Multimodal Language Analysis 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

Multimodal Language Analysis with Recurrent Multistage Fusion
Multimodal Language Modalities

**Language**
- Lexicon
- Syntax
- Pragmatics

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

**Acoustic**
- Prosody
- Vocal expressions
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: Intra-modal Interactions

a) Temporal sequences

**Speaker’s behaviors**

- “This movie is great”
- Smile
- Head nod

**Sentiment Intensity**

- ++
- +
Challenge 2: Cross-modal Interactions

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

Speaker’s behaviors

“This movie is great”
Smile
Loud voice

Sentiment Intensity

Paul Pu Liang

Multimodal Language Analysis 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)
Computational Model for Multistage Fusion

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

Computational Model
Multimodal Descriptors

Language

Visual

Acoustic

He’s
average

multimodal descriptors

...
Language Descriptors

Language: He’s

Visual

Acoustic

multimodal descriptors
neutral word

average

time
Visual Descriptors

Language

He’s

Visual

Acoustic

multimodal descriptors

neutral word

frown

shrug

average

...
Acoustic Descriptors

Language

He’s

Visual

Acoustic

multimodal descriptors

neutral word
frown

shrug

loud voice

speech elongation

...
Multistage Fusion

neutral word
frown
shrug
loud voice
speech elongation
...
Multistage Fusion

stage 1

HIGHLIGHT

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

- **Stage 2**
  - neutral word
  - frown
  - shrug
  - speech elongation

FUSE

HIGHLIGHT
Multistage Fusion

**Stage 1**
- neutral word
- frown
- shrug
- loud voice
- speech elongation
- ... 

- negative

**Stage 2**
- neutral word
- frown
- shrug
- ... 

- emphasis

- loud voice
- speech elongation
- ...
Multistage Fusion

stage 1
- negative
- negative

stage 2
- strongly negative
- emphasis

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

FUSE
- negative
- negative

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

...
Multistage Fusion

FUSE

stage 1
- negative
- negative

HIGHLIGHT

neutral word
frown
- shrug
- loud voice
- speech elongation

stage 2
- strongly negative
- emphasis

neutral word
frown
- shrug

neutral word
frown
- shrug

neutral word
frown

...
Multistage Fusion

- Stage 1: negative
- Stage 2: strongly negative, emphasis
- Stage 3: ambivalence

Highlight:
- Neutral word: frown
- Shrugs
- Loud voice
- Speech elongation

Strong emphasis: strongly

Negative emphasis: strongly
Multistage Fusion

- **Stage 1**: Negative
- **Stage 2**: Strongly negative
- **Stage 3**: Disappointed

**Highlight**
- Neutral word: Frown
- Shrug
- Loud voice
- Speech elongation

**FUSE**
Intra-modal Recurrent Networks
Multistage Fusion Process

\[ h_t^l h_t^v h_t^a \]
Multistage Fusion Process

Multistage Fusion Process

\( h_t^l \ h_t^v \ h_t^a \)

HIGHLIGHT

Paul Pu Liang

Multimodal Language Analysis with Recurrent Multistage Fusion
Multistage Fusion Process

stage 1

$\mathbf{h}_t^l \mathbf{h}_t^v \mathbf{h}_t^a$
Multistage Fusion Process
Multistage Fusion Process

Paul Pu Liang

Multimodal Language Analysis with Recurrent Multistage Fusion
Multistage Fusion Process

Highlight LSTM

Multistage Fusion Process

stage 1

stage 2

HIGHLIGHT

HIGHLIGHT

FUSE

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

Fuse LSTM

Highlight LSTM

$h_t^l, h_t^v, h_t^\alpha$
Multistage Fusion Process

Fuse LSTM

Highlight LSTM

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

Fuse LSTM

Highlight LSTM

HIGHLIGHT

SUMMARIZE

stage 1

stage 2

... stage $K$

$h_t^l$ $h_t^v$ $h_t^a$
Recurrent Multistage Fusion Network

Multistage Fusion Process

stage 1

FUSE

HIGHLIGHT

stage 2

FUSE

HIGHLIGHT

... stage K

FUSE

HIGHLIGHT

SUMMARIZE

Z_t

h_t^l \ h_t^v \ h_t^a

LSTHM \ l

LSTHM \ v

LSTHM \ a

time t

time t + 1
Recurrent Multistage Fusion Network
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)

- **MV-LSTM**
- **RMFN**
- **MFN**
- **RMFN**

POM Personality Traits (Multiclass Accuracy)

- **MV-LSTM**
- **RMFN**
- **MFN**
- **MARN**
- **RMFN**

IEMOCAP Happy Emotion (Binary Accuracy)

- **RMFN**
- **MARN**
- **RMFN**
- **RMFN**

IEMOCAP Sad Emotion (Binary Accuracy)

- **MV-LSTM**
- **MFN**
- **RMFN**

Best Baseline Model

RMFN
Results

IEMOCAP Neutral Emotion (Binary Accuracy)

Best Baseline Model
RMFN

Paul Pu Liang
Multimodal Language Analysis with Recurrent Multistage Fusion
Multiple Stages are Important

CMU-MOSI Sentiment Analysis
(Binary Accuracy)

CMU-MOSI Sentiment Analysis
(Multiclass Accuracy)

Number of stages

1 2 3 4 5

1 2 3 4 5
## Ablation Studies

| Dataset               | CMU-MOSI Sentiment |
|-----------------------|--------------------|
|                       | A2 ↑    | F1 ↑    | A7 ↑    | MAE ↓   | Corr ↑ |
| MARN                  | 77.1    | 77.0    | 34.7    | 0.968   | 0.625  |
| RMFN (no MFP)         | 76.5    | 76.5    | 30.8    | 0.998   | 0.582  |
| RMFN (no HIGHLIGHT)   | 77.9    | 77.9    | 35.9    | 0.952   | 0.666  |
| RMFN                  | 78.4    | 78.0    | 38.3    | 0.922   | 0.681  |
Interpretable Fusion

| Language       | Visual                  | Acoustic                |
|----------------|-------------------------|-------------------------|
| I thought it   |                         | (elongation)            |
| was fun        |                         | (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)
Interpretable Fusion

Language

Visual

Acoustic

I thought it was fun

(elongation)

(emphasis)
Across Stages

Language

Visual

Acoustic

I thought it was fun

(elongation)

(emphasis)
Across Time

Language: I thought it was fun

Visual:

Acoustic (elongation) (emphasis)
Multimodal Priors

Language: I thought it was fun
Visual
Acoustic (elongation) (emphasis)
Synchronized Interactions

| Language  | Visual | Acoustic |
|-----------|--------|----------|
| I thought it was fun | | (elongation) |
| | | (emphasis) |

- $h_t^l$: low
- $h_t^v$: high
- $h_t^a$: elongation
# Synchronized Interactions

| Language | Visual | Acoustic |
|----------|--------|----------|
| I thought it was fun |        | (elongation) (emphasis) |

- $h_t^l$ (Language)
- $h_t^v$ (Visual)
- $h_t^a$ (Acoustic)

Stages: 123

$t = 1$ vs $t = 5$

High vs Low

Examples:
- Language: *I thought it was fun*
- Visual: Facial expressions
- Acoustic: (elongation) (emphasis)
Synchronized Interactions

Language: I thought it was fun
Visual: (elongation)
Acoustic: (emphasis)
Asynchronous Trimodal Interactions

Language: He delivers a lot of intensity
Asynchronous Trimodal Interactions

Language: He delivers a lot of intensity

Visual: (smile) (smile)

Acoustic: (emphasis)
Asynchronous Trimodal Interactions

He delivers a lot of intensity

(smile) (smile)

(emphasis)
Bimodal Interactions

Language: It doesn’t give any insight or help

Visual

Acoustic: (emphasis), (soft), (disappointed)
Bimodal Interactions

It doesn’t give any insight or help. (emphasis) (soft) (disappointed)
Recurrent Multistage Fusion Network

Multistage Fusion Process

stage 1  stage 2  ...  stage K

FUSE  FUSE  FUSE  SUMMARIZE

h^l_t  h^v_t  h^α_t

h^l_t  h^v_t  h^α_t

LSTHM \textit{l}  LSTHM \textit{v}  LSTHM \textit{α}

time \textit{t}  time \textit{t} + 1
The End!

Website: www.cs.cmu.edu/~pliang

Email: pliang@cs.cmu.edu

Twitter: @pliang279

Multistage Fusion Process

stage 1  stage 2  ...  stage K

FUSE  FUSE  ...  FUSE

SUMMARIZE

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