Strong and Simple Baselines for Multimodal Utterance Embeddings

Paul Pu Liang*, Yao Chong Lim*, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov and Louis-Philippe Morency
Human Language is often multimodal

**Language**
- Word choice
- Syntax
- Pragmatics

**Acoustic**
- Tone
- Prosody
- Phrasing

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

**Sentiment**
- Positive/Negative
- Intensity

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

**Meaning**
- Sarcasm
- Humor
Human Language is often multimodal

“There movie is great” + Neutral expression

Sentiment Intensity
Human Language is often multimodal

"This movie is great" + Neutral expression

"This movie is great" + Smile

Sentiment Intensity

+ + + +
Challenges in Multimodal ML
Challenges in Multimodal ML

1. Intramodal interactions

\[
\text{Smile} + \text{Head nod} \quad \text{vs.} \quad \text{Smile} + \text{Head shake}
\]
Challenges in Multimodal ML

1. Intramodal interactions

Smile + Head nod vs. Smile + Head shake

2. Crossmodal interactions

Bimodal “This movie is great” + Smile
Challenges in Multimodal ML

1. Intramodal interactions

```
Smile + Head nod vs. Smile + Head shake
```

2. Crossmodal interactions

```
Bimodal: “This movie is great” + Smile

Trimodal: “This movie is GREAT” + Smile + “great” is emphasized, drawn-out
```

(Sarcasm)
Multimodal Language Embedding

“This is unbelievable!”

Intramodal + crossmodal interactions

Downstream Tasks
- Sentiment Analysis
- Emotion Recognition
- Speaker Trait Recognition
...

Utterance Embedding
Multimodal Language Embedding

“This is unbelievable!”

Intramodal + crossmodal interactions

Downstream Tasks
- Sentiment Analysis
- Emotion Recognition
- Speaker Trait Recognition

...
Why fast models?

• Applications
• Robots, virtual agents, intelligent personal assistants
• Processing large amounts of multimedia data
Research Question

Can we make principled but simple models for multimodal utterance embeddings that perform competitively?
Research Question

Can we make principled but simple models for multimodal utterance embeddings that perform competitively?

![Performance and Speed Diagram]

- **Current SOTA**
- **Our goal**
Research Question

Can we make principled but simple models for multimodal utterance embeddings that perform competitively?

Our models:
- Fewer parameters
- Has a closed-form solution
- Linear functions
- Competitive with SOTA!
A language-only solution

Arora et al. (2016, 2017):

Sentence embedding $m_s$

Word embeddings $w_1$, $w_2$, $w_3$, $w_4$

This manual is helpful
A language-only solution

Arora et al. (2016, 2017):

\[ p(w_i|m_s) \propto \exp(w_i \cdot m_s) \]

This manual is helpful
A language-only solution

Arora et al. (2016, 2017):

\[ p(w_i|m_s) \propto \exp(w_i \cdot m_s) \]

Fast: No learnable parameters.
MMB1: Representing intramodal interactions
MMB1: Representing intramodal interactions

(Arora et al)

It doesn’t give help

Utterance embedding $m_s$
MMB1: Representing intramodal interactions

Utterance embedding $m_s$

Utterance-level feature distributions:
- Visual
- Audio

(Arora et al)

Gaussian parameters
- Visual
  - $\mu_v$
  - $\sigma_v$
  - $v_1$, $v_2$, $v_3$, …, $v_n$

Gaussian parameters
- Audio
  - $\mu_a$
  - $\sigma_a$
  - $a_1$, $a_2$, $a_3$, …, $a_n$
MMB1: Representing intramodal interactions

Utterance embedding $m_s$

$\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, ..., $\mathbf{w}_n$

It, doesn’t, give, help

Arora et al.

Linear transformations

$\mu_v$, $\sigma_v$, Visual

$\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, ..., $\mathbf{v}_n$

$\mu_a$, $\sigma_a$, Audio

$\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, ..., $\mathbf{a}_n$
MMB1: Representing intramodal interactions

It doesn’t give help

Small number of additional parameters!
Crossmodal interactions

“\textit{It didn’t help}” + Neutral face

“\textit{It didn’t help}” + Sad face

Stable voice

Shaky voice

Disappointment

Sadness
MMB2: Incorporating crossmodal interactions

Unimodal

Utterance embedding $m_S$

W+A

[\begin{array}{c}
  w_1, a_1 \\
  \vdots \\
  w_n, a_n \\
\end{array}]

V+A

[\begin{array}{c}
  v_1, a_1 \\
  \vdots \\
  v_n, a_n \\
\end{array}]

W+V

[\begin{array}{c}
  w_1, v_1 \\
  \vdots \\
  w_n, v_n \\
\end{array}]

W+V+A

[\begin{array}{c}
  w_1, v_1, a_1 \\
  \vdots \\
  w_n, v_n, a_n \\
\end{array}]

Concatenated inputs
MMB2: Incorporating crossmodal interactions

Unimodal

Utterance embedding $m_s$

W+A

$\mu_{wa}$  $\sigma_{wa}$

$[w_1, a_1]$  ...  $[w_n, a_n]$  $[v_1, a_1]$  ...  $[v_n, a_n]$

V+A

$\mu_{va}$  $\sigma_{va}$

$[v_1, a_1]$  ...  $[v_n, a_n]$

W+V

$\mu_{wv}$  $\sigma_{wv}$

$[w_1, v_1]$  ...  $[w_n, v_n]$  $[w_1, v_1, a_1]$  ...  $[w_n, v_n, a_n]$

W+V+A

$\mu_{wva}$  $\sigma_{wva}$

$[w_1, v_1, a_1]$  ...  $[w_n, v_n, a_n]$
MMB2: Incorporating crossmodal interactions

\[
\begin{align*}
W+A: & \quad \mu_{wa}, \sigma_{wa}, [w_1, a_1], \ldots, [w_n, a_n] \\
V+A: & \quad \mu_{va}, \sigma_{va}, [v_1, a_1], \ldots, [v_n, a_n] \\
W+V: & \quad \mu_{wv}, \sigma_{wv}, [w_1, v_1], \ldots, [w_n, v_n] \\
W+V+A: & \quad \mu_{wva}, \sigma_{wva}, [w_1, v_1, a_1], \ldots, [w_n, v_n, a_n]
\end{align*}
\]
How do we optimize the model?

Coordinate ascent-style
How do we optimize the model?

Two steps each iteration:

- Visual
  - \( \mu \)
  - \( \sigma \)
  - \( v \)
  - \( v' \)
  - \( v_1 \), \( v_2 \), \( v_3 \), \ldots, \( v_n \)

- Audio
  - \( \mu \)
  - \( \sigma \)
  - \( a \)
  - \( a' \)
  - \( a_1 \), \( a_2 \), \( a_3 \), \ldots, \( a_n \)

- Words
  - \( w \)
  - \( w_1 \), \( w_2 \), \( w_3 \), \ldots, \( w_n \)

- Utterance embedding \( m_s \)

Coordinate ascent-style
How do we optimize the model?

Two steps each iteration:
1. Fix transformation parameters, solve for $m_s$

Coordinate ascent-style
How do we optimize the model?

Two steps each iteration:
1. Fix transformation parameters, solve for $m_s$
2. Fix $m_s$, update transformation parameters by gradient descent
Datasets

CMU-MOSI (Zadeh et al. 2016)
• Multimodal Sentiment Analysis dataset
• 2199 English opinion segments (monologues) from online videos

Language: I thought it was fun

Visual

Acoustic (elongation) (emphasis)
Datasets

POM (Park et al., 2014)
- Multimodal Speaker Traits Recognition
- 903 English videos annotated for speaker traits such as confidence, dominance, vividness, relaxed, nervousness, humor etc.
Compared Models

Deep neural models
• Early Fusion: EF-LSTM
• DF (Nojavanasghari et al., 2016)
• Multi-view Learning: MV-LSTM (Rajagopalan et al., 2016)
• Contextual LSTM: BC-LSTM (Poria et al., 2017)
• Tensor Fusion: TFN (Zadeh et al., 2017)
• Memory Fusion: MFN (Zadeh et al., 2018)
Experiments

CMU-MOSI Sentiment

<table>
<thead>
<tr>
<th>Model</th>
<th>Binary Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF-LSTM</td>
<td>74.6</td>
</tr>
<tr>
<td>DF</td>
<td>73.2</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>73.4</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>74.6</td>
</tr>
<tr>
<td>TFN</td>
<td>77.4</td>
</tr>
<tr>
<td>MFN</td>
<td>77.4</td>
</tr>
<tr>
<td>MMB1</td>
<td>75.1</td>
</tr>
<tr>
<td>MMB2</td>
<td>75.1</td>
</tr>
</tbody>
</table>

Deep neural models

Our baselines

Legend:
Experiments

POM Speaker Traits Recognition

MAE

0.71  0.72  0.73  0.74  0.75  0.76  0.77  0.78  0.79  0.8  0.81  0.82

EF-LSTM  DF  MV-LSTM  BC-LSTM  TFN  MFN  MMB1  MMB2

Deep neural models
Our baselines

Our baselines

0.774  0.746  0.785  0.746  0.774
Speed Comparisons

Average Inference Time (s)

- Deep neural models
- Our baselines
Conclusion

- Proposed two simple but strong baselines for learning embeddings of multimodal utterances
- Try strong baselines before working on complicated models!
The End!

Email:
pliang@cs.cmu.edu
yaochonl@cs.cmu.edu

Github: yaochie/multimodal-baselines

![Graph showing CMU-MOSI Accuracy (%) vs. Inferences per second for Deep neural models and Our baselines.](image)
Additional Results
<table>
<thead>
<tr>
<th>Dataset</th>
<th>CMU-MOSI Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>A (2)</td>
</tr>
<tr>
<td>Metric</td>
<td></td>
</tr>
<tr>
<td>Majority</td>
<td>50.2</td>
</tr>
<tr>
<td>RF</td>
<td>56.4</td>
</tr>
<tr>
<td>THMM</td>
<td>50.7</td>
</tr>
<tr>
<td>EF-HCRF(*)</td>
<td>65.3</td>
</tr>
<tr>
<td>MV-HCRF(*)</td>
<td>65.6</td>
</tr>
<tr>
<td>SVM-MD</td>
<td>71.6</td>
</tr>
<tr>
<td>C-MKL</td>
<td>72.3</td>
</tr>
<tr>
<td>DF</td>
<td>72.3</td>
</tr>
<tr>
<td>SAL-CNN</td>
<td>73.0</td>
</tr>
<tr>
<td>EF-LSTM(*)</td>
<td>74.3</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>73.9</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>73.9</td>
</tr>
<tr>
<td>TFN</td>
<td>74.6</td>
</tr>
<tr>
<td>MFN</td>
<td><strong>77.4</strong></td>
</tr>
<tr>
<td>MMB1</td>
<td>73.6</td>
</tr>
<tr>
<td>MMB2</td>
<td><strong>75.2</strong></td>
</tr>
<tr>
<td>Dataset</td>
<td>Task</td>
</tr>
<tr>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>Majority</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
</tr>
<tr>
<td>DF</td>
<td></td>
</tr>
<tr>
<td>EF-LSTM(*)</td>
<td></td>
</tr>
<tr>
<td>MV-LSTM</td>
<td></td>
</tr>
<tr>
<td>BC-LSTM</td>
<td></td>
</tr>
<tr>
<td>TFN</td>
<td></td>
</tr>
<tr>
<td>MFN</td>
<td></td>
</tr>
<tr>
<td>MMB2</td>
<td></td>
</tr>
<tr>
<td>Dataset</td>
<td>Con</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
</tr>
<tr>
<td>Majority</td>
<td>-0.041</td>
</tr>
<tr>
<td>SVM</td>
<td>0.063</td>
</tr>
<tr>
<td>DF</td>
<td>0.240</td>
</tr>
<tr>
<td>EF-LSTM(*)</td>
<td>0.221</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>0.358</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td><strong>0.359</strong></td>
</tr>
<tr>
<td>TFN</td>
<td>0.089</td>
</tr>
<tr>
<td>MFN</td>
<td><strong>0.395</strong></td>
</tr>
<tr>
<td>MMB2</td>
<td>0.350</td>
</tr>
</tbody>
</table>
Experiments

CMU-MOSI Sentiment

<table>
<thead>
<tr>
<th>Model</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF-LSTM</td>
<td>0.62</td>
</tr>
<tr>
<td>DF</td>
<td>0.50</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>0.60</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>0.55</td>
</tr>
<tr>
<td>TFN</td>
<td>0.65</td>
</tr>
<tr>
<td>MFN</td>
<td>0.63</td>
</tr>
<tr>
<td>MMB1</td>
<td>0.55</td>
</tr>
<tr>
<td>MMB2</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Legend:
- Blue: Deep neural models
- Red: Our baselines
Experiments

CMU-MOSI Sentiment

F1 score

- EF-LSTM
- DF
- MV-LSTM
- BC-LSTM
- TFN
- MFN
- MMB1
- MMB2

Deep neural models
Our baselines
Experiments

CMU-MOSI Sentiment

7-class Accuracy (%)

- EF-LSTM
- DF
- MV-LSTM
- BC-LSTM
- TFN
- MFN
- MMB1
- MMB2

Deep neural models

Our baselines
Experiments

CMU-MOSI Sentiment

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF-LSTM</td>
<td>0.85</td>
</tr>
<tr>
<td>DF</td>
<td>0.9</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>0.95</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>1</td>
</tr>
<tr>
<td>TFN</td>
<td>1.05</td>
</tr>
<tr>
<td>MFN</td>
<td>1.1</td>
</tr>
<tr>
<td>MMB1</td>
<td>1.15</td>
</tr>
<tr>
<td>MMB2</td>
<td>1.2</td>
</tr>
</tbody>
</table>

- Deep neural models
- Our baselines