Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling

Ming Hou\textsuperscript{1, *}, Jiajia Tang\textsuperscript{2,1, *}, Jianhai Zhang\textsuperscript{2}, Wansen Kong\textsuperscript{2} and Qibin Zhao\textsuperscript{1}
\textsuperscript{1}Tensor Learning Unit, RIKEN AIP, \textsuperscript{2}College of Computer Science, Hangzhou Dianzi University

1. Outline

Multimodal Learning for Recognition
- Multimodal recognition integrates features of multiple modalities (language, acoustic and visual) for yielding robust predictions.
- Multimodal fusion is a key step of multimodal recognition.
- Tensor-based fusion methods have achieved a great success.

Limitations of existing tensor fusion
- Restrict interaction among modalities to be linear w.r.t. each modality
- Ignore high-order statistical information

Our contributions
- Explicitly model nonlinear intra-modal and cross-modal interactions via high-order polynomial moments
- Directly model local interactions across mixed dimensions over time
- Significantly reduce heavy computation via using tensor network

2. Polynomial Tensor Pooling (PTP)

- PTP block first fuses M feature vectors using high-order moments and then transforms them into a joint representation.

\[ P = \prod_{i=1}^{P} f_i \]

- Fusion via high-order moment is obtained by P-order tensor product of concatenated features:

\[ f_i = [1, x_1, x_2, \ldots, x_{m_i}] \]

\[ F = \bigotimes f \]

- Transformation is performed by a weight tensor:

\[ z_h = \sum_{i_1, i_2, \ldots, i_P} W_{i_1, i_2, \ldots, i_P} \cdot F_{i_1, i_2, \ldots, i_P} \]

- Low-rank tensor networks is used to reduce large computation.

\[ z_h = \sum_{i_1, i_2, \ldots, i_P} \left[ \sum_{r=1}^{R} \alpha_{i_1, i_2, \ldots, i_P} \prod_{r=1}^{P} f_{i_r} \right] = \sum_{r=1}^{R} \prod_{p=1}^{P} \sum_{i_p=1}^{I} \alpha_{i_1, i_2, \ldots, i_P} f_{i_r} \]

3. Hierarchical Polynomial Fusion Network (HPFN)

- Features of multiple modalities are rearranged into a “feature map”.
- Single-layer HPFN is constructed by a global PTP operating on a “receptive window” across all time steps and modalities.

- Multi-layer HPFN is established by recursing PTP blocks layer by layer into a tree-structured architecture.
  - Local temporal-modality interactions most relevant to prediction can be transmitted to the global level.
  - PTP can be treated as a “fusion filter” analogous to a CNN filter.
  - CNN-style fusion framework with flexible design choices bring benefits.

- Incorporate dense connectivity to enhance the expressive capacity.

- Model complexity depends on total number of PTP filters and window size, parameters are larger than LMF but much smaller than TFN.

4. Experimental Results

- Datasets and tasks:
  - CMU-MOSI utterance level multimodal sentiment analysis with intensity range in [-3, 3]
  - IEMOCAP utterance level binary classification for emotions, including neutral, angry, happy and sad
  - Accuracy comparisons of HPFN with other methods

- Effect of fusion of one-layer HPFN with higher order polynomial

- Effect of fusion of HPFN with depth, connectivity

- Effect of fusion with and without the incorporation of temporal factors