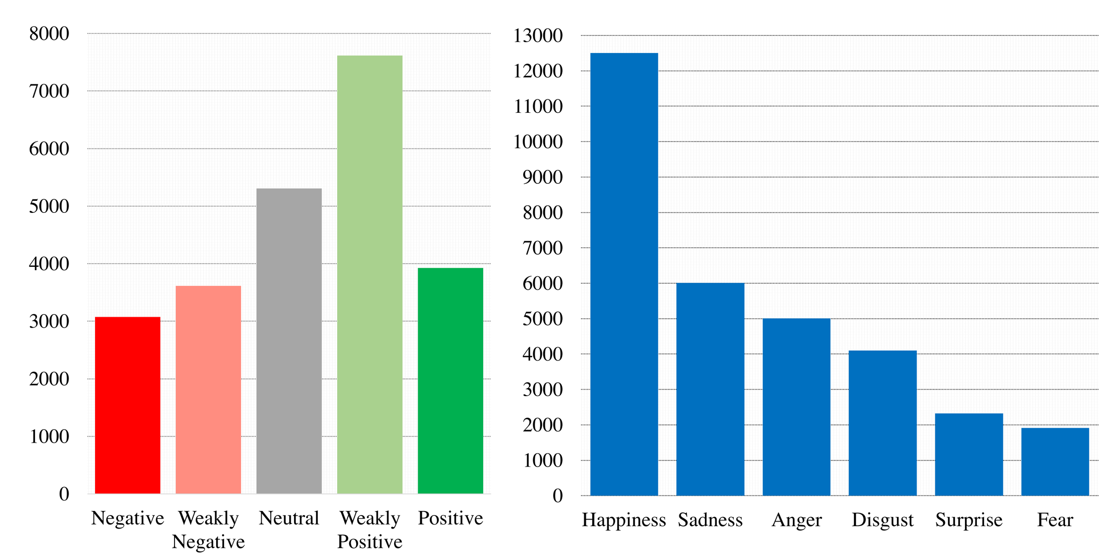
Topic Diversity: The 5 most frequent topics are: reviews (16.2%), debate (2.9%), consulting (1.8%), financial (1.8%) and speech (1.6%). The remaining topics are almost uniformly distributed at 0.5%-1.5% each.



Screenshots from MOSEI dataset (left) and distribution of topics (right).

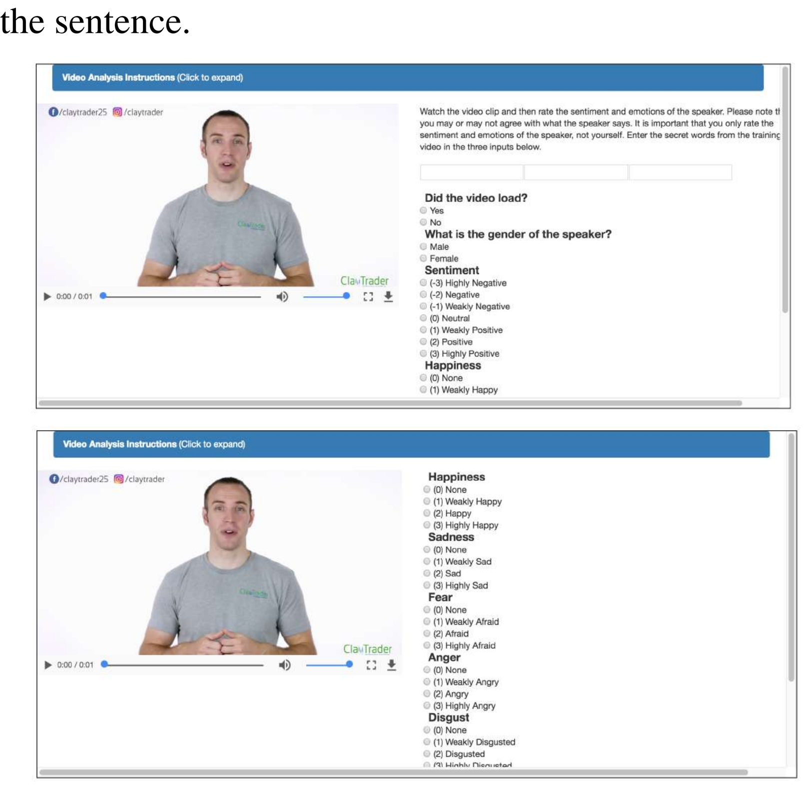
Annotations

The dataset is annotated on Amazon Mechanical Turk for **sentiment** on a [-3,3] scale and the presence of **emotions** happiness, sadness, anger, disgust, surprise and fear on a [0,3] scale.



To standardize annotations, we provide annotators with a 5 minute training video with the following definitions:

- **Sentiment**: the speaker's attitude towards the topic of his/her discussion.
- **Emotions**: the speaker's expressed state of mind and feeling while uttering the sentence.



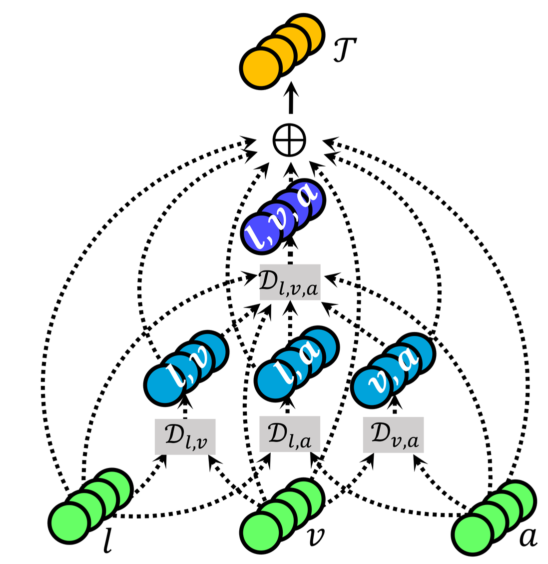
Screenshot of the **annotation user interface** for sentiment (top) and emotion (bottom) labeling.

Multimodal Features

- **Language**: GloVe word embeddings [4]
- **Visual**: FaceNet embeddings [5], FACET
- **Acoustic**: COVAREP [6]
- **Alignment**: P2FA between audio and transcripts.

CMU Multimodal Data SDK for fast data loading and alignment: https://github.com/A2Zadeh/CMU-MultimodalDataSDK.

Dynamic Fusion Graph



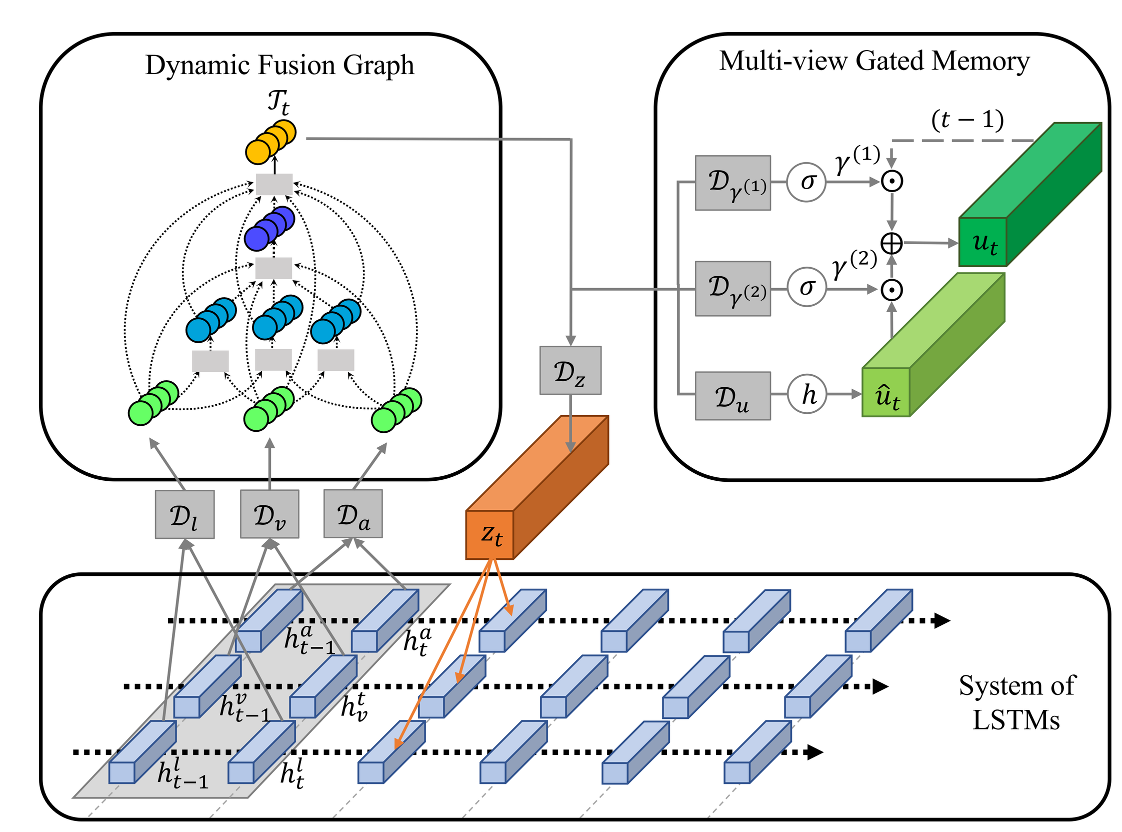
The structure of the Dynamic Fusion Graph for three modalities of $\{(l)anguage, (v)ision, (a)coustic\}$. Dashed lines represent dynamic connections between vertices controlled by efficacies.

The **Dynamic Fusion Graph** has the following properties that makes it suitable for multimodal fusion:

1. Explicitly models **unimodal**, **bimodal** and **trimodal** representations.
2. **Dynamically alter its structure** and choose the ideal fusion graph based on the importance of individual representations. This is performed by learning **efficacies** along each edge connection.
3. Efficacies allows us to **interpret the interactions** between modalities during fusion.

Terminal vertex $\mathcal{T}$ summarizes the unimodal, bimodal and trimodal representations.

Graph Memory Fusion Network

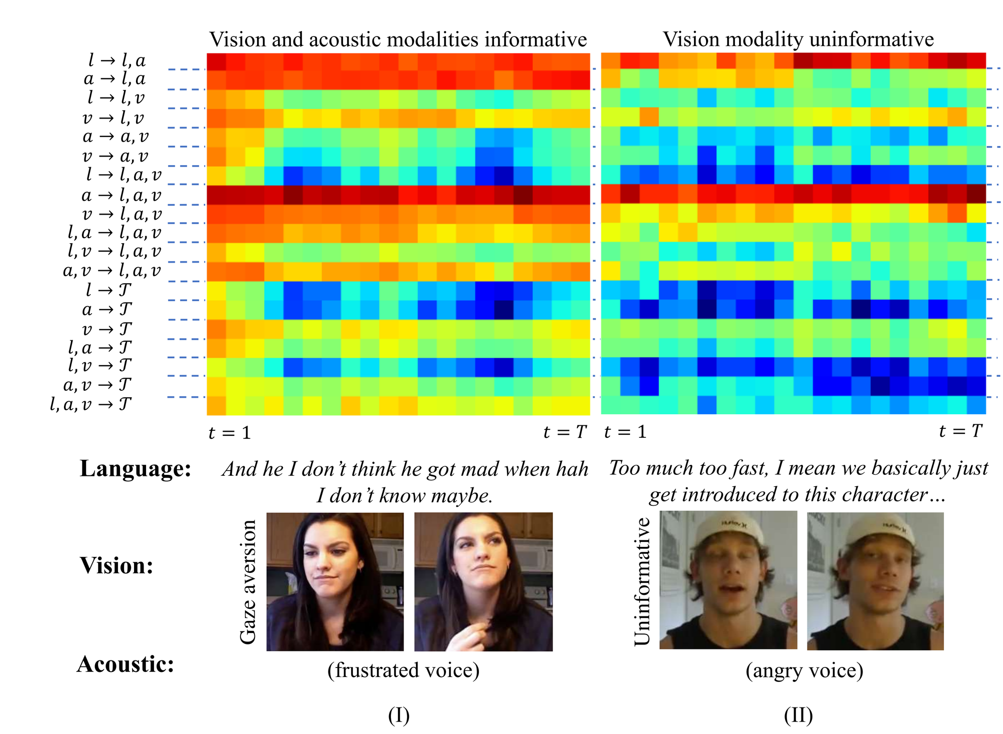


- The **Graph Memory Fusion Network** integrates the Dynamic Fusion Graph with the Memory Fusion Network [7].
- Each **Long Short-term Memory** encodes information from a single modality: language (l), visual (v) or acoustic (a).
- The **Dynamic Fusion Graph** learns multimodal representations from unimodal LSTM outputs h_t^l, h_t^v, h_t^a.
- The **Multi-view Gated Memory** u_t stores these multimodal representations and performs integration with the LSTM memories.
- The **outputs** of Graph Memory Fusion Network are the final state of the Multi-view Gated Memory and the outputs of each of the LSTMs.

Results on MOSEI

Dataset	MOSEI Sentiment					
Task	Sentiment					
Metric	A^2	F1	A^5	A^7	MAE	r
LANGUAGE						
SOTA2	74.1§	74.1$^{\triangleright}$	43.1$^{\wr}$	42.9$^{\wr}$	0.75§	0.46$^{\wr}$
SOTA1	74.3$^{\triangleright}$	74.1§	43.2§	43.2§	0.74$^{\triangleright}$	0.47§
VISUAL						
SOTA2	73.8§	73.5§	42.5$^{\triangleright}$	42.5$^{\triangleright}$	0.78$^{\wr}$	0.41$^{\heartsuit}$
SOTA1	73.9$^{\triangleright}$	73.7$^{\triangleright}$	42.7$^{\wr}$	42.7$^{\wr}$	0.78§	0.43$^{\wr}$
ACOUSTIC						
SOTA2	74.2$^{\wr}$	73.8$^{\triangle}$	42.1$^{\triangle}$	42.1$^{\triangle}$	0.78$^{\triangleright}$	0.43§
SOTA1	74.2$^{\triangle}$	73.9$^{\wr}$	42.4$^{\cap}$	42.4$^{\cap}$	0.74$^{\cap}$	0.43$^{\triangleright}$
MULTIMODAL						
SOTA2	76.6$^{\#}$	76.7$^{\blacksquare}$	44.5$^{\diamond}$	44.7$^{\diamond}$	0.71$^{\blacksquare}$	0.53$^{\blacksquare}$
SOTA1	76.7$^{\blacksquare}$	77.2$^{\flat}$	44.8$^{\blacksquare}$	44.7$^{\blacksquare}$	0.71$^{\#}$	0.54$^{\#}$
GMFN	**77.4**	**77.3**	**45.1**	**45.0**	**0.70**	**0.55**
Δ_{SOTA}	**↑ 0.7**	**↑ 0.1**	**↑ 0.3**	**↑ 0.3**	**↓ 0.01**	**↑ 0.01**

Interpretable Fusion



Visualization of Dynamic Fusion Graph efficacies across time. Dark red: high efficacies, dark blue: low efficacies.

1. **Multimodal Fusion is Volatile**. The Dynamic Fusion Graph dynamically adjusts efficacies of fusion depending on the given video.
2. **Efficacies to Terminal Vertex**. Unimodal efficacies to terminal vertex are low: model tends to rely on bimodal and trimodal representations.
3. **Priors of Human Communication**. High efficacies between language and acoustic modalities: natural priors of human communication.

Acknowledgements

Collaborator: Amir Zadeh. Helped in data collection and annotation: Jon Vanbriessen, Edmund Tong, Minghai Chen, Soujanya Poria and Erik Cambria. Helped in building CMU Multimodal SDK: Prateek Vij, Zhun Liu.

References

[1] R. Socher, *et al.* Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *EMNLP*, 2013.

[2] L.-P. Morency, *et al.* Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In *ICMI*, 2011.

[3] A. Mass, *et al.* Learning Word Vectors for Sentiment Analysis. In *ACL*, 2011.

[4] J. Pennington, *et al.* GloVe: Global Vectors for Word Representations. In *EMNLP*, 2014.

[5] F. Schroff, *et al.* FaceNet: A Unified Embedding for Face Recognition and Clustering. In *CVPR*, 2015.

[6] G. Degottex, *et al.* Covarep - A Collaborative Voice Analysis Repository for Speech Technologies. In *ICASSP*, 2014.

[7] A. Zadeh, *et al.* Memory Fusion Network for Multi-view Sequential Learning. In *AAAI*, 2018.