Language Technologies Institute

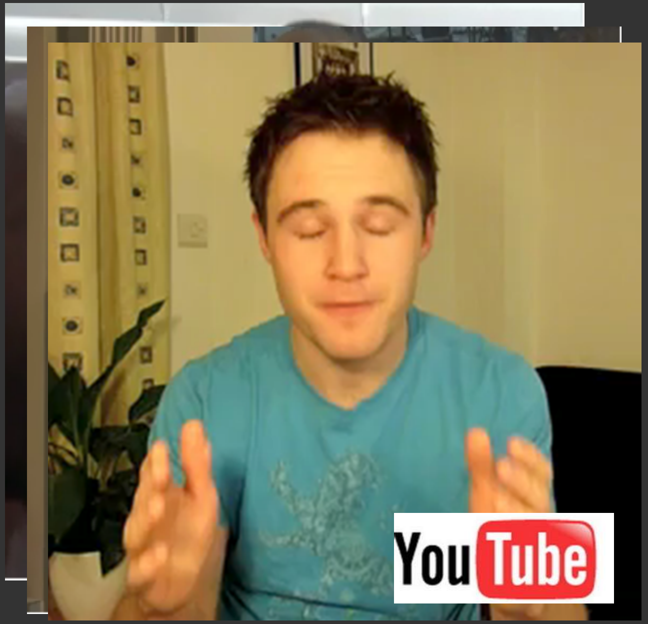
Carnegie Mellon University

Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning

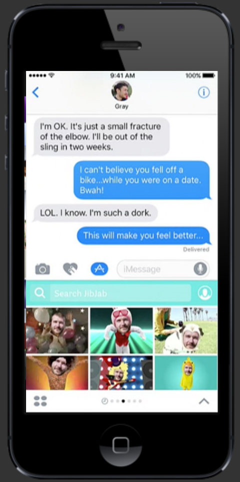
Minghai Chen*, Sen Wang*, Paul Pu Liang*,
Tadas Baltrusaitis, Amir Zadeh, Louis-Philippe Morency

Natural Computer Interaction

Parasocial Interactions
(e.g., multimedia content)



Intelligent Personal Assistant



Robots and Virtual Agents

Multimodal Communicative Behaviors

Verbal
- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

Vocal
- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Visual
- **Gestures**
  - Head gestures
  - Eye gestures
  - Arm gestures
- **Body language**
  - Body posture
  - Proxemics
- **Eye contact**
  - Head gaze
  - Eye gaze
- **Facial expressions**
  - FACS action units
  - Smile, frowning

Sentiment
- Positive
- Negative

Emotion
- Anger
- Disgust
- Fear
- Happiness
- Sadness
- Surprise

Social
- Empathy
- Engagement
- Dominance

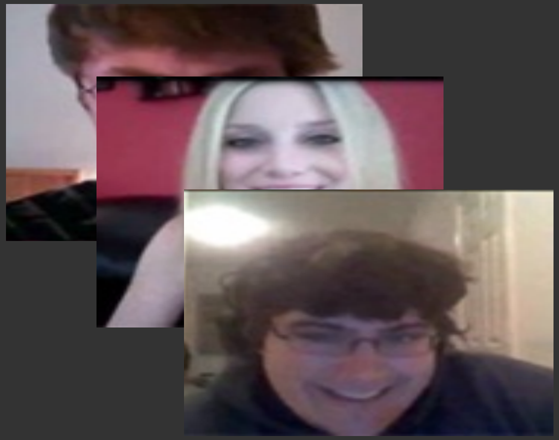
Multimodal Sentiment Analysis

Sentiment

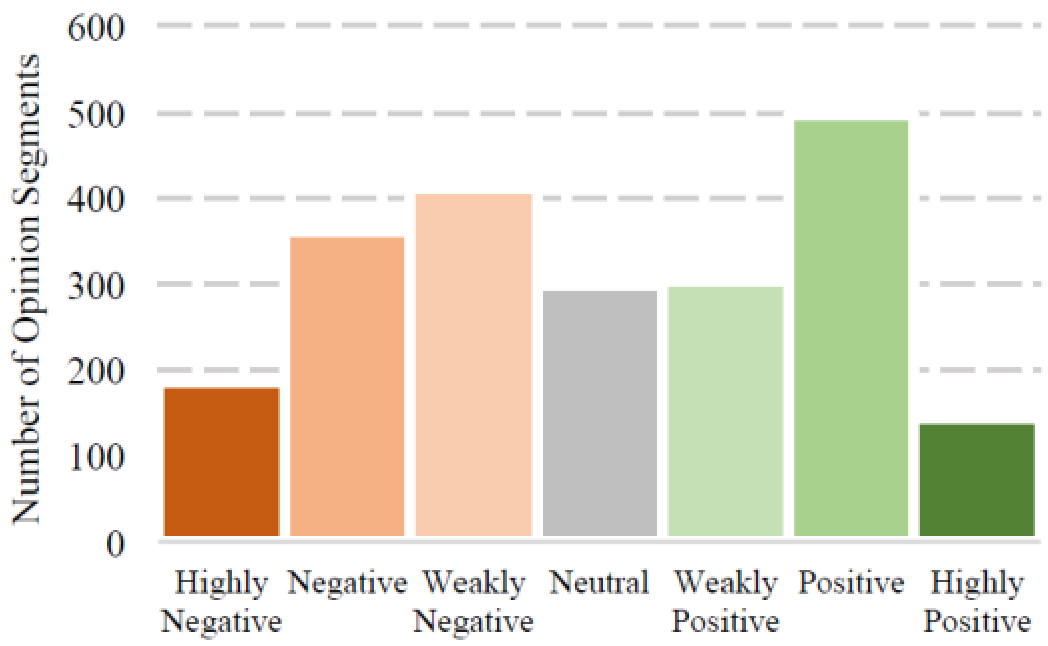
- Highly positive
- Positive
- Weakly positive
- Neutral
- Weakly negative
- Negative
- Highly negative

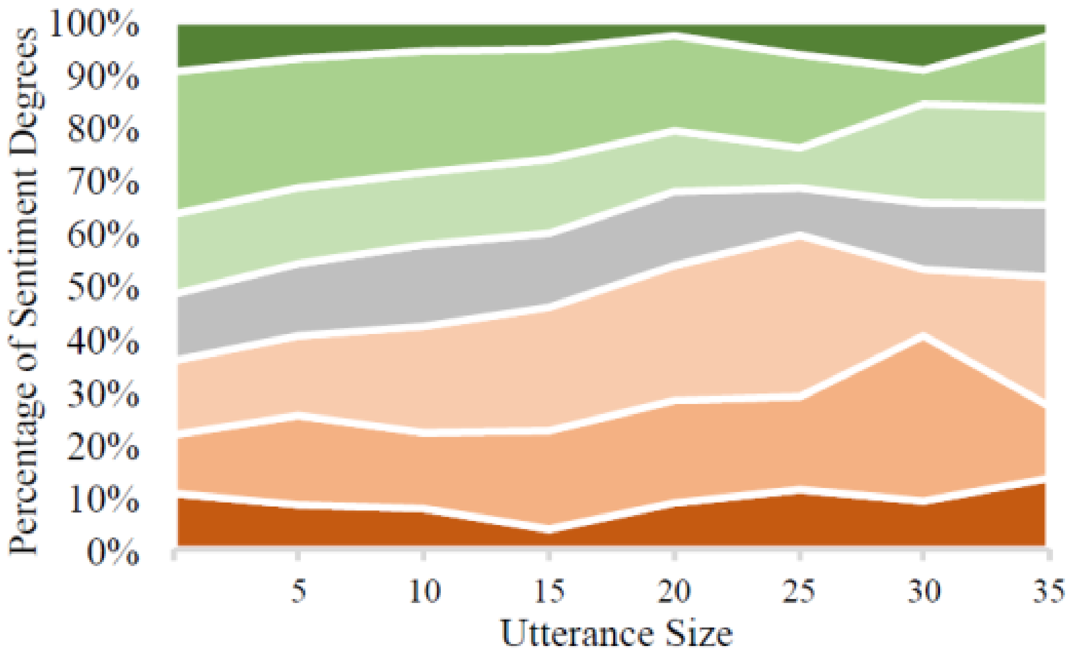
CMU-MOSI Dataset

- 93 videos of movie reviews
  - 89 distinct speakers
  - 48 male and 41 female speakers
- 2199 opinion segments
  - Average length: 4.2 sec
  - Average word count: 12
- 5 different annotators for each opinion segment
  - Krippendorf’s Alpha: 0.77



CMU-MOSI Dataset





Three Main Challenges Addressed in This Work

1 **What granularity should we use?**

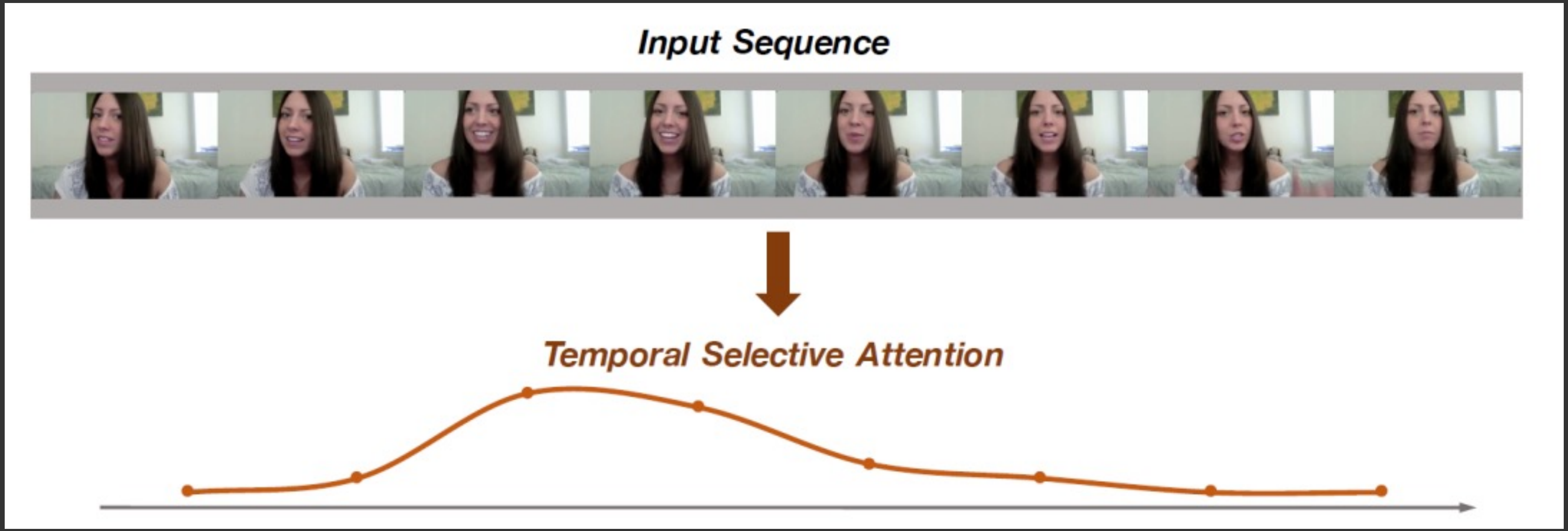
- Conventional approach summarizes features for the whole video
- But some multimodal interactions happen at the word level:
  - The word “crazy” with smile: Positive
  - The word “crazy with frown: Negative

Three Main Challenges Addressed in This Work

2 **What if a modality is noisy (e.g., occlusion)?**

Three Main Challenges Addressed in This Work

3 **What part of the video is relevant for the prediction task?**



Main Contributions

1. **What granularity should we use?**

   → Word-level feature representation

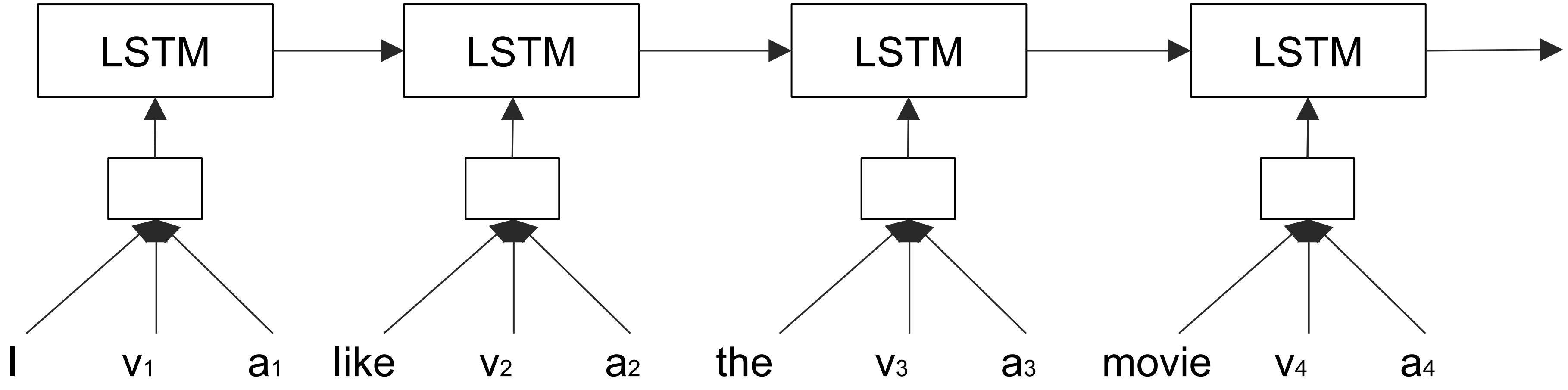
2. **What if a modality is noisy (e.g., occlusion)?**

   → Modality-specific “on/off gate”

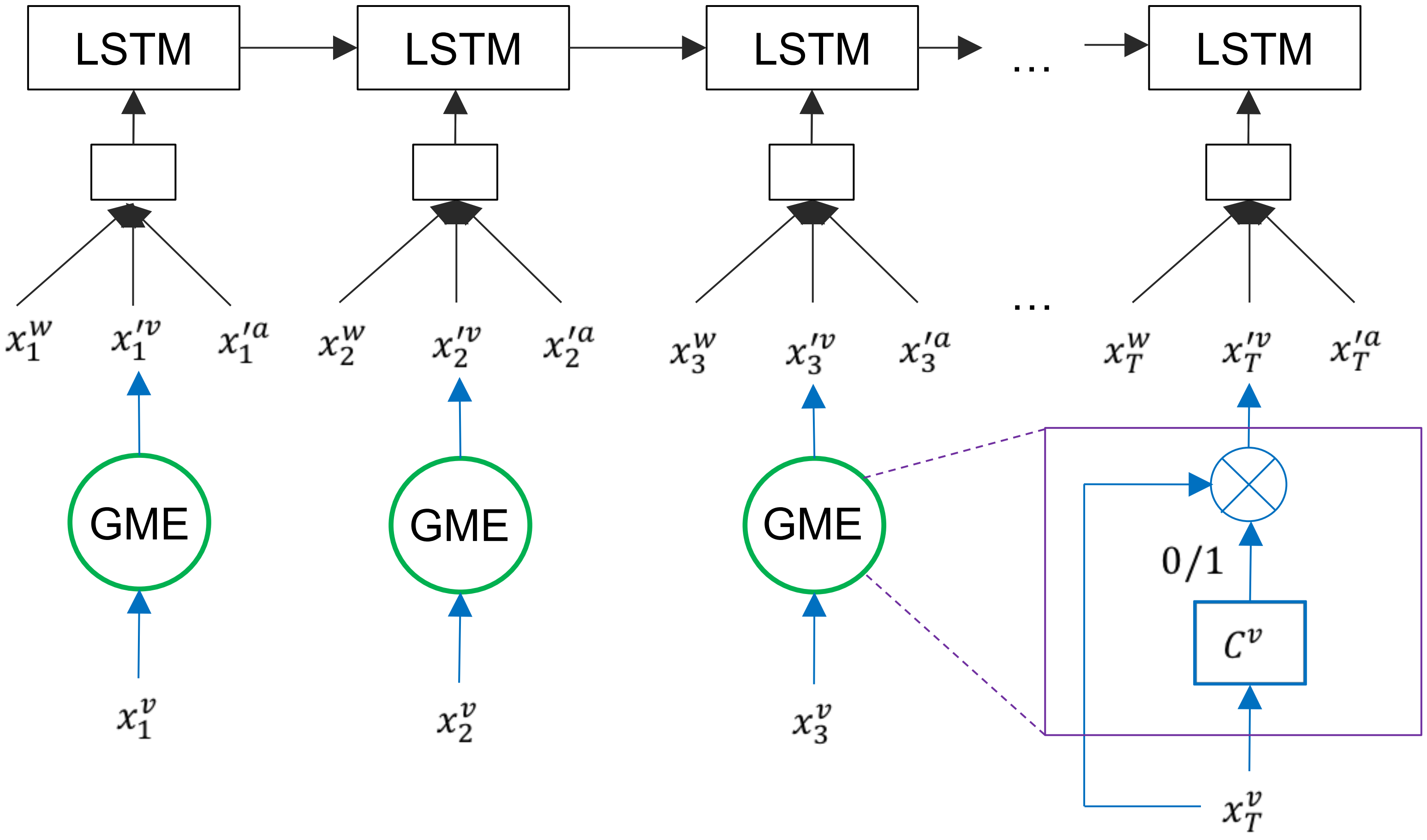
3. **What part of the video is relevant for the prediction task?**

   → Temporal attention

Challenge 1: LSTM with Word-Level Fusion



Challenge 2: Gated Multimodal Embedding (GME)

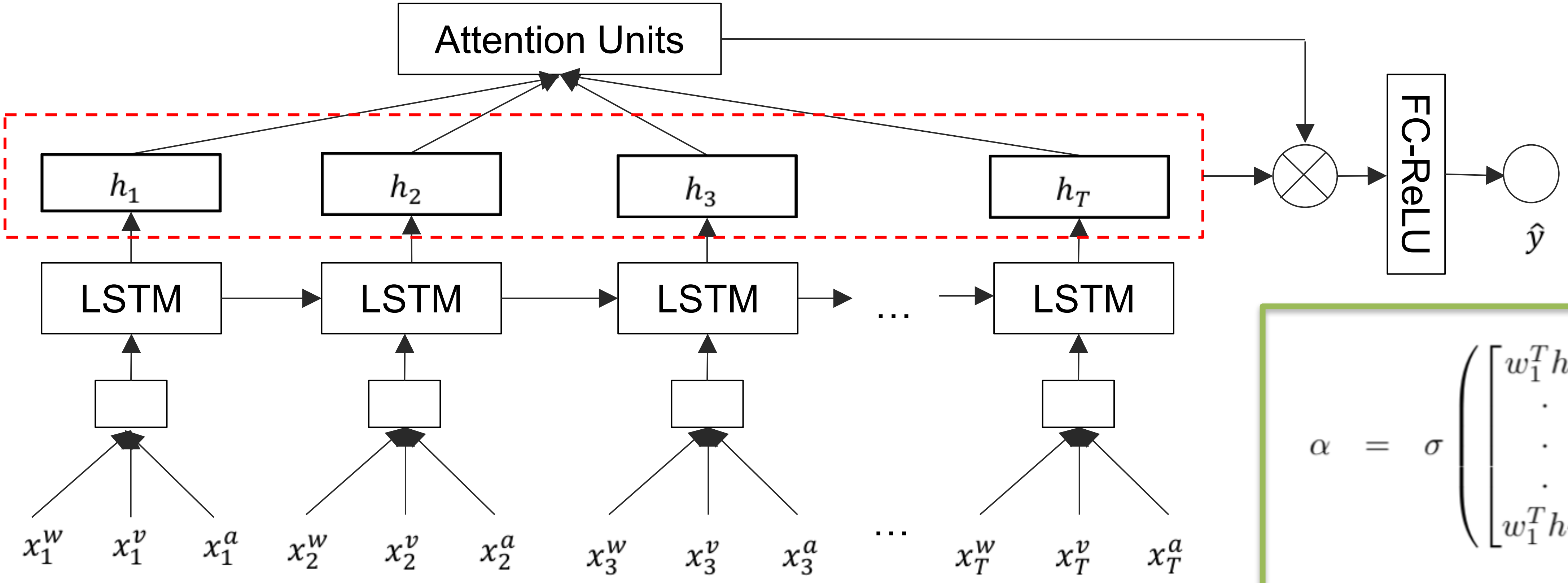


$$x_t'^w = x_t^w$$

$$x_t'^a = c_t^a \cdot x_t^a = C^a(x_t^a) \cdot x_t^a$$

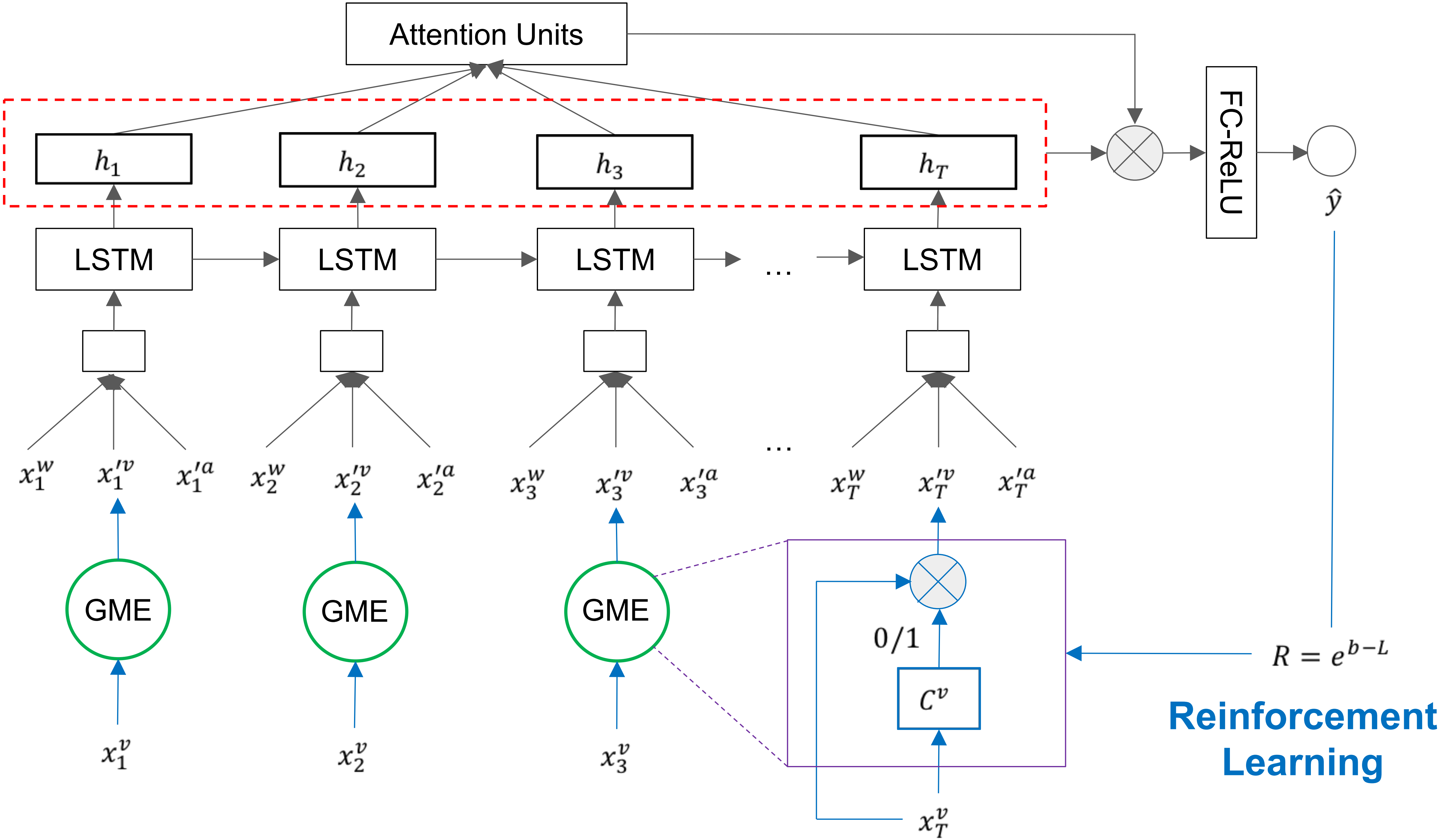
$$x_t'^v = c_t^v \cdot x_t^v = C^v(x_t^v) \cdot x_t^v$$

Challenge 3: LSTM with Temporal Attention



$$\alpha = \sigma\left(\begin{bmatrix} w_1^T h_1 \\ \cdot \\ \cdot \\ \cdot \\ w_1^T h_T \end{bmatrix}\right)$$

$$\hat{y} = w_2^T z \qquad z = H\alpha$$


Attention Units
h_1
h_2
h_3
h_T
LSTM
LSTM
LSTM
...
LSTM
FC-ReLU
$\hat{y}$
x_1^w $x_1'^v$ $x_1'^a$
x_2^w $x_2'^v$ $x_2'^a$
x_3^w $x_3'^v$ $x_3'^a$
...
x_T^w $x_T'^v$ $x_T'^a$
GME
GME
GME
0/1
C^v
$R = e^{b-L}$
Reinforcement Learning
x_1^v
x_2^v
x_3^v
x_T^v

Experiments

Text

- Transcripts of videos as well as pre-trained Glove word embeddings

Audio

- Covarep to extract acoustic features

Video

- Facet and Openface to extract facial landmarks, head pose, gaze tracking etc.

Baseline Models

- **C-MKL**: Convolutional Multi-Kernel Learning model. CNN to extract textual features and uses for fusion. (Poria et al., 2015)
- **SAL-CNN:** Select-Additive Learning. Reduces impact of identity-specific information. (Wang et al., 2016)
- **SVM-MD**: Support Vector Machine with Multimodal Dictionary. Multimodal features using early fusion. (Zadeh et al., 2016b)
- **RF**: Random Forest

Results – Multimodal Predictions

Method	Acc	F-score	MAE
Random	50.2	48.7	1.880
SAL-CNN	73.0	-	-
SVM-MD	71.6	72.3	1.100
C-MKL	73.5	-	-
RF	57.4	59.0	-
No Attention → LSTM	69.4	63.7	1.245
Without GME → LSTM(A)	75.7	72.1	1.019
Our model → **GME-LSTM(A)**	**76.5**	**73.4**	**0.955**
Human	85.7	87.5	0.710
Δ^{SOTA}	3.0 ↑	1.1 ↑	0.145 ↓

Results – Text Only

Method	Acc	F-score	MAE
RNTN	(73.7)	(73.4)	(0.990)
DAN	70.0	69.4	-
D-CNN	69.0	65.1	-
SAL-CNN text	73.5	-	-
SVM-MD text	73.3	72.1	1.186
RF text	57.6	57.5	-
LSTM text	67.8	51.2	1.234
LSTM(A) text	71.3	67.3	1.062
GME-LSTM(A)	**76.5**	**73.4**	**0.955**

LSTM with Word-Level Features

Modalities	**Acc**	**F-score**	**MAE**
text	67.8	51.2	1.234
audio	44.9	61.9	1.511
video	44.9	61.9	1.505
text+audio	66.8	55.3	**1.211**
text+video	63.0	**65.6**	1.302
text+audio+video	**69.4**	63.7	1.245

LSTM with Temporal Attention (LSTM(A))

Modalities	Acc	F-score	MAE
text	71.3	67.3	1.062
audio	55.4	63.0	1.451
video	52.3	57.3	1.443
text+audio	73.5	70.3	1.036
text+video	74.3	69.9	1.026
text+audio+video	**75.7**	**72.1**	**1.019**

Temporal Attention on Word features

But a lot of the footage was kind of ***unnecessary.***

And she really ***enjoyed*** *the film.*

I thought it was ***fun.***

So yes I really ***enjoyed*** *it.*

Example from LSTM with Temporal Attention

Transcript: *He's not gonna be looking like a chirper bright young man but early thirties really you **want** me to buy that.*

Visual modality: **Looks disappointed**

LSTM sentiment prediction: **1.24**

LSTM(A) sentiment prediction: **-0.94**

Ground truth sentiment: **-1.8**

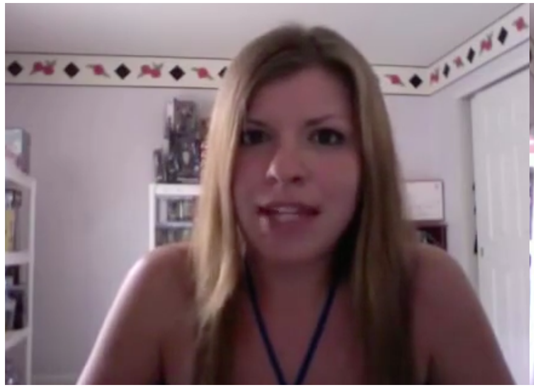
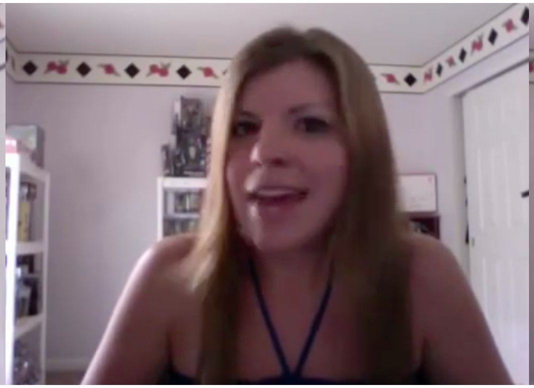
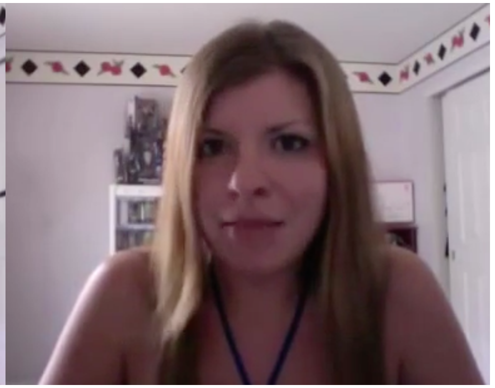
Example for Gated Multimodal Embedding

Transcript: *First of all I'd like to say little James or Jimmy he's so cute he's so xxx.*
LSTM(A) Attention: ***little*** (her mouth is covered by her hands)
GME-LSTM(A) Attention: ***cute***

LSTM(A) prediction: **-0.94**
GME-LSTM(A) prediction: **1.57**
Ground truth: **3.0**

Video example showing the effect of GME

GME Analysis

Visual RL Gate: Reject Pass Reject

LSTM(A) prediction: **-2.0032**

GME-LSTM(A) prediction: **1.4835**

Ground truth: **1.2**

Main Contributions

1 **What granularity should we use?**

→ Word-level feature representation

2 **What if a modality is noisy (e.g., occlusion)?**

→ Modality-specific “on/off gate”

3 **What part of the video is relevant for the prediction task?**

→ Temporal attention

MERCI !