Modeling Spatiotemporal Multimodal Language with Recurrent Multistage Fusion

Paul Pu Liang, Ziyin Liu, Amir Zadeh, Louis-Philippe Morency
School of Computer Science, Carnegie Mellon University
{pliang, ziyinl, abagherz, morency}@cs.cmu.edu

Artificial Intelligence and Multimodal Language

Giving Artificial Intelligence the power to model human language is a core research challenge.

Multimodal Language Modalities

Language
- Lexicon
- Syntax
- Pragmatics
- Prosody

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

Acoustic
- Prosoody
- Vocal expressions

Challenge 1: Intra-modal Interactions

Intra-modal interactions exist within each modality independent of other modalities (temporal interactions).

a) Temporal sequences

Challenge 2: Cross-modal Interactions

Cross-modal interactions refer to interactions between modalities (spatial interactions).

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

Recurrent Multistage Fusion

FUSE. The highlighted multimodal signals are simultaneously fused in a local fusion and then integrated with fusion representations from previous stages.

\[ a_t^{[k]} = f_t(b_t^{[k]}, a_{t-1}^{[k]}, \delta) \]  \hspace{1cm} (3)

SUMMARIZE. After completing \( K \) stages of HIGHLIGHT and FUSE, the SUMMARIZE operation generates a cross-modal representation using all final fusion representations \( a_T^{[k]} \).

\[ a_t = S(a_T^{[k]}) \theta \]  \hspace{1cm} (4)

Experiments

Datasets
- Multimodal Sentiment Analysis: CMU-MOSI
- Multimodal Emotion Recognition: IEMOCAP
- Multimodal Personality Traits Prediction: POM
- Language, visual and acoustic features extracted and aligned by P2FA

Baseline Models
- Non-temporal Models
- Multimodal Temporal Graphical Models
- Multimodal Temporal Neural Networks

Results

SUMMARIZE

Synchronized Interactions

Asynchronous Trimodal Interactions

Visualizations

- Attention weights change across multiple stages of fusion.
- Attention weights vary over time and adapt to the multimodal inputs.
- Language and acoustic modalities most commonly highlighted.

Conclusion

RMFN decomposes the multimodal fusion problem into multiple stages, each focused on a subset of multimodal signals. Multiple stages coordinate to capture both synchronous and asynchronous multimodal interactions.