Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph

Presenter: Paul Pu Liang
Amir Zadeh, Paul Pu Liang, Jonathan Vanbriessen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, Louis-Philippe Morency
Progress of Artificial Intelligence

Multimedia Content

Intelligent Personal Assistants

Robots and Virtual Agents

Paul Pu Liang

Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph
Throughout evolution language and nonverbal behaviors developed together.

Cries and Imitations

Modern Language
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
Multimodal Language Modalities

Language        Visual        Acoustic

Datasets

Sentiment
Emotion
Personality

Models
Multimodal Language Modalities

Language  Visual  Acoustic

Datasets  Models

Sentiment  Emotion  Personality

Paul Pu Liang
Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph
Multimodal Language Modalities

- Language
- Visual
- Acoustic

Datasets

- Large-scale
- Diverse

Models

- Sentiment
- Emotion
- Personality

Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph
Multimodal Language Modalities

- Language
- Visual
- Acoustic

Datasets

Models
- Word-level alignment
- Attention models
- Memory-based models

✓ Large-scale
✓ Diverse

Sentiment
Emotion
Personality
Multimodal Language Modalities

Datasets
- Large-scale
- Diverse

Models
- Word-level alignment
- Attention models
- Memory-based models

✓ Good Performance
✓ Interpretable
Datasets for Multimodal Language

- Require large and diverse amounts of data: (Novelty)
  - Diversity in samples
Datasets for Multimodal Language

- Require large and diverse amounts of data: (Novelty)
  - Diversity in samples
  - Diversity in topics
Datasets for Multimodal Language

- Require large and diverse amounts of data: (Novelty)
  - Diversity in **samples**
  - Diversity in **topics**
  - Diversity in **speakers**
Datasets for Multimodal Language

- Require large and diverse amounts of data: \((\text{Novelty})\)
- Diversity in **samples**
- Diversity in **topics**
- Diversity in **speakers**
- Diversity in **annotations**
New Dataset: CMU-MOSEI

23,000 video segments
3 modalities

Language: And he I don’t think he got mad when hah I don’t know maybe. Too much too fast, I mean we basically just get introduced to this character... All I can say is he’s a pretty average guy.

Vision: Gaze aversion

Acoustic: (frustrated voice) Uninformative (angry voice) Contradictory smile (disappointed voice)
CMU-MOSEI Dataset

1,000 speakers
250 topics
Annotation Distributions

![Bar chart showing annotation distributions]

- Negative
- Weakly Negative
- Neutral
- Weakly Positive
- Positive

Paul Pu Liang
Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph
Annotation Distributions

![Bar chart showing annotation distributions for different sentiments and emotions. The x-axis represents categories such as Negative, Weakly Negative, Neutral, Weakly Positive, Positive, Happiness, Sadness, Anger, Disgust, Surprise, Fear. The y-axis represents the number of annotations.]
Feature Extraction

Language
- Glove word embeddings

Visual
- Facet features
- MultiComp OpenFace
- Face embeddings

Acoustic
- COVAREP features
  - MFCCs
  - Pitch tracking

Sentiment
- Positive
- Negative

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

Alignment
- Word level
- P2FA
# CMU-MOSEI Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th># S</th>
<th># Sp</th>
<th>Mod</th>
<th>Sent</th>
<th>Emo</th>
<th>TL (hh:mm:ss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU-MOSEI</td>
<td>23,453</td>
<td>1,000</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✓</td>
<td>65:53:36</td>
</tr>
<tr>
<td>CMU-MOSI</td>
<td>2,199</td>
<td>98</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✗</td>
<td>02:36:17</td>
</tr>
<tr>
<td>ICT-MMMMO</td>
<td>340</td>
<td>200</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✗</td>
<td>13:58:29</td>
</tr>
<tr>
<td>YouTube</td>
<td>300</td>
<td>50</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✗</td>
<td>00:29:41</td>
</tr>
<tr>
<td>MOUD</td>
<td>400</td>
<td>101</td>
<td>{l, v, a}</td>
<td>✓</td>
<td>✗</td>
<td>00:59:00</td>
</tr>
<tr>
<td>IEMOCAP</td>
<td>10,000</td>
<td>10</td>
<td>{l, v, a}</td>
<td>✗</td>
<td>✓</td>
<td>11:28:12</td>
</tr>
<tr>
<td>SST</td>
<td>11,855</td>
<td>–</td>
<td>{l}</td>
<td>✓</td>
<td>✗</td>
<td>–</td>
</tr>
<tr>
<td>Cornell</td>
<td>2,000</td>
<td>–</td>
<td>{l}</td>
<td>✓</td>
<td>✗</td>
<td>–</td>
</tr>
<tr>
<td>HUMAINE</td>
<td>50</td>
<td>4</td>
<td>{v, a}</td>
<td>✗</td>
<td>✓</td>
<td>04:11:00</td>
</tr>
<tr>
<td>RECOLA</td>
<td>46</td>
<td>46</td>
<td>{v, a}</td>
<td>✗</td>
<td>✓</td>
<td>03:50:00</td>
</tr>
<tr>
<td>SEWA</td>
<td>538</td>
<td>408</td>
<td>{v, a}</td>
<td>✗</td>
<td>✓</td>
<td>04:39:00</td>
</tr>
<tr>
<td>SEMAINE</td>
<td>80</td>
<td>20</td>
<td>{v, a}</td>
<td>✗</td>
<td>✓</td>
<td>06:30:00</td>
</tr>
<tr>
<td>AFEW</td>
<td>1,645</td>
<td>330</td>
<td>{v, a}</td>
<td>✗</td>
<td>✓</td>
<td>02:28:03</td>
</tr>
</tbody>
</table>

## Multimodal

- **Multimodal Language**
- **Audio-visual**

Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph

Paul Pu Liang
Models for Multimodal Language

Multimodal Fusion

Transcript

Video clips

Visual gestures

Um...

...mm

this movie

is dumb.

Gaze Aversion

Frown

-
Models for Multimodal Language

multimodal

Multimodal Fusion

Interpretation
- Importance of each modality
- Interactions between modalities

Transcript

Um...

this movie

Video clips

Gaze Aversion

Frown

Visual gestures

is dumb.
Dynamic Fusion Graph (DFG)

Interpretation
- Importance of each modality
- Interactions between modalities

multimodal

unimodal

Transcript

Um...

...mm

this movie

is dumb.

Video clips

Gaze Aversion

Frown

Frustration

Visual gestures
Dynamic Fusion Graph (DFG)

**multimodal**

**bimodal**

**unimodal**

**Interpretation**

- Importance of each modality
- Interactions between modalities

Transcript

- Um...
- ...mm
- this movie
- is dumb.

Video clips

Visual gestures

Gaze Aversion

Frown

Frustration
Dynamic Fusion Graph (DFG)

- **multimodal**
- **trimodal**
- **bimodal**
- **unimodal**

**Interpretation**
- Importance of each modality
- Interactions between modalities

**Transcript**

\[\text{Um...} \quad \ldots \text{mm} \quad \text{this movie} \quad \text{is dumb.}\]

**Video clips**

- Gaze Aversion
- Frown

**Visual gestures**

- Frustration
Dynamic Fusion Graph (DFG)

**Interpretation**
- Importance of each modality
- Interactions between modalities

**fusion weights**

**multimodal**

**trimodal**

**bimodal**

**unimodal**

Transcript

<table>
<thead>
<tr>
<th>Transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>Um...</td>
</tr>
<tr>
<td>...mm</td>
</tr>
<tr>
<td>this movie</td>
</tr>
<tr>
<td>is dumb.</td>
</tr>
</tbody>
</table>

Video clips

Gaze Aversion

Frown

Visual gestures

Frustration
Dynamic Fusion Graph (DFG)

**multimodal**

**trimodal**

**bimodal**

**unimodal**

**Interpretation**
- Importance of each modality
- **Interactions between modalities**

```
\[
\sum_t \text{trimmed}
\]
```

```
\sum_i \text{trimmed}
```

```
\sum_i \text{bimodal}
```

```
\sum_i \text{unimodal}
```

```
\sum_i \text{multimodal}
```

Transcript: 
```
Um...
...mm
this movie
is dumb.
```

Video clips:
- Gaze Aversion
- Frown
- Frustration
Dynamic Fusion Graph (DFG)

$t = 1$

**multimodal**

**trimodal**

**bimodal**

**unimodal**

Transcript

Video clips

Visual gestures

Gaze Aversion

...mm

this movie

is dumb.
Dynamic Fusion Graph (DFG)

multimodal
trimodal
bimodal
unimodal

Transcript
Umm...
this movie
is dumb.

Video clips

Visual gestures
Gaze Aversion
Frown

$t = 1$

$t = 2$
Dynamic Fusion Graph (DFG)

- **multimodal**
- **trimodal**
- **bimodal**
- **unimodal**

Transcript: "Um... ...mm this movie is dumb.

Video clips:
- Gaze Aversion
- Frown
- Frustration
Dynamic Fusion Graph (DFG)

Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph

Trimodal

Bimodal

Unimodal

Transcript

<table>
<thead>
<tr>
<th>t = 1</th>
<th>t = 2</th>
<th>t = 3</th>
<th>t = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Um...</td>
<td>...mm</td>
<td>this movie</td>
<td>is dumb.</td>
</tr>
</tbody>
</table>

Video clips

Gaze Aversion  Frown  -  Frustration

Visual gestures
Dynamic Fusion Graph (DFG)

**Interpretation**
- Importance of each modality
- Interactions between modalities

**fusin weights**

**Multimodal**

**Trimodal**

**Bimodal**

**Unimodal**

**Transcript**

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Video clips</th>
<th>Visual gestures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Um...</td>
<td>...mm</td>
<td>this movie</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>is dumb.</td>
</tr>
</tbody>
</table>

**Gaze Aversion**

**Frown**

**Frustration**
Dynamic Fusion Graph (DFG)

**Multimodal**
- Importance of each modality
- Interactions between modalities

**Trimodal**
- Construction of bimodal and trimodal representations

**Bimodal**
- Construction weights

**Unimodal**
- Fusion weights

Transcript: *Um...* ...

Video clips: 
- *this movie*
- *is dumb.*

Visual gestures: 
- Gaze Aversion
- Frown
- Frustration
Dynamic Fusion Graph (DFG)

**Interpretation**
- Importance of each modality
- Interactions between modalities

**Fusion Weights**

**Construction Weights**

**Multimodal**

**Trimodal**

**Bimodal**

**Unimodal**

Transcript:
- Um...
- ...mm
- this movie
- is dumb.

Video clips:
- Gaze Aversion
- Frown

Visual gestures:
- Frustration
Dynamic Fusion Graph (DFG)

**Interpretation**
- Importance of each modality
- Interactions between modalities

**Construction of bimodal and trimodal representations**
- Construction weights
- Fusion weights

**Multimodal**
- Importance of each modality
- Interactions between modalities

**Trimodal**
- Construction of bimodal and trimodal representations

**Bimodal**
- Construction weights
- Fusion weights

**Unimodal**
- Importance of each modality
- Interactions between modalities
Dynamic Fusion Graph (DFG)

**multimodal**

**trimodal**

**bimodal**

**unimodal**

**Interpretation**

- Importance of each modality
- Interactions between modalities

**Construction weights**

**fusion weights**

- Construction of bimodal and trimodal representations
Dynamic Fusion Graph (DFG)

\[ \begin{align*}
\text{multimodal} & \quad \text{trimodal} & \quad \text{bimodal} & \quad \text{unimodal} \\
\text{t = 1} & \quad \text{t = 2} & \quad \text{t = 3} & \quad \text{t = 4}
\end{align*} \]

Transcript:
- Um...
- ...mm
- this movie
- is dumb.

Video clips:
- [Image of video clips]

Visual gestures:
- Gaze Aversion
- Frown
- -
- Frustration
Graph-Memory Fusion Network (Graph-MFN)

**multimodal**

**trimodal**

**bimodal**

**unimodal**

Transcript

*Um...*  
...*mm*  
*this movie*  
*is dumb.*

Video clips

Gaze Aversion  
Frown  
-

Visual gestures

Frustration
Graph-Memory Fusion Network (Graph-MFN)

- multimodal
- trimodal
- bimodal
- unimodal

Transcript

Um...

...mm

this movie

is dumb.

Video clips

Gaze Aversion

Frown

Visual gestures

Frustration
Graph-Memory Fusion Network (Graph-MFN)

- **multimodal**
- **trimodal**
- **bimodal**
- **unimodal**

Transcript:

- *Um...*
- *...mm*
- *this movie*
- *is dumb.*

Video clips:

- Gaze Aversion
- Frown

Visual gestures:

- Frustration
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
   - 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
   - MFN (Zadeh et al., 2018)
State-of-the-art Results

CMU-MOSEI Sentiment (Binary Accuracy)

76.9%

Baseline Models
Graph-MFN

SVM | DF | EF-RHN | EF-LSTM | TFN | BC-LSTM | MVLSTM | MARN | MFN | Graph-MFN

73 | 73.5 | 74 | 74.5 | 75 | 75.5 | 76 | 76.5 | 77
State-of-the-art Results

- **Sentiment (Correlation)**
  - MV-LSTM: 0.46
  - Graph-MFN: 0.51
  - TFN: 0.47

- **Sentiment (Multiclass Accuracy)**
  - MV-LSTM: 60.5
  - Graph-MFN: 63
  - TFN: 60

- **Anger Emotion (Binary Accuracy)**
  - MV-LSTM: 60
  - Graph-MFN: 63
  - TFN: 60.5

- **Disgust Emotion (Binary Accuracy)**
  - MV-LSTM: 60
  - Graph-MFN: 70
  - MN: 63

**Legend:**
- Blue: Best Baseline Model
- Red: Graph-MFN

Paul Pu Liang

Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph
Interpretable Fusion

$\quad t = 1 \quad t = T \quad t = 1 \quad t = T$

*And he I don’t think he got mad when hah
I don’t know maybe.*

Gaze aversion
(frustrated voice)

*Too much too fast, I mean we basically just get introduced to this character…*

Uninformative
(angry voice)
Interpretable Fusion

$t = 1$

And he I don’t think he got mad when hah
I don’t know maybe.

(t = 1)

(t = T)

Gaze aversion
(frustrated voice)

$t = 1$

Too much too fast, I mean we basically just
get introduced to this character...

(t = T)

Uninformative
(angry voice)
Interpretable Fusion

fusion weights

$t = 1$

And he I don’t think he got mad when hah
I don’t know maybe.

Gaze aversion
(frustrated voice)

$t = T$

Too much too fast, I mean we basically just get introduced to this character...

Uninformative
(angry voice)

$t = 1$

$t = T$
Interpretable Fusion

- **Fusion weights**
- **Construction weights**

**unimodal**

**bimodal**

**trimodal**

Text:

And he I don’t think he got mad when hah I don’t know maybe.

Too much too fast, I mean we basically just get introduced to this character...

Gaze aversion (frustrated voice)

Uninformative (angry voice)
Multimodal Fusion has a Dynamic Nature

And he I don’t think he got mad when hah I don’t know maybe.

Too much too fast, I mean we basically just get introduced to this character...

Uninformative

(frustrated voice)

(ungauged voice)
Priors in Human Multimodal Language

---

**Fusion weights**

**Construction weights**

---

T = 1  T = T  T = 1  T = T

**unimodal**

**bimodal**

**trimodal**

---

*And he I don’t think he got mad when hah, I don’t know maybe.*

*Too much too fast, I mean we basically just get introduced to this character...*

---

Gaze aversion (frustrated voice)

Uninformative (angry voice)
Priors in Human Multimodal Language

And he I don’t think he got mad when hah I don’t know maybe.

Too much too fast, I mean we basically just get introduced to this character...

Gaze aversion
(frustrated voice)

Uninformative
(angry voice)
Dynamic Selection of Modalities

- fusion weights
- construction weights

And he I don’t think he got mad when hah... I don’t know maybe.

Too much too fast, I mean we basically just get introduced to this character...

Gaze aversion (frustrated voice)

Uninformative (angry voice)

all modalities are informative
Dynamic Selection of Modalities

And he I don’t think he got mad when hah I don’t know maybe.

Too much too fast, I mean we basically just get introduced to this character...

Gaze aversion
(frustrated voice)

visual modality uninformative

Uninformative
(angry voice)
Computational Modeling of Multimodal Language

1. CMU-MOSEI Dataset

✓ Large-scale
✓ Diverse
Computational Modeling of Multimodal Language

1. CMU-MOSEI Dataset
   - Large-scale
   - Diverse

2. Dynamic Fusion Graph
   - Good Performance
   - Interpretable

Paul Pu Liang
Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph
The End!

Data: https://github.com/A2Zadeh/CMU-MultimodalSDK
Website: www.cs.cmu.edu/~pliang
Email: pliang@cs.cmu.edu
Twitter: @pliang279
The End!

Data: https://github.com/A2Zadeh/CMU-MultimodalSDK
Website: www.cs.cmu.edu/~pliang
Email: pliang@cs.cmu.edu
Twitter: @pliang279

Workshop @ 20 July 9am – 3pm, Room 217
First Grand Challenge and Workshop on
Human Multimodal Language