Efficient Low-rank Multimodal Fusion
With Modality-specific Factors

Zhun Liu, Ying Shen,
Varun Bharadwaj, Paul Pu Liang,
Amir Zadeh, Louis-Philippe Morency
Artificial Intelligence
Sentiment and Emotion Analysis

Speaker’s behaviors

"This movie is sick"

Smile

Loud

Sentiment Intensity

?
Multimodal Sentiment and Emotion Analysis

Speaker’s behaviors

“This movie is sick”

Sentiment Intensity

?

Multimodal Representation
(Multimodal Fusion)

① Intra-modal Interactions
② Cross-modal Interactions
③ Computational Efficiency
Multimodal Fusion using Tensor Representation

Intra-modal interactions

Cross-modal interactions

Computational efficiency

“This movie is sick”

Multimodal Representation

Unimodal

Bimodal

$h$

$h$

$Z = [Z_v] \otimes [Z_l] = [Z_v] 1 \otimes [Z_v] 1$

“Tensor Fusion Network for Multimodal Sentiment Analysis” by Zadeh, A., et al. (2017)
Computational Complexity – Tensor Product

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

\[ O(d_1 \times d_2 \times d_3) \]

\[ O(d_1 \times d_2) \]

\[ M=2 \quad M=3 \]

Number of Modalities

Computationally Complex

\[ \mathbb{Z} \]
CORE CONTRIBUTIONS

Low-rank Multimodal Fusion (LMF)
From Tensor Representation to Low-rank Fusion

Visual  
Language

① Decomposition of weight $W$.

② Decomposition of input tensor $Z$.

③ Rearrange the computation of $h$.

Tensor Fusion Networks

Low-rank Multimodal Fusion
Canonical Polyadic (CP) Decomposition of tensors

\[ W = W_{v}^{(1)} \otimes W_{l}^{(1)} + W_{v}^{(2)} \otimes W_{l}^{(2)} + \cdots \]

Rank of tensor \( W \): minimum number of vector tuples needed for exact reconstruction
Canonical Polyadic (CP) Decomposition of 3D tensors

\( \mathcal{W} = |h| \otimes \mathcal{L} + \otimes \mathcal{L} + \cdots \)
Modality-specific Decomposition

\[ \mathcal{W} = W_{v}^{(1)} \times w_{l}^{(1)} + W_{v}^{(2)} \times w_{l}^{(2)} + \ldots \]

Retain the dimension for the multimodal representation $h$ during decomposition
 Decomposition of weight tensor $W$
Decomposition of weight tensor $W$

$$\mathbf{Z} = \mathbf{w}_v^{(1)} \otimes \mathbf{w}_l^{(1)} + \mathbf{w}_v^{(2)} \otimes \mathbf{w}_l^{(2)} + \cdots = \mathbf{h}$$
\[Z = h\]

\[Z = w^{(1)}_v \times w^{(1)}_l + w^{(2)}_v \times w^{(2)}_l + \cdots\]
③ Rearranging computation

\[
\begin{align*}
\left[ w_v^{(1)} + w_v^{(2)} + \cdots + w_v^{(r)} \right] \cdot [1] \cdot \left[ w_l^{(1)} + w_l^{(2)} + \cdots + w_l^{(r)} \right] = h
\end{align*}
\]
Low-rank Multimodal Fusion

$x_v \rightarrow f_v$

$z_v$

$x_l \rightarrow f_l$

$z_l$

Low-rank Multimodal Fusion

$h$

Prediction

Task output

Low-rank factors

$w^{(1)}_v + w^{(2)}_v + \ldots + w^{(r)}_v \cdot z_v$

Low-rank factors

$w^{(1)}_l + w^{(2)}_l + \ldots + w^{(r)}_l \cdot z_l$
Easily scales to more modalities

- Intra-modal interactions
- Cross-modal interactions
- Computational complexity
EXPERIMENTS AND RESULTS
Datasets

CMU-MOSI

Sentiment Analysis
2199 video segments
• Single-speaker
• From 93 Movie reviews

Segment level annotations
• Sentiment
• Real-valued

POM

Speaker Trait Recognition
1000 full video clips
• Single-speaker
• Movie reviews

Video level annotations
• 16 types of speaker traits
• Categorical annotations

IEMOCAP

Emotion Recognition
10039 video segments
• Dyadic interaction
• From 302 videos

Segment level annotations
• 10 classes of emotions
• Categorical annotations
Compare to full rank tensor fusion

CMU-MOSI

Low-rank Multimodal Fusion (Our Model)

Tensor Fusion Networks (Zadeh, et al., 2017)
Compare to full rank tensor fusion

|                | CMU-MOSI | POM | IEMOCAP |
|----------------|---------|-----|---------|
| **MAE ↓**      | 0.67    | 0.90| 86.0    |
| **Correlation ↑** | 0.91   | 0.89| 85.8    |
| **F1-Happy ↑**  | 0.97    | 1.0 | 83.6    |
| **F1-Sad ↑**    | 0.67    | 0.40| 85.9    |

*Data includes CMU-MOSI, POM, and IEMOCAP datasets.*
Compare with State-of-the-Art Approaches

CMU-MOSI

Low-rank Multimodal Fusion (our model)

Memory Fusion Networks (Zadeh, et al., 2018)

Multi-attention Recurrent Networks (Zadeh, et al., 2018)

Tensor Fusion Networks (Zadeh, et al., 2017)

Multi-view LSTM (Rajagopalan, et al., 2016)

Deep Fusion (Nojavanasghari, et al., 2016)
Compare with Top 2 State-of-the-Art Approaches

| Dataset       | CMU-MOSI | POM | IEMOCAP |
|---------------|----------|-----|---------|
| MAE           |          |     |         |
| LMF           | 0.912    | 0.805 | 85.9   |
| MFN           | 0.965    | 0.396 | 84.3   |
| MARN          | 0.968    | 0.349 | 84.2   |
| TFN           | 0.67     | 0.270 | 81.0   |
| MV-LSTM       | 0.668    | 0.886 | 90.0   |
| Correlation↑  | 0.632    | 0.633 | 89.0   |
| MAE↓          | 0.60     | 0.805 | 84.2   |
| Correlation↑  | 0.796    | 0.396 | 82.8   |
| F1-Angry↑     | 0.67     | 0.349 | 82.1   |
| F1-Sad↑       | 0.89     | 0.270 | 82.1   |

Legend:
- **LMF**
- **MFN**
- **MARN**
- **TFN**
- **MV-LSTM**
**Efficiency Metric**: Number of data samples processed per second

- Training Efficiency
- Testing Efficiency
Conclusions

- Intra-modal interactions
- Cross-modal interactions
- Computational complexity
- State-of-the-art results
Thank you!

Code: https://github.com/Justin1904/Low-rank-Multimodal-Fusion

http://multicomp.cs.cmu.edu/