Computational Modeling of Human Multimodal Language

Paul Pu Liang

Joint work with: Amir Zadeh, Yao-Hung Hubert Tsai, Zhun Liu, Ying Shen, Hai Pham, Varun Lakshminarayanan, Edmund Tong, Jon Vanbriessen, Ruslan Salakhutdinov, Louis-Philippe Morency
Contents

- Human Multimodal Language
Contents

- Human Multimodal Language
- 5 directions:
  - Intra-modal and Cross-modal
  - Unimodal, Bimodal and Trimodal
  - Direct and Relative
  - Multimodal Representation Learning
  - Robust Multimodal Representation Learning
Contents

- Human Multimodal Language
- 5 directions:
  - Intra-modal and Cross-modal
  - Unimodal, Bimodal and Trimodal
  - Direct and Relative
  - Multimodal Representation Learning
  - Robust Multimodal Representation Learning
- New Multimodal Dataset: MOSEI
Human Multimodal Language

5 directions:
- Intra-modal and Cross-modal
- Unimodal, Bimodal and Trimodal
- Direct and Relative
- Multimodal Representation Learning
- Robust Multimodal Representation Learning

New Multimodal Dataset: MOSEI

Future directions
Progress of Artificial Intelligence

Multimedia Content

Intelligent Personal Assistants

Robots and Virtual Agents
Multimodal Communicative Behaviors

- **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
Direction 1: Intra-modal and Cross-modal
Challenge 1: Intra-modal dynamics

Speaker’s behaviors

“This movie is great”

Sentiment Intensity

+++

Paul Pu Liang
Computational Modeling of Human Multimodal Language
Challenge 1: Intra-modal dynamics

**Speaker’s behaviors**

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

**Sentiment Intensity**

- ++
- +

Intra-modal time

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Challenge 1: Intra-modal dynamics

Speaker’s behaviors

“This movie is great”

Smile  Head nod

Sentiment Intensity

++

++

Paul Pu Liang
Computational Modeling of Human Multimodal Language
Challenge 2: Cross-modal Dynamics

a) Multiple co-occurring interactions

Speaker’s behaviors

- “This movie is great”
- Smile
- Loud voice

Sentiment Intensity

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Challenge 2: Cross-modal Dynamics

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

Speaker’s behaviors:

- “This movie is **fair**”
- Smile
- Loud voice

Sentiment Intensity:

Over time, the combination of these behaviors indicates a positive sentiment.
Challenge 2: Cross-modal Dynamics

a) Multiple co-occurring interactions
b) Different weighted combinations
c) Multiple prediction targets

Speaker’s behaviors
- “This movie is great”
- Raised Eyebrows
- Loud voice

Emotions
- Happy
- Surprised

Paul Pu Liang
Computational Modeling of Human Multimodal Language
Multi-attention Recurrent Network (MARN)

1. Modeling intra-modal dynamics
   
   Set of Long-short Term Memories
Multi-attention Recurrent Network (MARN)

1. Modeling intra-modal dynamics
   - Set of Long-short Term Memories

2. Modeling cross-modal dynamics
   - Set of Long-short Term Hybrid Memories + Single-attention Block
Multi-attention Recurrent Network (MARN)

1. Modeling intra-modal dynamics
   - Set of Long-short Term Memories

2. Modeling cross-modal dynamics
   - Set of Long-short Term Hybrid Memories + Single-attention Block
   - Modeling multiple cross-modal dynamics
   - Set of Long-short Term Hybrid Memories + Multi-attention Block
Challenge 1: Intra-modal Dynamics

- **Language**
  - LSTM $l$
    - This
    - movie
    - is
    - great

- **Visual**
  - LSTM $v$
    - -
    - -
    - -
    - (smile)

- **Acoustic**
  - LSTM $a$
    - -
    - -
    - -
    - (loud voice)
Challenge 2: Cross-modal Dynamics

How do we capture cross-modal dynamics continuously across time?
Challenge 2: Single-attention Block

Captures cross-modal dynamics.

$LSTM_l$

$LSTM_v$

$LSTM_a$
Challenge 2: Single-attention Block

\[ \mathcal{A} h^2_2 \]

\[ \mathcal{A} h^2_3 \]

\[ \mathcal{A} h^2_4 \]

\[ \mathcal{A} h^2_5 \]

\[ a^1_t \]
Challenge 2: Single-attention Block

\[ h_t^l \]  \[ A \]  \[ h_t^v \]  \[ h_t^a \]  \[ a_t^1 \]  \[ \tilde{h}_t^1 \]  

LSTM \( l \)  \[ ]  LSTM \( l \)  
LSTM \( v \)  \[ ]  LSTM \( v \)  
LSTM \( a \)  \[ ]  LSTM \( a \)
Challenge 2: Single-attention Block
Challenge 2: Single-attention Block

\[
\begin{align*}
\mathcal{A} & \quad h^l_t \\
& \quad h^v_t \\
& \quad h^a_t
\end{align*}
\]

\[
\begin{align*}
\mathcal{A} & \quad a^1_t \\
& \quad \tilde{h}^1_t
\end{align*}
\]

\[
\begin{align*}
C_l & \quad s^l_t \\
C_v & \quad s^v_t \\
C_a & \quad s^a_t
\end{align*}
\]

\[
z_t
\]
Challenge 2: Single-attention Block

\[
\begin{align*}
\text{LSTM}^l_t & \quad \text{LSTM}^v_t & \quad \text{LSTM}^a_t \\
\begin{array}{c}
\cdots \\
\cdots \\
\cdots \\
\end{array}
\end{align*}
\]

\[
W^lx^l_{t+1} + U^lh^l_t + b^l
\]
Challenge 2: Long-short Term Hybrid Memory

LSTHM update

\[ W^l x_{t+1}^l + U^l h_t^l + V^l z_t^l + b^l \]
Challenge 2: Multi-attention Block
Multi-attention Recurrent Network (MARN)
Experiments

Language
- Glove word embeddings

Visual
- Facet features
  - FACS action units
  - Emotions

Acoustic
- COVAREP features
  - MFCCs
  - Pitch tracking

Sentiment
- Positive
- Negative

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

Personality
- Confidence
- Persuasion
- Passion

Alignment
- Word level
- P2FA

Facet features
- FACS action units
- Emotions

Language
- Visual
- Acoustic

Sentiment
- Positive
- Negative

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

Personality
- Confidence
- Persuasion
- Passion

Alignment
- Word level
- P2FA
State-of-the-art Results

CMU-MOSI Sentiment Analysis

Baseline Models

Multi-attention Recurrent Network (MARN)
State-of-the-art Results

Sentiment Analysis

Emotion Recognition

Personality Trait Prediction

- CMU-MOSI
- ICT-MMMO
- MOUD
- YouTube
- IEMOCAP
- POM Confidence
- POM Persuasion
- POM Passion
- POM Credibility

State-of-the-art Baseline

Multi-attention Recurrent Network (MARN)
Multi-attention Block is Important

Sentiment Analysis

Emotion Recognition

Personality Trait Prediction

No Multi-attention Block

Multi-attention Recurrent Network (MARN)
Multiple Attentions are Important

CMU-MOSI Sentiment Analysis

YouTube Sentiment Analysis
Visualization

Attentions *show diversity* and are sensitive to different cross-modal dynamics.
Some attentions *always inactive*

- Carry only intra-modal dynamics
- No cross-modal dynamics
Visualization

Attentions change behaviors across time, some changes are more drastic than others.
Visualization

Different attentions focus on different modalities.
Multi-attention Recurrent Network (MARN)

1. Modeling intra-modal dynamics
   - Set of Long-short Term Memories

2. Modeling cross-modal dynamics
   - Set of Long-short Term Hybrid Memories + Single-attention Block

   Modeling multiple cross-modal dynamics
   - Set of Long-short Term Hybrid Memories + Multi-attention Block
Direction 2: Unimodal, Bimodal and Trimodal
“This movie is sick”

“This movie is fair”

Smile

Loud voice

Sentiment Intensity

Ambiguous!

Unimodal cues

Ambiguous!
This movie is sick

Speaker’s behaviors

Unimodal

“This movie is sick”
Smile
Loud voice

“This movie is fair”

“This movie is sick”

“This movie is sick”

“This movie is sick”

Sentiment Intensity

?  →  Ambiguous!

+  →  Unimodal cues

+  →  Ambiguous!

?  →  Still Ambiguous!

Resolves ambiguity (bimodal interaction)

Different trimodal interactions!

Bimodal

“This movie is sick”
Smile

“This movie is sick”
Frown

“This movie is sick”

Loud voice

Trimodal

“This movie is sick”
Smile
Loud voice

“This movie is fair”
Smile
Loud voice

?  →  Ambiguous!

?  →  Still Ambiguous!

Resolves ambiguity (bimodal interaction)

Different trimodal interactions!
Simple Neural Network

Joint Multimodal Representation

Simply concatenates all three individual representations:

\[ h_m = f(W \cdot [h_x, h_y, h_z]) \]

➤ Similar to early fusion
Models both unimodal and bimodal interactions:

$$h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} = \begin{bmatrix} h_x \\ h_x \otimes h_y \end{bmatrix}$$

Important!
Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

$$h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} = \begin{bmatrix} h_x \\ h_x \otimes h_y \\ h_y \end{bmatrix}$$

Important!
Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

\[ h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} = \begin{bmatrix} h_x \\ h_x \otimes h_y \\ h_y \end{bmatrix} \]

Important!
Multimodal Tensor Fusion Network (TFN)

Can be extended to three modalities:

\[ h_m = [h_x] \otimes [h_y] \otimes [h_z] \]

Explicitly models **unimodal**, **bimodal** and **trimodal** interactions!
Number of Parameters

\[ O \left( d_y \times \sum_{m=1}^{M} d_m \right) \]
Number of Parameters

\[ d_m \cdot w = h \cdot d_y \]

\[ O \left( d_y \times \prod_{m=1}^{M} d_m \right) \]
Number of Parameters

\[ d_m \]

\[ d_m \]

\[ d_m \]

\[ O \left( d_y \times \prod_{m=1}^{M} d_m \right) \]

\[ 64 \times (32+32+32) = 6000 \]
Number of Parameters

\[
\begin{align*}
O & \left( d_y \times \prod_{m=1}^{M} d_m \right) = 64 \\
64 \times (32\times32\times32) & = 2000000
\end{align*}
\]
Low-rank Tensor Approximation

- Rank-$r$ approximation $O\left(d_y \times \prod_{m=1}^{M} d_m\right)$

$$w = \left\{ w_p^{(1)} d_m, w_p^{(2)} d_m \right\} + \ldots$$

$r$ summations
Low-rank Tensor Approximation

- Rank-r approximation \( O \left( d_y \times \prod_{m=1}^{M} d_m \right) \)

\[
\begin{align*}
d_m & \quad z_v \\
d_m & \quad z_l
\end{align*}
\]

\[
= \begin{cases} 
w_v^{(1)} & \quad d_{m} \\
+ w_{l}^{(1)} & \quad d_{m} \\
+ \ldots
\end{cases}
\]

\( r \ll d_m \text{ summations} \)
Low-rank Tensor Approximation

- Rank-\(r\) approximation \(O\left(d_y \times \prod_{m=1}^{M} d_m\right)\)

\[
\begin{align*}
  d_m & \quad z_v \\
  d_m & \quad z_t
\end{align*}
\]

\[
\begin{align*}
  \mathbf{Z} & \quad \mathbf{w} \\
  \mathbf{Z} & \quad \mathbf{w}_t^{(2)}
\end{align*}
\]

\[
\begin{align*}
  h & \quad d_y
\end{align*}
\]
Low-rank Tensor Approximation

- Rank-\( r \) approximation

\[
O \left( d_y \times \prod_{m=1}^{M} d_m \right)
\]

\[
O \left( d_y \times r \times \sum_{m=1}^{M} d_m \right)
\]
Low-rank Multimodal Fusion
# Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CMU-MOSI</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>Corr</td>
<td>Acc-2</td>
<td>F1</td>
<td>Acc-7</td>
<td>MAE</td>
<td>Corr</td>
<td>Acc</td>
<td>F1-Happy</td>
<td>F1-Sad</td>
</tr>
<tr>
<td>SVM</td>
<td>1.864</td>
<td>0.057</td>
<td>50.2</td>
<td>50.1</td>
<td>17.5</td>
<td>0.887</td>
<td>0.104</td>
<td>33.9</td>
<td>81.5</td>
<td>78.8</td>
</tr>
<tr>
<td>DF</td>
<td>1.143</td>
<td>0.518</td>
<td>72.3</td>
<td>72.1</td>
<td>26.8</td>
<td>0.869</td>
<td>0.144</td>
<td>34.1</td>
<td>81.0</td>
<td>81.2</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>1.079</td>
<td>0.581</td>
<td>73.9</td>
<td>73.9</td>
<td>28.7</td>
<td>0.840</td>
<td>0.278</td>
<td>34.8</td>
<td>81.7</td>
<td>81.7</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>1.019</td>
<td>0.601</td>
<td>73.9</td>
<td>74.0</td>
<td>33.2</td>
<td>0.891</td>
<td>0.270</td>
<td>34.6</td>
<td>81.3</td>
<td>74.0</td>
</tr>
<tr>
<td>MARN</td>
<td>0.968</td>
<td>0.625</td>
<td>77.1</td>
<td>77.0</td>
<td>34.7</td>
<td>-</td>
<td>-</td>
<td>39.4</td>
<td>83.6</td>
<td>81.2</td>
</tr>
<tr>
<td>MFN</td>
<td>0.965</td>
<td>0.632</td>
<td><strong>77.4</strong></td>
<td><strong>77.3</strong></td>
<td><strong>34.1</strong></td>
<td>0.805</td>
<td>0.349</td>
<td>41.7</td>
<td>84.0</td>
<td>82.1</td>
</tr>
<tr>
<td>TFN</td>
<td>0.970</td>
<td>0.633</td>
<td>73.9</td>
<td>73.4</td>
<td>32.1</td>
<td>0.886</td>
<td>0.093</td>
<td>31.6</td>
<td>83.6</td>
<td>82.8</td>
</tr>
<tr>
<td>LMF</td>
<td><strong>0.912</strong></td>
<td><strong>0.668</strong></td>
<td>76.4</td>
<td>75.7</td>
<td>32.8</td>
<td><strong>0.796</strong></td>
<td><strong>0.396</strong></td>
<td><strong>42.8</strong></td>
<td><strong>85.8</strong></td>
<td><strong>85.9</strong></td>
</tr>
</tbody>
</table>
Dynamic Fusion Graph
Dynamic Fusion Graph

bimodal

unimodal
Dynamic Fusion Graph

trimodal
bimodal
unimodal
Dynamic Fusion Graph

- **multimodal representation**
  - **trimodal**
  - **bimodal**
  - **unimodal**
Interpretable Fusion

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

(angry voice)
Direction 3: Direct and Relative
Person-independent Features

- Universal emotion expressions
Person-independent Features

- Universal emotion expressions
- Absolute emotions can be **directly** inferred from these observed behaviors
Person-dependent Features

- Emotions are also expressed with idiosyncratic behaviors
Person-dependent Features

- Emotions are also expressed with idiosyncratic behaviors
- Estimate relative changes by comparing behaviors
Multimodal Local Ranking

- Video centered at 2 random indices $j, k$
- $w$ window size

\[ t = j \ldots t = k \ldots \]
Multimodal Local Ranking

- Video centered at 2 random indices $k, l$
- $w$ window size
Multimodal Local Ranking

- **w** window size
- **m** local comparison pairs

\[ r_{j,k} = \mathbb{I}[y_j > y_k] \]
Global Ranking

- Bayesian ranking algorithm

\[ \hat{\epsilon}_j \geq \hat{\epsilon}_k \geq \hat{\epsilon}_l \]

\[ \hat{r}_{j,k} \geq \hat{r}_{k,l} \]

Global Ranking

observed

unobserved
Global Ranking

- Bayesian ranking algorithm

\[ r_{j,k} = \mathbb{I}[y_j > y_k] \]

\[ p(r_{j,k} = 1|e_j, e_k) = p(e_j > e_k) \]
Direct-Relative Fusion

\[
\hat{e}_j \quad \cdots \quad \hat{e}_k \quad \cdots \quad \hat{e}_l \quad \text{global emotion ranks}
\]

\[
\text{relative emotion label} \quad t = j \quad \text{emotion label} \quad t = k \quad \text{emotion label} \quad t = l
\]
Direct-Relative Fusion

multimodal data

\[ t = j \]

\[ t = k \]

\[ t = l \]

Direct-Relative Fusion

LSTM

emotion label
\[ t = j \]

direct

\( \hat{e}_j \)

relative

\( \hat{e}_k \)

\( \hat{e}_l \)

global emotion ranks
## Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AVEC16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arousal</td>
</tr>
<tr>
<td></td>
<td>CCC</td>
</tr>
<tr>
<td>EF-(/S/B/SB)LSTM [9, 11, 29]</td>
<td>0.4327</td>
</tr>
<tr>
<td>Gated-LSTM [38]</td>
<td>0.3210</td>
</tr>
<tr>
<td>MV-LSTM, view-specific [27]</td>
<td>0.4530</td>
</tr>
<tr>
<td>MV-LSTM, coupled [27]</td>
<td>0.4300</td>
</tr>
<tr>
<td>MV-LSTM, hybrid [27]</td>
<td>0.4729</td>
</tr>
<tr>
<td>MV-LSTM, fully connected [27]</td>
<td>0.4293</td>
</tr>
<tr>
<td>MLLR-500</td>
<td>0.4732</td>
</tr>
<tr>
<td>MLLR-1000</td>
<td><strong>0.5049</strong></td>
</tr>
<tr>
<td>Improvement over baselines</td>
<td><strong>↑ 0.032</strong></td>
</tr>
</tbody>
</table>
# Effect of Window Size

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AVEC16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>Arousal CCC</td>
</tr>
<tr>
<td>Metric</td>
<td></td>
</tr>
<tr>
<td>MLRF-500 w=10</td>
<td>0.4165</td>
</tr>
<tr>
<td>MLRF-500 w=50</td>
<td>0.4168</td>
</tr>
<tr>
<td>MLRF-500 w=100</td>
<td>0.4196</td>
</tr>
<tr>
<td>MLRF-500 w=200</td>
<td><strong>0.4732</strong></td>
</tr>
</tbody>
</table>
Effect of Direct and Relative Approaches

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AVEC16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arousal</td>
</tr>
<tr>
<td>Task</td>
<td>CCC</td>
</tr>
<tr>
<td>Metric</td>
<td></td>
</tr>
<tr>
<td>MLRF-500 direct predictions only</td>
<td>0.4327</td>
</tr>
<tr>
<td>MLRF-500 relative predictions only</td>
<td>0.3646</td>
</tr>
<tr>
<td>MLRF-500</td>
<td>0.4732</td>
</tr>
<tr>
<td>MLRF-1000 direct predictions only</td>
<td>0.4327</td>
</tr>
<tr>
<td>MLRF-1000 relative predictions only</td>
<td>0.4297</td>
</tr>
<tr>
<td>MLRF-1000</td>
<td>0.5049</td>
</tr>
</tbody>
</table>
Direction 4: Multimodal Representation Learning
Representation Learning

- Discriminative: $P(Y|X_1, \ldots, X_M)$
Representation Learning

- Discriminative: $P(Y|X_1, \ldots, X_M)$
- Generative: $P(X_1, \ldots, X_M)$
Representation Learning

- Discriminative: $P(Y|X_1, \ldots, X_M)$
- Generative: $P(X_1, \ldots, X_M)$
- Specificity: modality-specific and multimodal
Factorized Representations
Multimodal Factorization Model

Modality-specific generative factors
Multimodal Factorization Model

Modality-specific generative factors  Multimodal discriminative factor
Generative-Discriminative Objective

Modality-specific generative factors       Multimodal discriminative factor

\[
\sum_{i=1}^{M} c_X(x_i, F(G_{a_i}(z_{a_i}), G_Y(z_Y)))
\]
Generative-Discriminative Objective

Modality-specific generative factors  Multimodal discriminative factor

\[
\sum_{i=1}^{M} c_{X_i} \left( X_i, F\left( G_{a_i}(Z_{a_i}), G_{y}(Z_y) \right) \right) + c_{Y} \left( Y, D\left( G_{y}(Z_y) \right) \right)
\]
Generative-Discriminative Objective

Modality-specific generative factors

Multimodal discriminative factor

\[
\sum_{i=1}^{M} c_{X_i} \left( X_i, F(G_{a_i}(Z_{a_i}), G_y(Z_y)) \right) + c_Y \left( Y, D(G_y(Z_y)) \right) + \lambda \text{MMD}(Q_z, P_z),
\]
Unimodal Generation Results

za: style  
zy: label 0-9

MNIST

SVHN
Unimodal Generation Results

$\mathbf{z}_a$: style  $\mathbf{z}_y$: label 0-9

$\mathbf{f}(x_{z\alpha})$
Unimodal Generation Results

za: style  
zy: label 0-9

fix \( z_a \)

fix \( z_y \) or \( y \)

Paul Pu Liang
Computational Modeling of Human Multimodal Language
Multimodal Generative Results

Neural Networks

$Z_{a1}$ → $F_{a1}$ → 9

$Z_{a2}$ → $F_{a2}$ → 9

$Z_{y}$ → $F_{y}$ → nine (label)
Multimodal Generative Results

- $Z_{a1}$: SVHN style
- $Z_{a2}$: MNIST style
- $Z_y$: label
- $F_{a1}$
- $F_{a2}$
- $F_y$

The output is labeled "nine (label)".
Multimodal Generative Results

za1: SVHN style  
zy: label  
za2: MNIST style

Modality 1 (SVHN)
Multimodal Generative Results

\[ \begin{align*}
Z_{a1} & \rightarrow F_{a1} \\
Z_{a2} & \rightarrow F_{a2} \\
Z_y & \rightarrow F_y
\end{align*} \]

\text{nine (label)}

Modality 1 (SVHN)

Modality 2 (MNIST)

za1: SVHN style  \hspace{1cm} zy: label  \hspace{1cm} za2: MNIST style

Fix \[ Z_{a2} \]
Multimodal Generative Results

**Za1**: SVHN style  
**Zy**: label  
**Za2**: MNIST style

- **Modality 1 (SVHN)**: $\text{Fix } Z_y$
- **Modality 2 (MNIST)**

Paul Pu Liang  
Computational Modeling of Human Multimodal Language
Multimodal Discriminative Results

Table 2: Results for ablation studies on CMU-MOSI. Best results in bold. All the components in MFM are necessary for best performance.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>M_E</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>–</td>
<td>76.1</td>
<td>76.0</td>
</tr>
<tr>
<td>M_D</td>
<td>no</td>
<td>no</td>
<td>–</td>
<td>–</td>
<td>74.6</td>
<td>74.7</td>
</tr>
<tr>
<td>M_C</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>76.5</td>
<td>76.5</td>
</tr>
<tr>
<td>M_B</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>–</td>
<td>74.9</td>
<td>75.0</td>
</tr>
<tr>
<td>M_A</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>75.1</td>
<td>75.1</td>
</tr>
<tr>
<td>MFM</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td><strong>77.3</strong></td>
<td><strong>77.2</strong></td>
</tr>
</tbody>
</table>
Direction 5: Robust Multimodal Representation Learning
Learning Joint Representations: 2 modalities

<table>
<thead>
<tr>
<th>Traditional Methods</th>
<th>Visual Modality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language Modality</strong></td>
<td><strong>Visual Modality</strong></td>
</tr>
<tr>
<td><em>Today was a great day!</em></td>
<td>![Images of a person speaking]</td>
</tr>
</tbody>
</table>

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Learning Joint Representations: 2 modalities

Traditional Methods

Language Modality

Today was a great day!

Joint Representation

Visual Modality

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Learning Joint Representations: 2 modalities

Traditional Methods

Language Modality

*Today was a great day!*

Joint Representation

Visual Modality

Sentiment Prediction
Learning Joint Representations: 2 modalities

Both modalities required at test time!
Sensitive to missing/noisy visual modality.
Learning Robust Joint Representations: 2 modalities

Bimodal Cyclic Translations

Language Modality

Today was a great day!

Joint Representation

forward

forward

Visual Modality

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Learning Robust Joint Representations: 2 modalities

Bimodal Cyclic Translations

**Language Modality**

*Today was a great day!*

**Joint Representation**

forward

backward

**Visual Modality**

forward

backward
Learning Robust Joint Representations: 2 modalities

Bimodal Cyclic Translations

Language Modality

*Today was a great day!*

Joint Representation

forward
backward
forward
backward

Visual Modality

Sentiment Prediction
Learning Robust Joint Representations: 2 modalities

Bimodal Cyclic Translations

Language Modality

\textit{Today was a great day!}

Joint Representation

\textbf{Language Modality}

\hspace{1cm} forward

\hspace{1cm} backward

\textbf{Visual Modality}

\hspace{1cm} forward

\hspace{1cm} backward

Sentiment Prediction

Only language modality required at test time!
Learning Robust Joint Representations: 3 modalities

Trimodal Cyclic Translations

**Language Modality**

*Today was a great day!*

**Visual Modality**

**Acoustic Modality**
Learning Robust Joint Representations: 3 modalities

Trimodal Cyclic Translations

Language Modality

*Today was a great day!*

Visual Modality

Acoustic Modality
Learning Robust Joint Representations: 3 modalities

Trimodal Cyclic Translations

Language Modality

Today was a great day!

Visual Modality

forward

backward

forward

backward

Acoustic Modality

Joint Representation
Learning Robust Joint Representations: 3 modalities

Trimodal Cyclic Translations

Language Modality

*Today was a great day!*

Visual Modality

Forward

Forward

Acoustic Modality

Sentiment Prediction

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Learning Robust Joint Representations: 3 modalities

Trimodal Cyclic Translations

Language Modality

Today was a great day!

Joint Representation

Visual Modality

Acoustic Modality

Sentiment Prediction

Only language modality required at test time!
Multimodal Cyclic Translation Network

\[
\begin{align*}
\text{Source } & \quad X^S \\
\text{Target } & \quad X^T
\end{align*}
\]
Multimodal Cyclic Translation Network

Source $X^S$

Encoder RNN $f_{\theta_e}$

Target $X^T$

Forward Translation

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Multimodal Cyclic Translation Network

\[ \text{Target } X^T \]

\[ \text{Embedded Representation } \mathcal{E}_{S \leftrightarrow T} \]

\[ \text{Encoder RNN } f_{\theta_e} \]

\[ \text{Source } X^S \]

1 Forward Translation

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Multimodal Cyclic Translation Network

Source $X^S$

Encoder RNN $f_{\theta_e}$

Embedded Representation $\mathcal{E}_{S\leftrightarrow T}$

Decoder RNN $f_{\theta_d}$

Target $X^T$

1. Forward Translation
2. Forward Translation

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Multimodal Cyclic Translation Network

Source $X^S$

Encoder RNN $f_{\theta_e}$

Embedded Representation $E_{S \leftrightarrow T}$

Decoder RNN $f_{\theta_d}$

Target $X^T$

Forward Translation

Seq2Seq

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Multimodal Cyclic Translation Network

Source $X^S$

Backward Translation 3

Target $X^T$

Forward Translation 2

Seq2Seq

Encoder RNN $f_{\theta_e}$

Embedded Representation $E_{S\leftrightarrow T}$

Decoder RNN $f_{\theta_d}$

Forward Translation 1

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Multimodal Cyclic Translation Network

```
<table>
<thead>
<tr>
<th></th>
<th>Backward Translation</th>
<th>Forward Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Source</td>
<td>X^S</td>
</tr>
<tr>
<td>2</td>
<td>Target</td>
<td>X^T</td>
</tr>
<tr>
<td>3</td>
<td>Backward Translation</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Forward Translation</td>
<td></td>
</tr>
</tbody>
</table>
```

**Seq2Seq**

- Encoder RNN: \( f_{\theta_e} \)
- Embedded Representation: \( \mathcal{E}_{S \leftarrow T} \)
- Decoder RNN: \( f_{\theta_d} \)
Multimodal Cyclic Translation Network

- **Seq2Seq**
  - Encoder RNN $f_{\theta_e}$
  - Embedded Representation $\mathcal{E}_{S\leftrightarrow T}$
  - Decoder RNN $f_{\theta_d}$

- Forward Translation
  - Source $X^S$

- Backward Translation
  - Target $X^T$

- Prediction RNN $g_w$
- Sentiment Prediction

Paul Pu Liang
Computational Modeling of Human Multimodal Language
Multimodal Cyclic Translation Network

Seq2Seq

$\text{Encoder RNN } f_{\theta_e}$

$\text{Embedded Representation } \mathcal{E}_{S \leftrightarrow T}$

$\text{Decoder RNN } f_{\theta_d}$

Target $X^T$

Backward Translation

Forward Translation

Source $X^S$

Backward Translation

Forward Translation

Sentiment Prediction

Prediction RNN $g_w$
Coupled Translation-Prediction Objective

Forward translation loss $\mathcal{L}_t = \mathbb{E}[\ell_X(X^T, \hat{X}^T)]$
Coupled Translation-Prediction Objective

- **Forward translation loss** \( L_t = \mathbb{E}[\ell_{X^T} (\hat{X}^T, X^T)] \)
- **Cycle consistent loss** \( L_c = \mathbb{E}[\ell_{X^S} (\hat{X}^S, X^S)] \)

**Diagram:**

- Source: \( X^S \)
- Encoder RNN: \( f_{\theta_e} \)
- Embedded Representation: \( E_{S \rightarrow T} \)
- Decoder RNN: \( f_{\theta_d} \)
- Predict RNN: \( g_w \)
- Target: \( X^T \)
- Sentiment Prediction: \( y \)

- Backward Translation: 4
- Forward Translation: 1
- Forward Translation: 2
- Backward Translation: 3
- Forward Translation: 5
Coupled Translation-Prediction Objective

\[ \mathcal{L} = \lambda_t \mathcal{L}_t + \lambda_c \mathcal{L}_c + \mathcal{L}_p \]

- Forward translation loss
  \[ \mathcal{L}_t = \mathbb{E}[\ell_X(\hat{X}^T, X^T)] \]
- Cycle consistent loss
  \[ \mathcal{L}_c = \mathbb{E}[\ell_X(\hat{X}^S, X^S)] \]
- Prediction loss
  \[ \mathcal{L}_p = \mathbb{E}[\ell_Y(\hat{y}, y)] \]
Hierarchical Multimodal Cyclic Translation Network

Target2 $X^{T_2}$

Target1

Source $X^S$
Hierarchical Multimodal Cyclic Translation Network
Hierarchical Multimodal Cyclic Translation Network
Hierarchical Multimodal Cyclic Translation Network

Source $X^S$

1. Forward Translation
2. Forward Translation
3. Backward Translation
4. Backward Translation

Encoder RNN $f^{1}_{\theta_e}$

Embedded Representation $\mathcal{E}_{S \leftrightarrow T_1}$

Decoder RNN $f^{1}_{\theta_d}$

Target1 $X^{T_1}$

Target2 $X^{T_2}$

Embedded Representation $\mathcal{E}_{(S \leftrightarrow T_1) \rightarrow T_2}$

Encoder RNN $f^{2}_{\theta_e}$

Seq2Seq 1

Seq2Seq 2

Forward Translation

Backward Translation
Hierarchical Multimodal Cyclic Translation Network
Hierarchical Multimodal Cyclic Translation Network

1. Source $X^S$
2. Forward Translation
3. Backward Translation
4. Backward Translation
5. Forward Translation
6. Forward Translation
7. Prediction RNN $g_{w}$

Decoder RNN $f^{1}_{\theta_{d}}$
Encoder RNN $f^{1}_{\theta_{e}}$

Embedded Representation $\mathcal{E}_{S \leftrightarrow T_1}$

Seq2Seq 1

Seq2Seq 2

Sentiment Prediction

Sentiment $Y$
State-of-the-art Results: CMU-MOSI

<table>
<thead>
<tr>
<th>Dataset Model</th>
<th>Test Inputs</th>
<th>Acc(↑)</th>
<th>FI(↑)</th>
<th>MAE(↓)</th>
<th>Corr(↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>{ℓ, v, a}</td>
<td>56.4</td>
<td>56.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SVM</td>
<td>{ℓ, v, a}</td>
<td>71.6</td>
<td>72.3</td>
<td>1.100</td>
<td>0.559</td>
</tr>
<tr>
<td>THMM</td>
<td>{ℓ, v, a}</td>
<td>50.7</td>
<td>45.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EF-HCRF</td>
<td>{ℓ, v, a}</td>
<td>65.3</td>
<td>65.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MV-HCRF</td>
<td>{ℓ, v, a}</td>
<td>65.6</td>
<td>65.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DF</td>
<td>{ℓ, v, a}</td>
<td>74.2</td>
<td>74.2</td>
<td>1.143</td>
<td>0.518</td>
</tr>
<tr>
<td>EF-LSTM</td>
<td>{ℓ, v, a}</td>
<td>74.3</td>
<td>74.3</td>
<td>1.023</td>
<td>0.622</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>{ℓ, v, a}</td>
<td>73.9</td>
<td>74.0</td>
<td>1.019</td>
<td>0.601</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>{ℓ, v, a}</td>
<td>75.2</td>
<td>75.3</td>
<td>1.079</td>
<td>0.614</td>
</tr>
<tr>
<td>TFN</td>
<td>{ℓ, v, a}</td>
<td>74.6</td>
<td>74.5</td>
<td>1.040</td>
<td>0.587</td>
</tr>
<tr>
<td>GME-LSTM(A)</td>
<td>{ℓ, v, a}</td>
<td>76.5</td>
<td>73.4</td>
<td>0.955</td>
<td>-</td>
</tr>
<tr>
<td>MARN</td>
<td>{ℓ, v, a}</td>
<td>77.1</td>
<td>77.0</td>
<td>0.968</td>
<td>0.625</td>
</tr>
<tr>
<td>MFN</td>
<td>{ℓ, v, a}</td>
<td>77.4</td>
<td>77.3</td>
<td>0.965</td>
<td>0.632</td>
</tr>
<tr>
<td>LMF</td>
<td>{ℓ, v, a}</td>
<td>76.4</td>
<td>75.7</td>
<td>0.912</td>
<td>0.668</td>
</tr>
<tr>
<td>RMFN</td>
<td>{ℓ, v, a}</td>
<td>78.4</td>
<td>78.0</td>
<td>0.922</td>
<td>0.681</td>
</tr>
<tr>
<td>MCTN</td>
<td>{ℓ}</td>
<td>79.3</td>
<td>79.1</td>
<td>0.909</td>
<td>0.676</td>
</tr>
</tbody>
</table>
State-of-the-art Results: ICT-MMMO and YouTube

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test Inputs</th>
<th>ICT-MMMO</th>
<th></th>
<th>YouTube</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc(†)</td>
<td>F1(†)</td>
<td>Acc(†)</td>
<td>F1(†)</td>
</tr>
<tr>
<td>RF</td>
<td>{ℓ, v, a}</td>
<td>70.0</td>
<td>69.8</td>
<td>33.3</td>
<td>32.3</td>
</tr>
<tr>
<td>SVM</td>
<td>{ℓ, v, a}</td>
<td>68.8</td>
<td>68.7</td>
<td>42.4</td>
<td>37.9</td>
</tr>
<tr>
<td>THMM</td>
<td>{ℓ, v, a}</td>
<td>53.8</td>
<td>53.0</td>
<td>42.4</td>
<td>27.9</td>
</tr>
<tr>
<td>EF-HCRF</td>
<td>{ℓ, v, a}</td>
<td>73.8</td>
<td>73.1</td>
<td>45.8</td>
<td>45.0</td>
</tr>
<tr>
<td>MV-HCRF</td>
<td>{ℓ, v, a}</td>
<td>68.8</td>
<td>67.1</td>
<td>44.1</td>
<td>44.0</td>
</tr>
<tr>
<td>DF</td>
<td>{ℓ, v, a}</td>
<td>65.0</td>
<td>58.7</td>
<td>45.8</td>
<td>32.0</td>
</tr>
<tr>
<td>EF-LSTM</td>
<td>{ℓ, v, a}</td>
<td>72.5</td>
<td>70.9</td>
<td>44.1</td>
<td>43.6</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>{ℓ, v, a}</td>
<td>72.5</td>
<td>72.3</td>
<td>45.8</td>
<td>43.3</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>{ℓ, v, a}</td>
<td>70.0</td>
<td>70.1</td>
<td>45.0</td>
<td>45.1</td>
</tr>
<tr>
<td>TFN</td>
<td>{ℓ, v, a}</td>
<td>72.5</td>
<td>72.6</td>
<td>45.0</td>
<td>41.0</td>
</tr>
<tr>
<td>MARN</td>
<td>{ℓ, v, a}</td>
<td>71.3</td>
<td>70.2</td>
<td>48.3</td>
<td>44.9</td>
</tr>
<tr>
<td>MFN</td>
<td>{ℓ, v, a}</td>
<td>73.8</td>
<td>73.1</td>
<td>51.7</td>
<td>51.6</td>
</tr>
<tr>
<td>MCTN</td>
<td>{ℓ}</td>
<td>81.3</td>
<td>80.8</td>
<td>51.7</td>
<td>52.4</td>
</tr>
</tbody>
</table>
Bimodal Variations

MCTN Bi

Simple Bi

No-Cycle Bi

Double Bi

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Bimodal Variations

MCTN Bi

Simple Bi

No-Cycle Bi

Double Bi

Step 1

Step 2

+ Cyclic translations
+ Parameter sharing

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Trimodal Variations

MCTN Tri

Simple Tri

Double Tri

Concat Tri

Paired Tri

\[ \dot{X}^{T_2} \]

Enc

\[ \mathcal{E}(S \rightarrow T_1) \rightarrow T_2 \]

Dec

\[ \mathcal{E}(S \rightarrow T_1) \rightarrow T_2 \]

Enc

\[ \mathcal{E}(S \rightarrow T_1) \rightarrow T_2 \]

Dec

Double Tri

\[ \mathcal{E}(S \rightarrow T_1, T_1 \rightarrow S) \rightarrow T_2 \]

Enc

\[ \mathcal{E}(S \rightarrow T) \rightarrow T_2 \]

Enc

\[ \mathcal{E}(S \rightarrow T) \rightarrow T_2 \]

Enc

Concat Tri

\[ \mathcal{E}(X^S, X^{T_1}) \rightarrow T_3 \]

Dec

\[ \mathcal{E}(X^S, X^{T_1}) \rightarrow T_3 \]

Dec

Paired Tri

\[ X^{T_1} \]

Dec1

\[ \mathcal{E}(S \rightarrow T_1) \]

Enc

\[ X^{T_2} \]

Dec2

\[ \mathcal{E}(S \rightarrow T_2) \]

Enc

Paul Pu Liang

Computational Modeling of Human Multimodal Language
Trimodal Variations

- MCTN Tri
- Simple Tri
- Double Tri
- Concat Tri
- Paired Tri

+ Cyclic translations
+ Parameter sharing
+ Hierarchical structure
## Adding More Modalities

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Translation</th>
<th>CMU-MOSI</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Translation</td>
<td>Acc</td>
<td>F1</td>
<td>MAE</td>
<td>Corr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>V ⇋ A</td>
<td>53.1</td>
<td>53.2</td>
<td>1.420</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>MCTN Bi (Fig. 4a)</td>
<td>T ⇋ A</td>
<td>76.4</td>
<td>76.4</td>
<td>0.977</td>
<td>0.636</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T ⇋ V</td>
<td>76.8</td>
<td>76.8</td>
<td>1.034</td>
<td>0.592</td>
</tr>
<tr>
<td></td>
<td>(V ⇋ A) → T</td>
<td></td>
<td>56.4</td>
<td>56.3</td>
<td>1.455</td>
<td>0.151</td>
</tr>
<tr>
<td></td>
<td>MCTN Tri (Fig. 4e)</td>
<td>(T ⇋ A) → V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>78.7</td>
<td>78.8</td>
<td>0.960</td>
<td>0.650</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(T ⇋ V) → A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>79.3</td>
<td>79.1</td>
<td>0.909</td>
<td>0.676</td>
</tr>
</tbody>
</table>
Adding More Modalities

Bimodal MCTN  
without cyclic translation

Bimodal MCTN  
with cyclic translation

Trimodal MCTN  
with cyclic translation

Paul Pu Liang  
Computational Modeling of Human Multimodal Language
New Multimodal Dataset: MOSEI
New Dataset: 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
- Uninformative
- Contradictory smile

Acoustic:
- (frustrated voice)
- (angry voice)
- (disappointed voice)
MOSEI Dataset

1,000 speakers

250 topics
Annotation Distributions

[Bar charts showing distribution of annotations for different sentiments and emotions]
Future Directions

- Learning from limited/missing multimodal data

That’s a zebra!

That’s a horse!
Future Directions

- Learning from limited/missing multimodal data

That’s a horse!
Future Directions

- Learning from limited/missing multimodal data

A zebra is a horse with stripes.
Future Directions

- Learning from limited/missing multimodal data

“A zebra is a horse with stripes.”

“stripes”
Future Directions

- Learning from limited/missing multimodal data
Future Directions

- Learning from unstructured, semi-supervised multimodal data
Future Directions

- Learning from unstructured, semi-supervised multimodal data

"horse family"

"donkey"

image
Future Directions

- Multimodal generation, style transfer, video prediction
Computational Modeling of Multimodal Language

1. 5 Directions
   - Intra-modal and Cross-modal
   - Unimodal, Bimodal and Trimodal
   - Direct and Relative
   - Multimodal Representation Learning
   - Robust Multimodal Representation Learning

2. MOSEI Dataset
   - Diversity in samples, topics, speakers and annotations
The End!

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