Memory Fusion Network for Multi-view Sequential Learning

Amir Zadeh, Paul Pu Liang, Navonil Manzumder, Soujanya Poria, Erik Cambria, Louis-Philippe Morency

Presenter: Amir Zadeh
Multi-view Sequential Data in AI

Multimedia

Self Driving Cars

Intelligent Personal Assistants

Robotics
Multi-view Sequential Data

Data that exhibits **Multiview** and **Sequential** properties.
Multi-view Sequential Data

Data that exhibits **Multiview** and **Sequential** properties.

**Multiview:** views with different distributions.
Multi-view Sequential Data

Data that exhibits **Multiview** and **Sequential** properties.

**Sequential**: data in each view is chronological.

\[ t = 1 \quad t = 2 \quad t = 3 \quad t = 4 \quad t = \ldots \quad t = T \]

Interesting, here’s why I think the whole so much fun.
Research Goal

Supervised Multi-view Sequential Modeling

Multiview sequential data is mapped to a sequence of labels

\[ t = 1 \quad t = 2 \quad t = 3 \quad t = 4 \quad t = \ldots \quad t = T \]

Interesting, here's why I think the whole so much fun.

Happiness

Multiview Data

Sequence of Labels
Challenge 1: Intra-view Dynamics

Visual View

$t = 1$  
$t = 2$  
$t = 3$  
$t = 4$  
$t = \cdots$  
$t = T$

Smile  
Smile  
Expressive body language

Language View

$t = 1$  
$t = 2$  
$t = 3$  
$t = 4$  
$t = \cdots$  
$t = T$

Interesting, here’s why I think the whole so much fun.

Positive word Positive phrase
Challenge 2: Cross-view Dynamics

**Visual View**

$t = 1$ | $t = 2$ | $t = 3$ | $t = 4$ | $t = \cdots$ | $t = T$
---|---|---|---|---|---
Smile | Smile | Expressive body language

**Language View**

$t = 1$ | $t = 2$ | $t = 3$ | $t = 4$ | $t = \cdots$ | $t = T$
---|---|---|---|---|---

here’s why | I think | the whole | so much fun.
**Challenge 2**: Cross-view Dynamics

**Visual View**

- $t = 1$
- $t = 2$
- $t = 3$
- $t = 4$
- $t = \cdots$
- $t = T$

Smile

Smile

Expressive body language

**Language View**

- $t = 1$
- $t = 2$
- $t = 3$
- $t = 4$
- $t = \cdots$
- $t = T$

**Positive word**

Interesting, here’s why

I think the whole

Positive phrase

so much fun
Multi-view Sequential Modeling Challenges

**Challenge 1**
Intra-view Dynamics
The dynamics within each view

**Challenge 2**
Cross-view Dynamics
The dynamics across views
System of RNNs:

- Parallel RNNs where each RNN encodes one view.
- Alignment has been defined between them.
Memory Fusion Framework

A two step fusion framework:

- **Delta-memory Attention Network**: Identifies cross-view dynamics across different modalities in each iteration of System of RNNs.

- **Multi-view Gated Memory**: Store the cross-view dynamics in a neural memory component.

**Challenge 2**
Memory Fusion Network (MFN)

Delta-memory Attention Network

Multi-view Gated Memory

System of LSTMs
Datasets

6 publicly available datasets

**CMU-MOSI**: utterance level multimodal sentiment analysis.
**POM**: video level speaker traits analysis.
**IEMOCAP**: utterance level multimodal emotion recognition.
**ICT-MMIMO**: video level multimodal sentiment analysis.
**YouTube**: video level multimodal sentiment analysis.
**MOUD**: utterance level Spanish multimodal sentiment analysis.

*language, vision and audio views*
Baselines

More than 20 baselines

Notable Baselines:

**MV-LSTM**: an extension of LSTM for multi-view learning.

**MV-HCRF, MV-LDHCRF, MV-HSSHCRF**: graphical models for multi-view learning.

**Early Fusion LSTM**: LSTM that concatenates all views into one single vector.

**Dataset Specific Baselines**: Tensor Fusion Network (TFN), Convolutional Multiple Kernel Learning (CMKL), Deep Fusion Network.
Experiments

Binary Classification: Accuracy, F1

Multi-class Classification: Accuracy, F1

Regression: MAE, Correlation
Multimodal Sentiment Analysis

CMU-MOSI BINARY ACCURACY SENTIMENT

ICT-MM MO BINARY ACCURACY SENTIMENT

YOUTUBE BINARY ACCURACY SENTIMENT

MOUD BINARY ACCURACY SENTIMENT

MFN  SOTA1  SOTA2
Multimodal Speaker Traits and Emotion Recognition

- POM CONFIDENT 7-CLASS
  - MFN: 34.5
  - SOTA1: 26.6
  - SOTA2: 26.6

- POM DOMINANT 7-CLASS
  - MFN: 41.9
  - SOTA1: 35.1
  - SOTA2: 35.1

- POM PERSUASIVE 7-CLASS
  - MFN: 34.1
  - SOTA1: 28.1
  - SOTA2: 27.6

- IEMOCAP 9-CLASS EMOTIONS
  - MFN: 36.5
  - SOTA1: 36
  - SOTA2: 35.9
Delta-memory Attention

![Bar chart showing comparison between different sentiment analysis methods.](chart.png)

- CMU-MOSI Binary Sentiment
- ICT-MMMO Binary Sentiment
- MOUD Binary Sentiment
- POM Confident 7-class

**Legend:**
- MFN
- MFN (no Delta)
Multi-view Gated Memory

- CMU-MOSI Binary Sentiment
- ICT-MMMO Binary Sentiment
- MOUD Binary Sentiment
- POM Confident 7-class

Graph showing performance with MFN and MFN (no Delta) for different binary sentiment tasks.
ACL 2018 Workshop:
First Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language
http://multicomp.cs.cmu.edu/acl2018multimodalchallenge/

CMU-MOSEI Dataset:
Largest dataset of multimodal sentiment analysis and emotion recognition

Contact: abagherz@cs.cmu.edu
Code: https://github.com/A2Zadeh/MFN
Training data: https://github.com/A2Zadeh/CMU-MultimodalDataSDK
Thank you!

Delta-memory Attention Network

Multi-view Gated Memory

System of LSTMs