Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling

Ming Hou[1,*], Jiajia Tang[2,1,*], Jianhai Zhang[2], Wanzen Kong[2] and Qibin Zhao[1]

[1] Tensor Learning Unit, RIKEN AIP, [2] College of Computer Science, Hangzhou Dianzi University

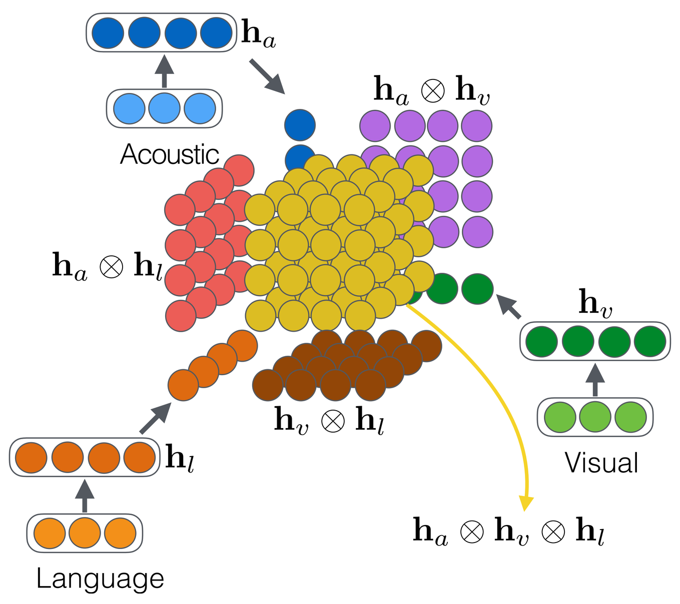
1. Outline

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

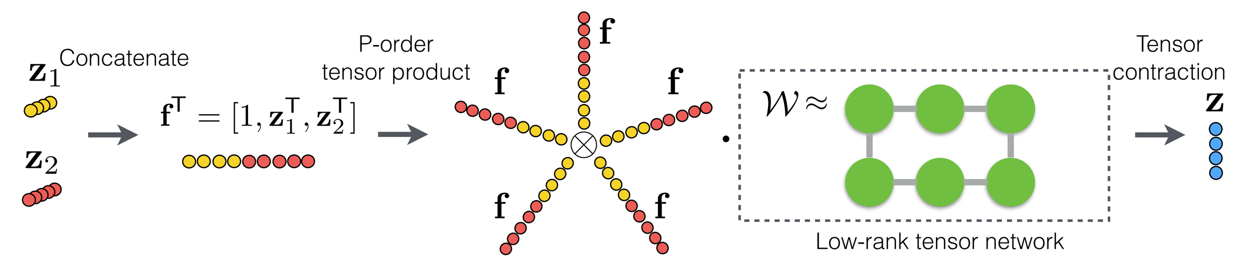


Our contributions

- 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.



- Fusion via high-order moment is obtained by P-oder tensor product of concatenated features: $\mathbf{f}^{\mathsf{T}} = [1, \mathbf{z}_1^{\mathsf{T}}, \mathbf{z}_2^{\mathsf{T}}, \cdots, \mathbf{z}_M^{\mathsf{T}}]$

$$\mathcal{F} = \underbrace{\mathbf{f} \otimes \mathbf{f} \otimes \cdots \otimes \mathbf{f}}_{\text{P-order}}$$

- Transformation is performed by a weight tensor : $\mathcal{W} = [\mathcal{W}^1, ..., \mathcal{W}^h, ..., \mathcal{W}^H]$

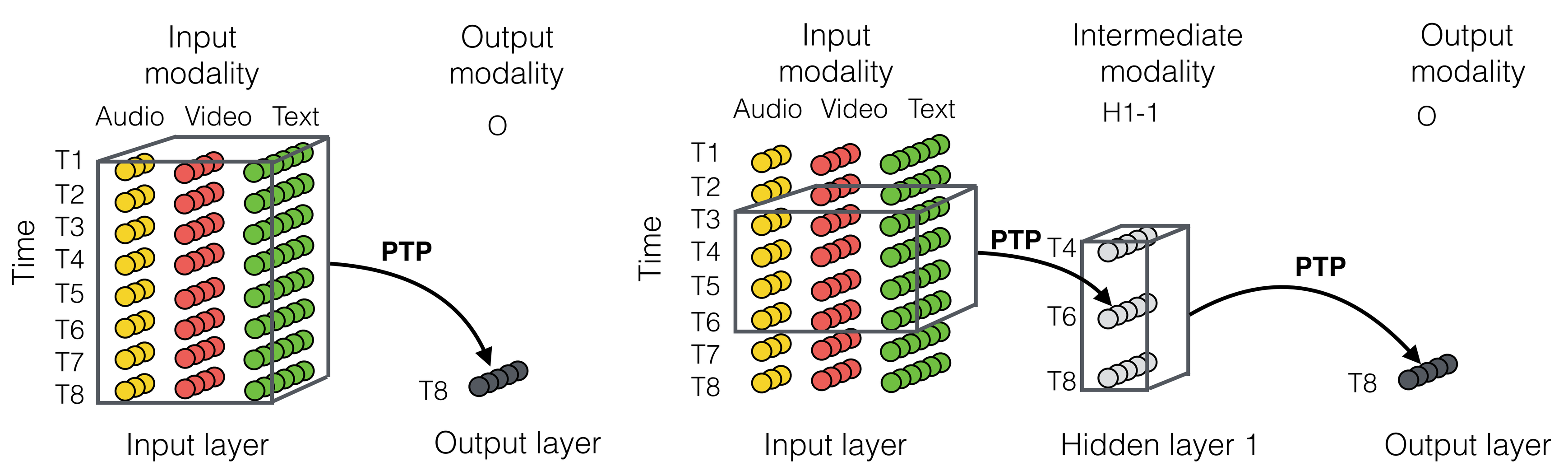
$$z_h = \sum_{i_1, i_2, \cdots, i_P} \mathcal{W}^h_{i_1 i_2 \cdots i_P} \cdot \mathcal{F}_{i_1 i_2 \cdots i_P}$$

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

$$z_h = \sum_{i_1, i_2, \cdots, i_P} \left[\left(\sum_{r=1}^{R} a_r^h \prod_{p=1}^{P} \mathbf{w}^h_{r\, i_p} \right) \left(\prod_{p=1}^{P} \mathbf{f}_{i_p} \right) \right] = \sum_{r=1}^{R} a_r^h \prod_{p=1}^{P} \sum_{i_p}^{I} \mathbf{w}^h_{r\, i_p} \mathbf{f}_{i_p}$$

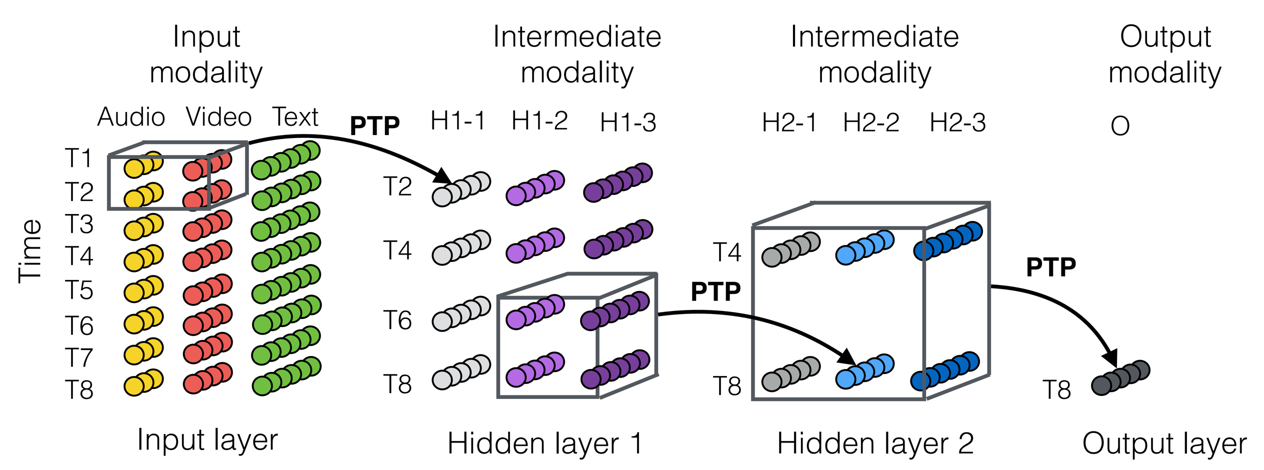
3. Hierarchical Polynomial Fusion Network (HPFN)

- 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.

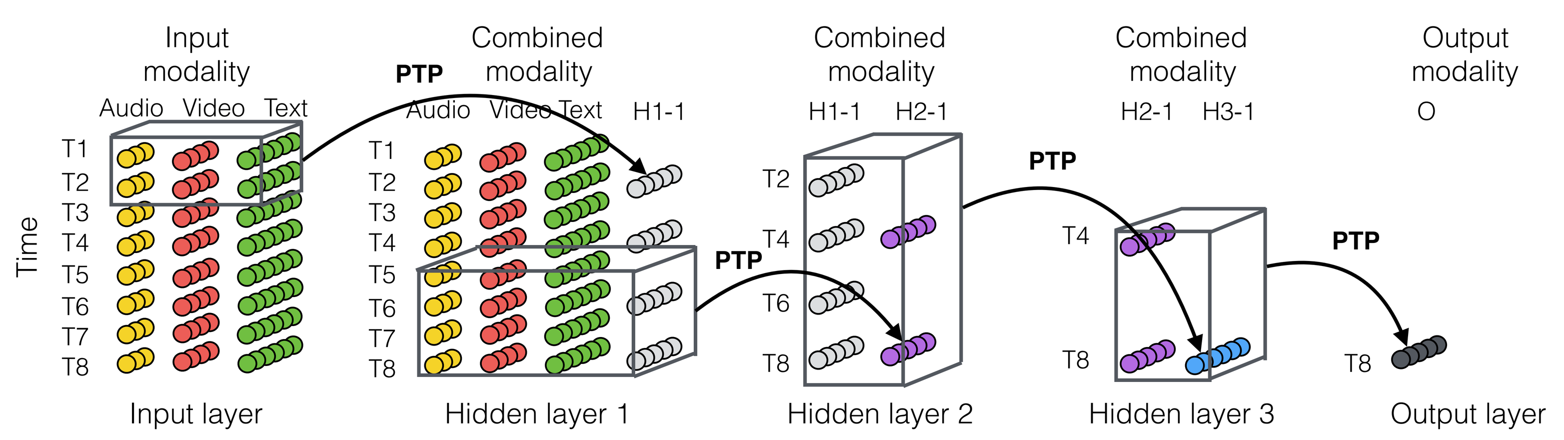


(a) one-layer HPFN (b) two-layer HPFN

- 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

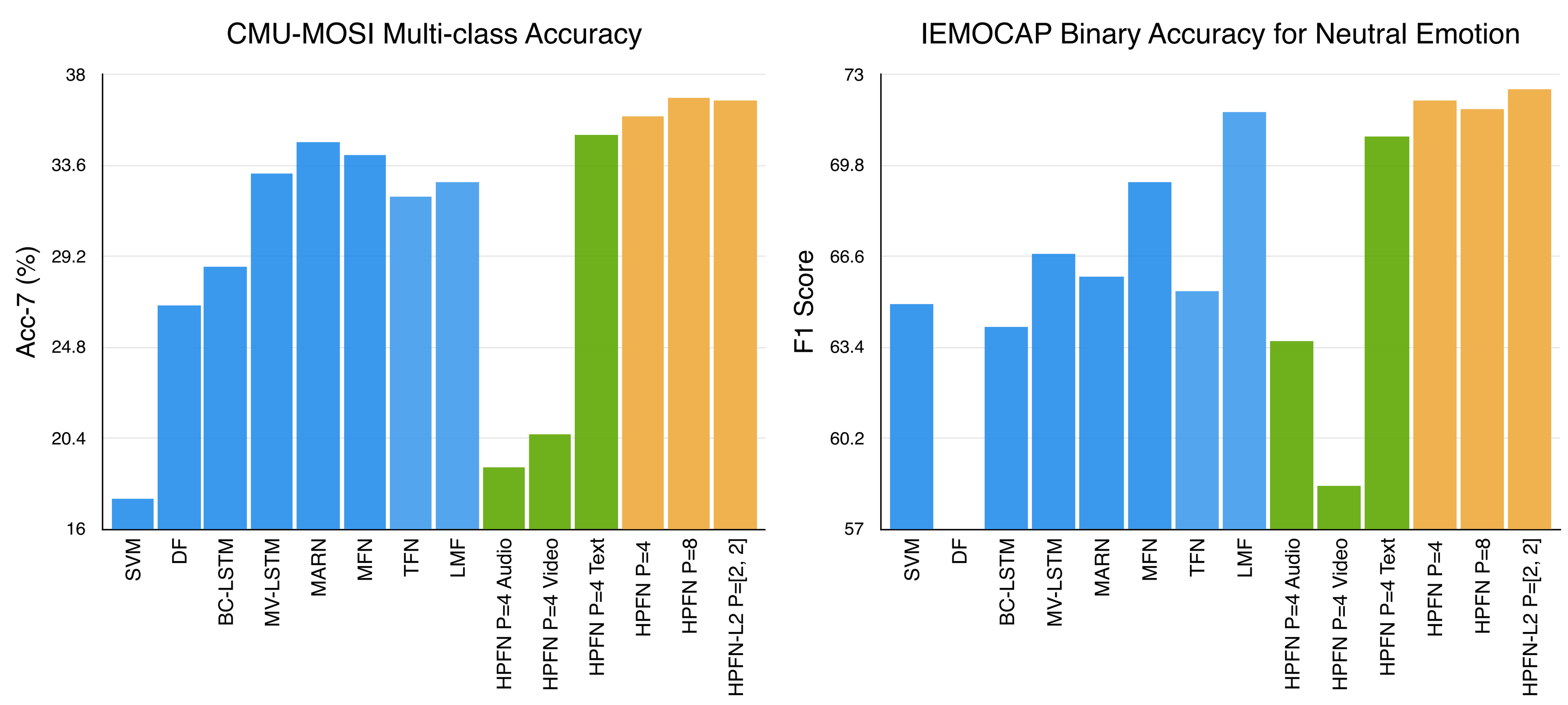


  - Model complexity depends on total number of PTP filters and window size, parameters are larger than LMF but much smaller than TFN.

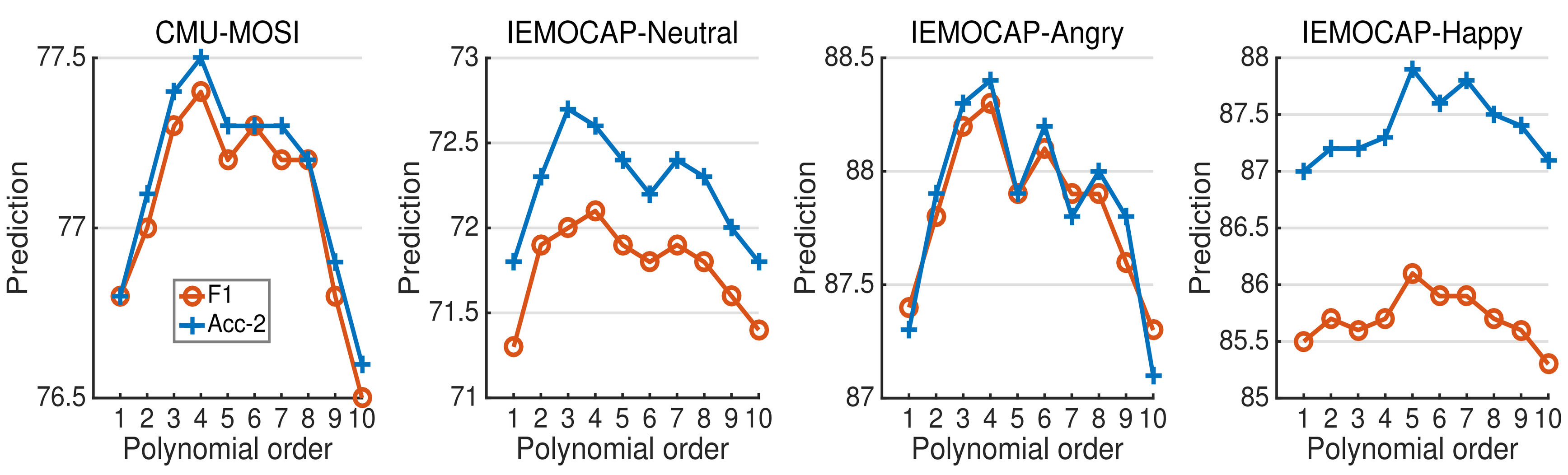
Model	TFN [non-temporal]	LMF [non-temporal]	PTP [temporal]	HPFN (L layers) [temporal]
Param.	$\mathcal{O}(I_y \prod_{m=1}^{M} I_m)$	$\mathcal{O}(I_y R(\sum_{m=1}^{M} I_m))$	$\mathcal{O}(I_y R(\sum_{t=1}^{T} \sum_{m=1}^{S} I_{t,m}))$	$\mathcal{O}(I_y R(\sum_{l=1}^{L} N_l)(\sum_{t=1}^{T} \sum_{m=1}^{S} I_{t,m}))$

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 oder polynomial



- Effect of fusion of HPFN with depth, connectivity
- Effect of fusion with and without the incorporation of temporal factors