Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion

Presenter: Paul Pu Liang
Paul Pu Liang, Ziyin Liu, Amir Zadeh, Louis-Philippe Morency
Progress of Artificial Intelligence

Multimedia Content

Intelligent Personal Assistants

Robots and Virtual Agents

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Multimodal Language Modalities

**Language**
- Lexicon
- Syntax
- Pragmatics

**Visual**
- Gestures
- Body language
- Eye contact
- Facial expressions

**Acoustic**
- Prosody
- Vocal expressions

Paul Pu Liang
Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Multimodal Language Modalities

- **Language**
  - Lexicon
  - Syntax
  - Pragmatics

- **Visual**
  - Gestures
  - Body language
  - Eye contact
  - Facial expressions

- **Acoustic**
  - Prosody
  - Vocal expressions

- **Sentiment**
  - Positive
  - Negative

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

- **Personality**
  - Confidence
  - Persuasion
  - Passion
Challenge 1: Temporal Intra-modal Interactions

a) Temporal sequences

<table>
<thead>
<tr>
<th>Speaker’s behaviors</th>
<th>Sentiment Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>“This movie is great”</td>
<td>++</td>
</tr>
<tr>
<td>Smile</td>
<td>Head nod</td>
</tr>
</tbody>
</table>

Intra-modal

time

Paul Pu Liang

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Challenge 2: Spatial Cross-modal Interactions

a) Multiple co-occurring interactions
b) Different weighted combinations

Speaker’s behaviors

“This movie is great”
Smile
Loud voice

Sentiment Intensity

Carnegie Mellon University

Paul Pu Liang

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Multistage Aggregation in Humans

(Parsini et al. 2015, Taylor et al. 2017)
Multistage Aggregation in Humans

(Parsini et al. 2015, Taylor et al. 2017)
Multistage Aggregation in Humans

(Parsini et al. 2015, Taylor et al. 2017)

wide smile loud voice → positive reaction positive words → excitement joyous
Computational Model for Multistage Fusion

- wide smile
- loud voice
- positive reaction
- positive words
- excitement
- joyous

Computational Model
Multimodal Descriptors

Language

He’s

average

Visual

Acoustic

multimodal descriptors

...
Language Descriptors

Language: He’s

Visual

Acoustic

multimodal descriptors
neutral word

average
Visual Descriptors

Language
He’s

Visual

Acoustic

multimodal descriptors
neutral word
frown
shrug

average

time...
Acoustic Descriptors

Language

Visual

Acoustic

He’s

average

multimodal descriptors

neutral word

frown

shrug

loud voice

speech elongation

...
Multistage Fusion
Multistage Fusion

stage 1

Neutral word
frown

shrug
loud voice
speech elongation
...
Multistage Fusion

Stage 1

- negative
- negative

Highlight

- neutral word
- frown

- shrug
- loud voice
- speech elongation
Multistage Fusion

- Stage 1:
  - FUSE
  - Negative

- Stage 2:
  - Highlight
  - Neutral word
  - Frown
  - Shrugs
  - Loud voice
  - Speech elongation
  - Neutral word
  - Frown
  - Shrugs
  - Loud voice
  - Speech elongation
Multistage Fusion

Stage 1
- negative
- neutral word
- shrug
- loud voice
- speech elongation

Stage 2
- emphasis
- neutral word
- shrug
- loud voice
- speech elongation
Multistage Fusion

Stage 1
- negative
- negative

Stage 2
- strongly negative
- emphasis

Highlights:
- neutral word
- frown
- shrug
- loud voice
- speech elongation

Emphasis:
- strongly negative
- shrug
- loud voice
- speech elongation
Multistage Fusion
Multistage Fusion

- **Stage 1**:
  - Negative
  - Neutral word
  - Frown
  - Shrug
  - Loud voice
  - Speech elongation

- **Stage 2**:
  - Strongly negative
  - Emphasis
  - Neutral word
  - Frown
  - Shrug
  - Loud voice
  - Speech elongation

- **Stage 3**:
  - Ambivalence
  - Neutral word
  - Frown
  - Shrug
  - Loud voice
  - Speech elongation

HIGHLIGHT

FUSE
Multistage Fusion

stage 1
- negative
- negative

stage 2
- strongly negative
- emphasis

stage 3
- disappointed
- ambivalence

HIGHLIGHT
- neutral word
- frown
- shrug
- loud voice
- speech elongation

...
Intra-modal Recurrent Networks
Multistage Fusion Process

Paul Pu Liang

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Multistage Fusion Process

Paul Pu Liang

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Multistage Fusion Process

Paul Pu Liang

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Multistage Fusion Process
Multistage Fusion Process

Paul Pu Liang

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Multistage Fusion Process

\[ h_t^l h_t^v h_t^a \]

Highlight LSTM

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Multistage Fusion Process

Fuse LSTM

Highlight LSTM

$h_t^l$, $h_t^v$, $h_t^a$
Multistage Fusion Process

Paul Pu Liang

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Multistage Fusion Process

Fuse LSTM

Highlight LSTM

$h_t^l$, $h_t^v$, $h_t^a$

Paul Pu Liang

Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion
Recurrent Multistage Fusion Network

Multistage Fusion Process

\( \mathbf{h}_t, \mathbf{h}^v_t, \mathbf{h}^a_t \)

stage 1 \( \rightarrow \) stage 2 \( \rightarrow \cdots \) stage \( K \)

FUSE \( \rightarrow \) FUSE \( \rightarrow \) FUSE \( \rightarrow \) SUMMARIZE

\( Z_t \)

LSTHM \( l \)

LSTHM \( v \)

LSTHM \( a \)

time \( t \)

time \( t + 1 \)
Recurrent Multistage Fusion Network

Multistage Fusion Process

stage 1

stage 2

... stage $K$

FUSE

FUSE

FUSE

SUMMARIZE

$h_t^l, h_t^v, h_t^a$

$Z_t$

LSTHM $l$

LSTHM $v$

LSTHM $a$

time $t$

time $t + 1$
Baseline Models

1. Non-temporal Models
   - SVM (Cortes and Vapnik, 1995), DF (Nojavanasghari et al., 2016)
2. Early Fusion
   - EF-LSTM (Hochreiter and Schmidhuber, 1997), EF-RHN (Zilly et al., 2016)
3. Late Fusion
   - LMF (Liu et al., 2018), TFN (Zadeh et al., 2017), BC-LSTM (Poria et al., 2017)
4. Multi-view Learning
   - MV-LSTM (Rajagopalan et al., 2016)
5. Memory-based models
   - MARN, MFN (Zadeh et al., 2018)
State-of-the-art Results

CMU-MOSI Sentiment (Binary Accuracy)

Baseline Models

RMFN

78.4%
State-of-the-art Results

CMU-MOSI Sentiment (Correlation)

POM Personality Traits (Multiclass Accuracy)

IEMOCAP Happy Emotion (Binary Accuracy)

IEMOCAP Sad Emotion (Binary Accuracy)

Best Baseline Model

RMFN
Results

IEMOCAP Neutral Emotion (Binary Accuracy)

- **Best Baseline Model**
- **RMFN**
Multiple Stages are Important

CMU-MOSI Sentiment Analysis (Binary Accuracy)

CMU-MOSI Sentiment Analysis (Multiclass Accuracy)
<table>
<thead>
<tr>
<th><strong>Language</strong></th>
<th>I thought it was fun</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Visual</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Acoustic</strong></td>
<td>(elongation)</td>
</tr>
<tr>
<td></td>
<td>(emphasis)</td>
</tr>
</tbody>
</table>
Interpretable Fusion

Language
I thought it was fun

Visual

Acoustic
(elongation)
(emphasis)
Interpretable Fusion

Language

I thought it was fun

Visual

Acoustic

(elongation)

(emphasis)
Interpretable Fusion

Language
I thought it was fun

Visual

Acoustic
(elongation)
(emphasis)
Across Stages

Language  Visual  Acoustic

I thought it was fun

(elongation)

(emphasis)
Across Time

Language

Visual

Acoustic

I thought it was fun

(elongation)

(emphasis)
Multimodal Priors

Language

Visual

Acoustic

I thought it was fun

(elongation)

(emphasis)
### Synchronized Interactions

<table>
<thead>
<tr>
<th>Language</th>
<th>Visual</th>
<th>Acoustic</th>
</tr>
</thead>
<tbody>
<tr>
<td>I thought it was fun</td>
<td>(elongation)</td>
<td>(emphasis)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stages</th>
<th>t = 1</th>
<th>t = 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The diagram shows the synchronization of interactions across stages with indicators for high and low activity levels.
Synchronized Interactions

Language

Visual

Acoustic

I thought it was fun

(elongation)

(emphasis)
Synchronized Interactions

Language

I thought it was fun

Visual

Acoustic

(elongation)

(emphasis)
Asynchronous Trimodal Interactions

Language: He delivers a lot of intensity

Diagram with stages 123 for t = 1 and t = 6, showing high and low values.
Asynchronous Trimodal Interactions

Language: He delivers a lot of intensity

Visual: (smile) (smile)

Acoustic: (emphasis)
Asynchronous Trimodal Interactions

Language:
He delivers a lot of intensity

Visual:
(smile) (smile)

Acoustic:
(emphasis)
Recurrent Multistage Fusion Network

Multistage Fusion Process

stage 1  stage 2  ...  stage $K$

FUSE  FUSE  FUSE  SUMMARIZE

$\mathbf{h}_t^l, \mathbf{h}_t^v, \mathbf{h}_t^a$

LSTHM $l$

LSTHM $v$

LSTHM $a$

time $t$

$Z_t$

time $t + 1$