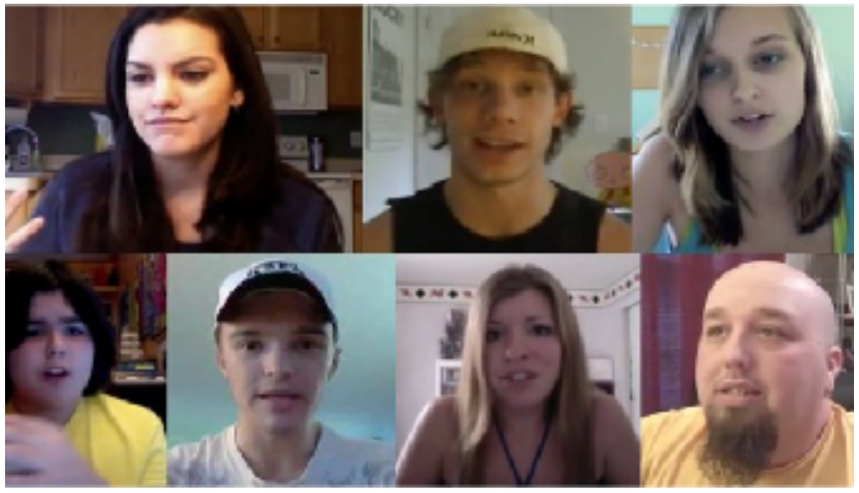
Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling

Multimodal Data in Artificial Intelligence

Multimedia Content

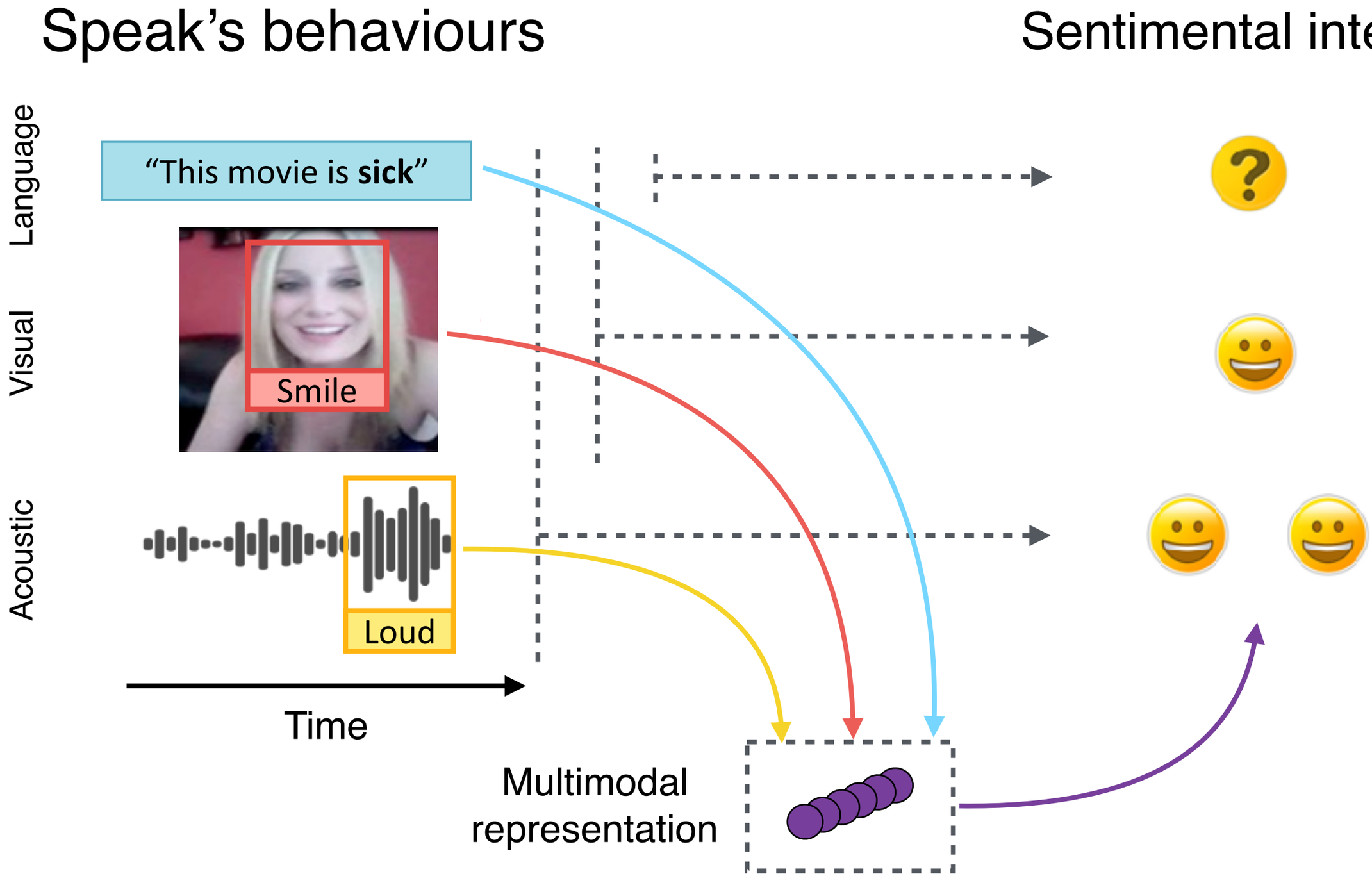


Intelligent Personal Assistants

Self-driving Cars

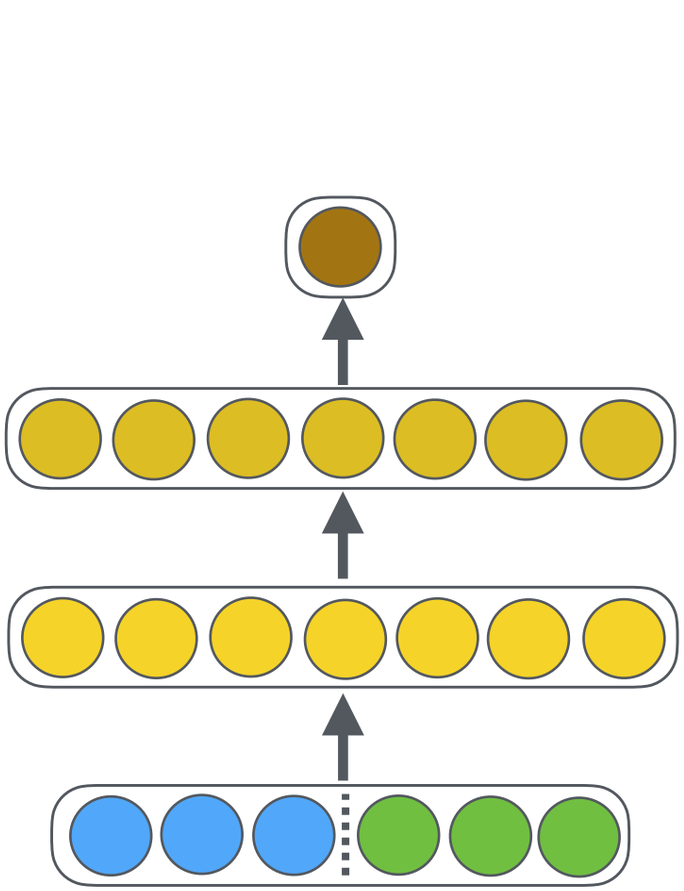
Robotics

Example: Multimodal Sentimental Analysis

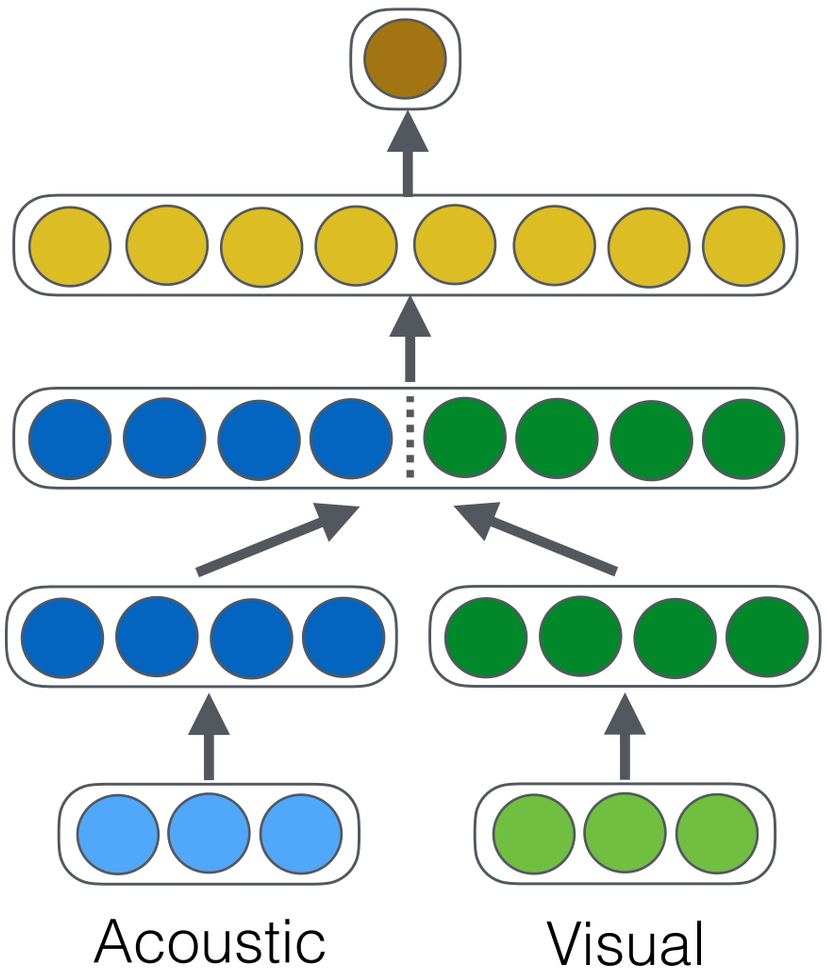


Conventional Multimodal Feature Fusion

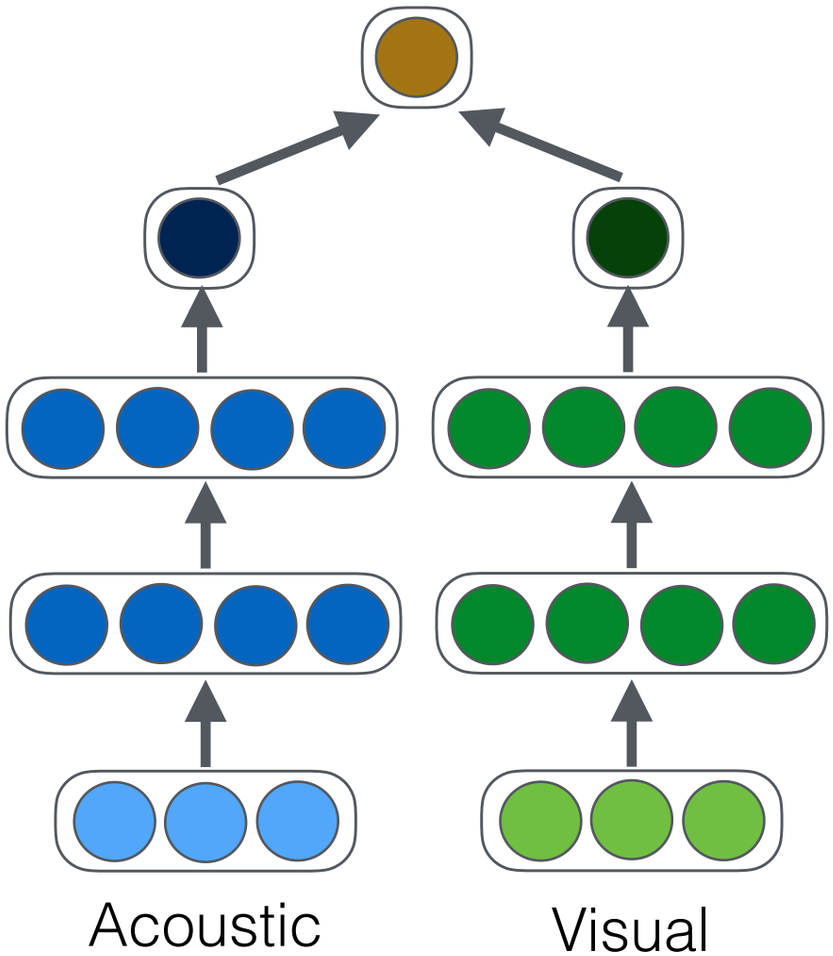
Early fusion



Model-level fusion



Late fusion

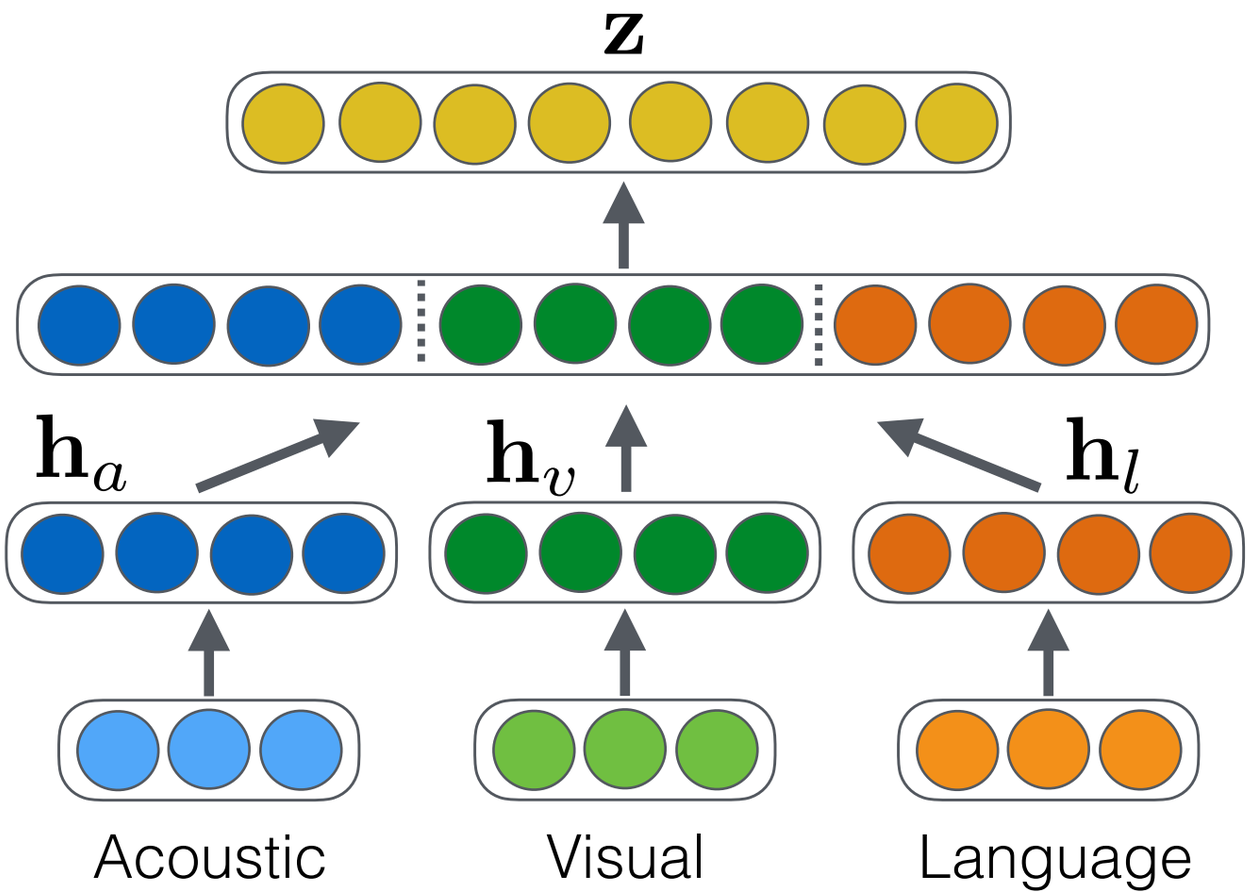


Fused Multimodal Representation

Joint multimodal representation

- Modal-level fusion
- Simply by linear transformation of concatenated individual features:

$$\mathbf{z} = f(\mathbf{W}[\mathbf{h}_a, \mathbf{h}_v, \mathbf{h}_l])$$

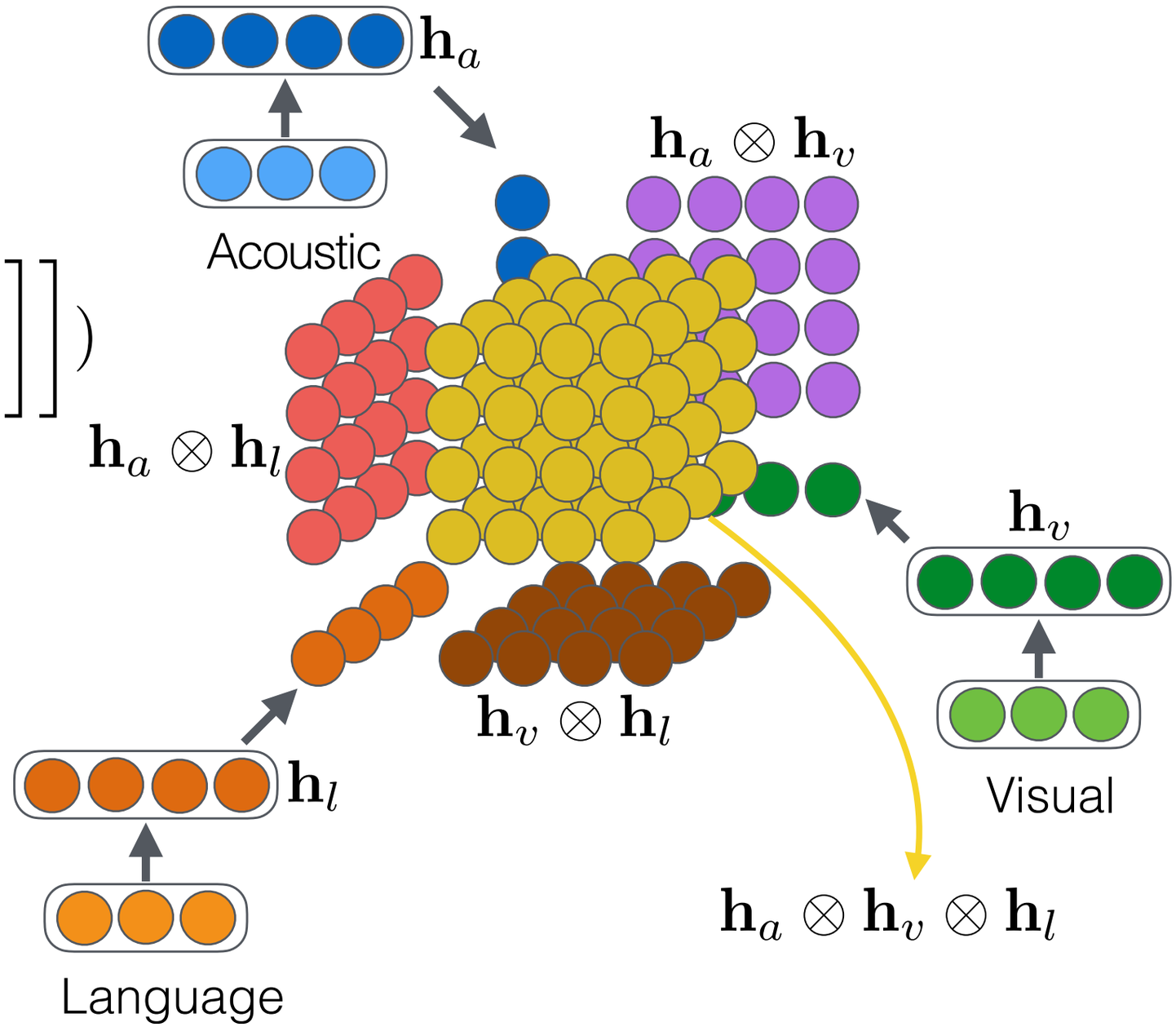


Tensor Fused Multimodal Representation

Bimodal and trimodal representation

- Explicitly capture linear, bilinear and trilinear interactions
- Obtained by tensor product of individual features:

$$\mathbf{z} = f(\mathcal{W} \cdot \left[\begin{bmatrix}\mathbf{h}_a \\ 1\end{bmatrix} \otimes \begin{bmatrix}\mathbf{h}_v \\ 1\end{bmatrix} \otimes \begin{bmatrix}\mathbf{h}_l \\ 1\end{bmatrix}\right])$$



Limitations of Existing Fusion Strategies

Interaction among multiple modalities is always linear w.r.t. each modality

- Consider only up to bilinear interactions for two modalities
- Consider only up to trilinear interactions for three modalities

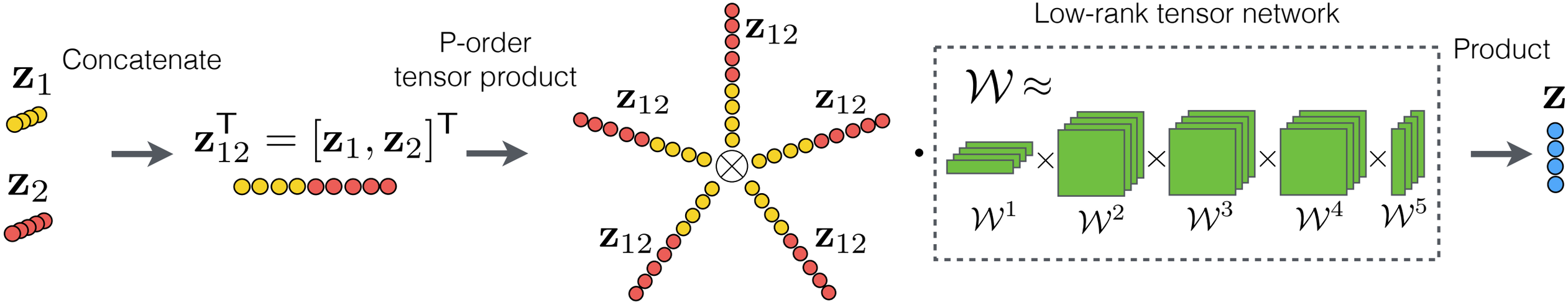
Fuse multimodal features all at once in a global manner, the complex local dynamics of interactions are ignored

- Unable to model sequential or temporal data
- Unable to capture the evolving temporal-modality intercorrelations

Tensor Polynomial Pooling (PTP)

Polynomial tensor pooling (PTP) block is able to capture high-order nonlinear interactions among features

- Explicitly model up to P-order polynomial interactions among multimodal features
- Serves as a basic building block for our hierarchical fusion framework



PTP Block

PTP block is to fuse a collection of feature vectors $\{\mathbf{z}_m\}_{m=1}^M$ into joint representation $\mathbf{z}$ by using high-order moments

- Concatenate $\{\mathbf{z}_m\}_{m=1}^M$ into a long feature vector $\mathbf{z}_{1\cdots M}$:

$$\mathbf{z}_{12\cdots M}^\top = [1, \mathbf{z}_1^\top, \mathbf{z}_2^\top, \cdots, \mathbf{z}_M^\top]$$

- P-polynomial feature tensor $\mathcal{Z}^P$ is obtained by constructing P-order tensor product of $\mathbf{z}_{1\cdots M}$:

$$\mathcal{Z}^P = \mathbf{z}_{12\cdots M} \otimes_1 \mathbf{z}_{12\cdots M} \otimes_2 \cdots \otimes_P \mathbf{z}_{12\cdots M}$$

- P-polynomial interactions among features is measured by pooling weight tensor $\mathcal{W}^h$

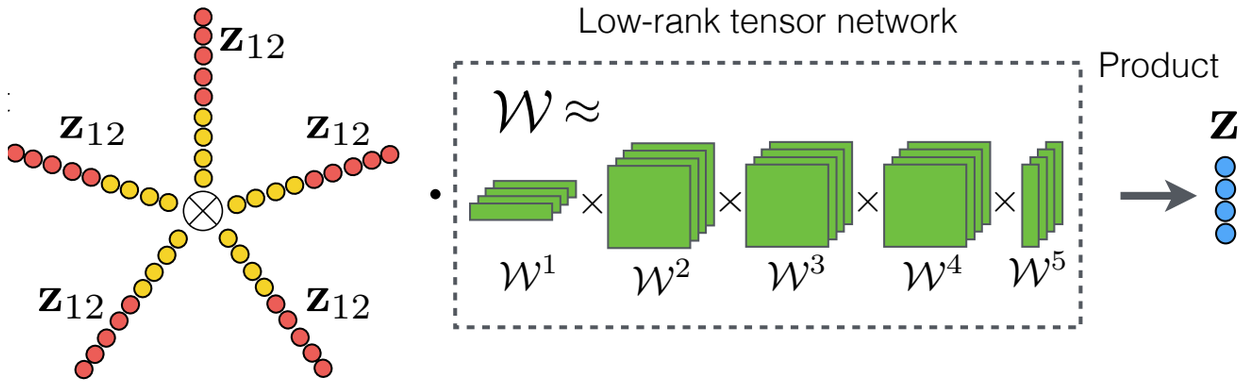
$$z_h = \sum_{i_1, i_2, \cdots, i_P} \mathcal{W}^h_{i_1 i_2 \cdots i_P} \cdot \mathcal{Z}^P_{i_1 i_2 \cdots i_P}$$

Computationally prohibitive for large P!

PTP Block Cont

PTP adopts low-rank tensor networks (TNs) to efficiently approximate $\mathcal{W}^h$

- If $\mathcal{W}^h$ admits rank-R CP format:



$$z_h = \sum_{i_1,i_2,\cdots,i_P} \mathcal{W}^h_{i_1 i_2 \cdots i_P} \left(\prod_{p=1}^{P} \mathbf{z}_{12\cdots m;i_p}\right)$$

$$= \sum_{i_1,i_2,\cdots,i_P} \left(\sum_{r=1}^{R} a_r^h \prod_{p=1}^{P} \mathbf{w}^{(p)}_{r;i_p}\right)\left(\prod_{p=1}^{P} \mathbf{z}_{12\cdots m;i_p}\right) = \sum_{r=1}^{R} a_r^h \prod_{p=1}^{P} \sum_{i_p}^{I} \mathbf{w}^{(p)}_{r;i_p} \mathbf{z}_{12\cdots m;i_p}$$

- If $\mathcal{W}^h$ admits low-rank TR format:

$$z_h = \sum_{i_1,i_2,\cdots,i_P} \mathcal{W}^h_{i_1 i_2 \cdots i_P} \left(\prod_{p=1}^{P} \mathbf{z}_{12\cdots m;i_p}\right) = \sum_{i_1,i_2,\cdots,i_P} \left(\sum_{r_1 r_2 \cdots r_P} \prod_{p=1}^{P} \mathcal{G}^{(p)}_{r_p;i_p;r_{p+1}}\right)\left(\prod_{p=1}^{P} \mathbf{z}_{12\cdots m;i_p}\right)$$

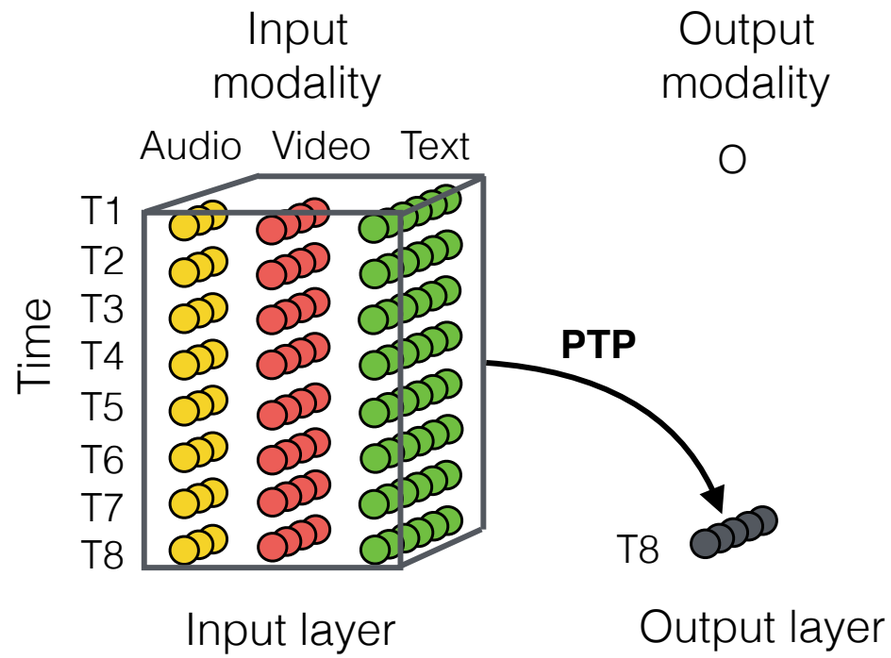
$$= \sum_{r_1 r_2 \cdots r_P} \prod_{p=1}^{P} \sum_{i_p}^{I} \mathcal{G}^{(p)}_{r_p;i_p;r_{p+1}} \mathbf{z}_{12\cdots m;i_p}$$

Computationally efficient for large P!

Hierarchical Polynomial Fusion Network (HPFN)

Hierarchical polynomial fusion network (HPFN) is able to model local interactions across both temporal and modality domains

- One-layer HPFN (HPFN-L1) is simply a PTP
- HPFN-L1 arranges the aligned temporal multimodal features just like a 2D image
- HPFN-L1 capture interactions within a receptive window (or scanning window) across all time steps and modalities

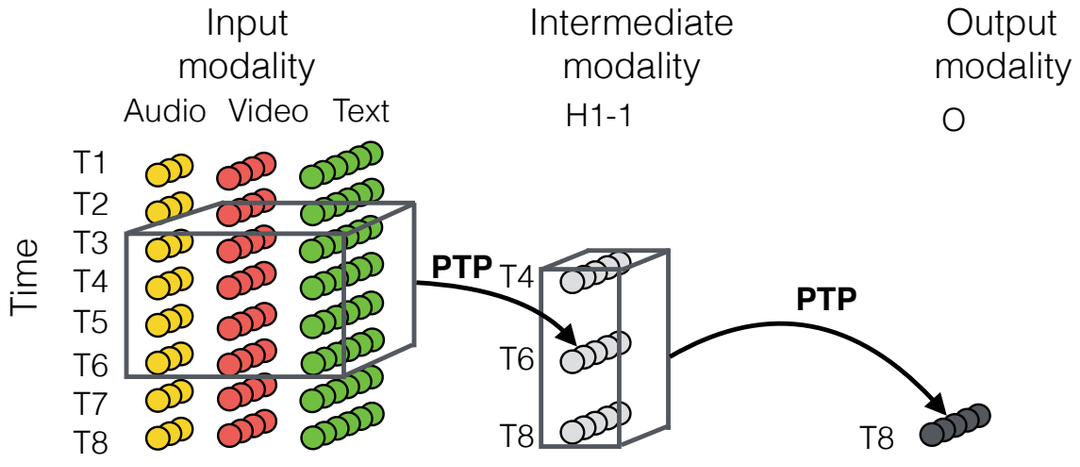


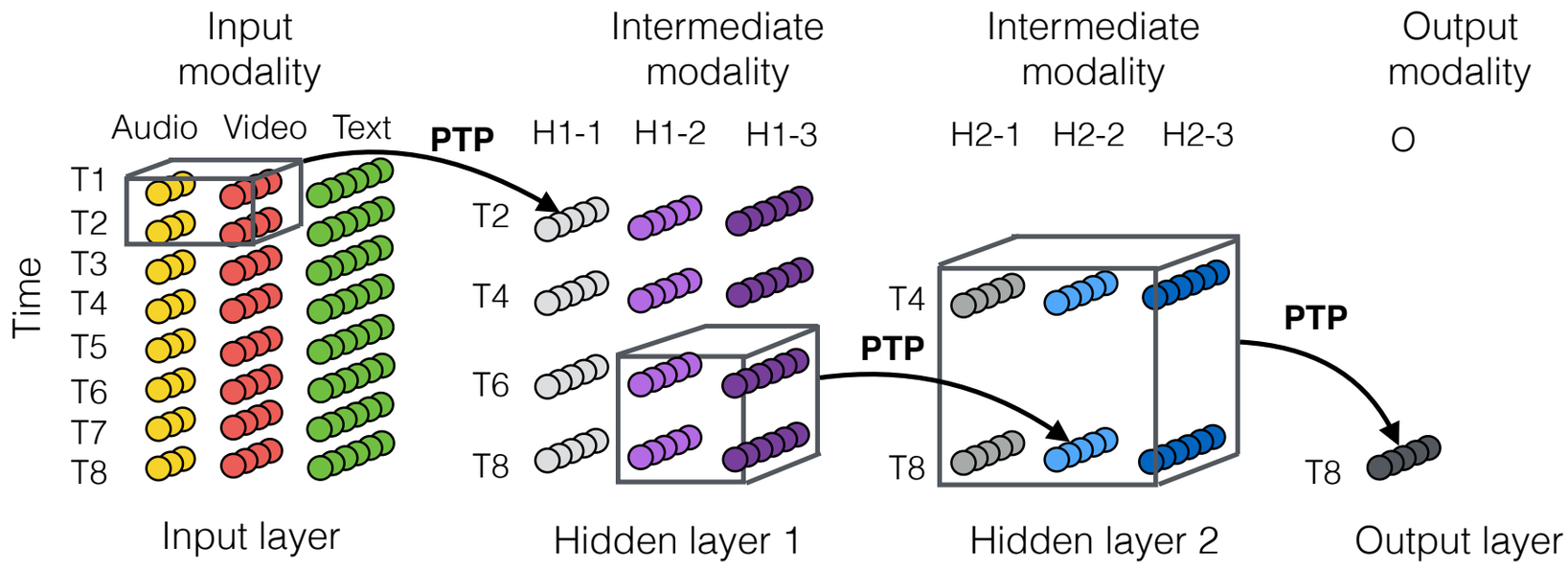
- Then a PTP block can be thought of as a fusion filter analogous to a CNN filter

HPFN Framework

Multi-layer HPFN is able to hierarchically fuse local features into a global feature by stacking PTP blocks layer by layer

- Dominant local temporal-modality intercorrelations can be transmitted into global scale
- Is a CNN-style fusion framework
  - Flexible architecture design choices:
    - window size
    - overlapped window
    - shared a fusion filter along multiple windows
    - associate a window with multiple fusion filters

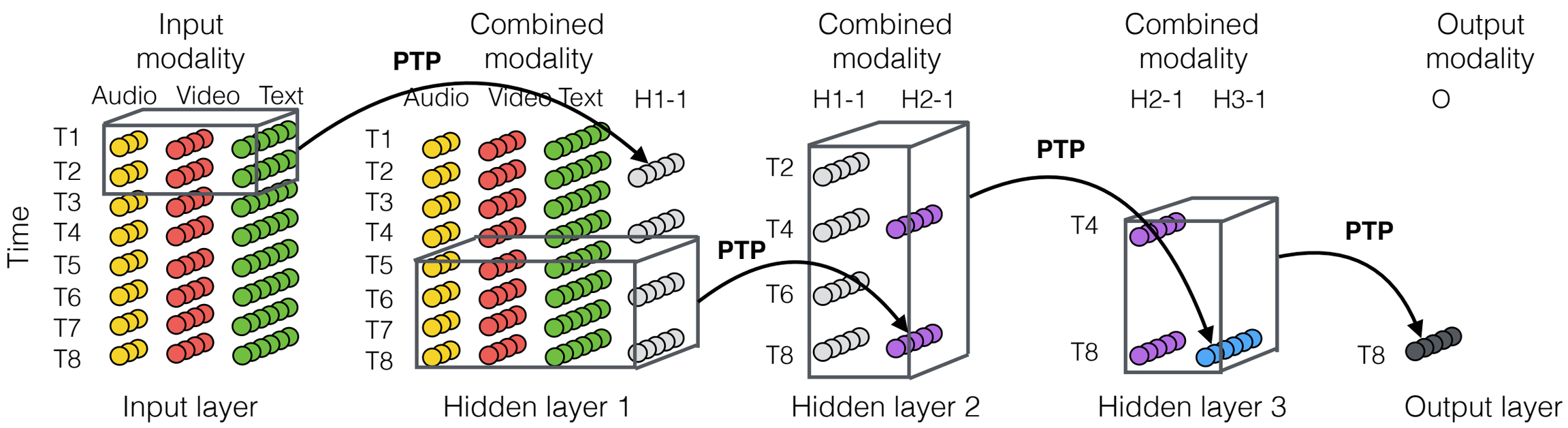




HPFN Framework Cont

More variations: densely-connected multi-layer HPFN

- Incorporate dense connectivity brings further enhanced expressive capacity to fusion model
- Can be beneficial in dealing with sequential signals

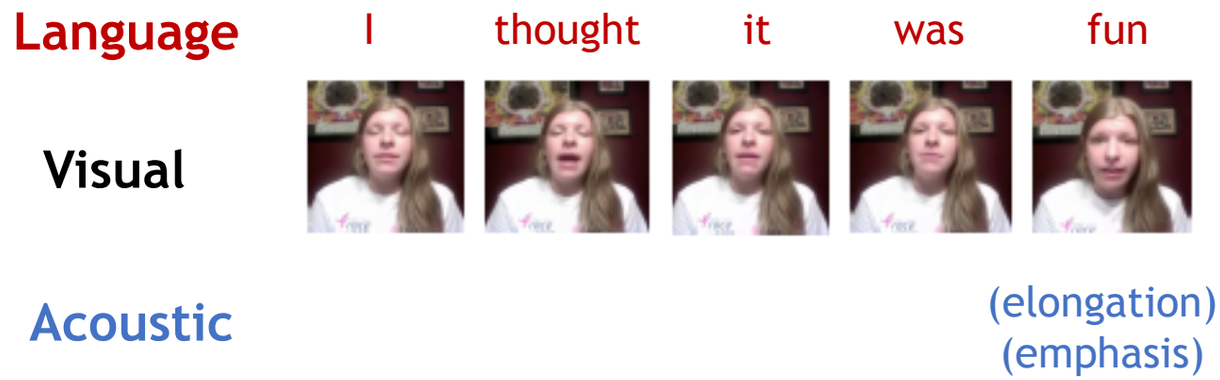


Densely-connected four-layer HPFN

Datasets

CMU-MOSI

- Utterance level multimodal sentiment analysis
- 2,199 english opinion segments (monologues) from 93 movie reviewers
- Sentimental intensity range in [-3, 3]



IEMOCAP

- Utterance level multimodal emotion recognition
- 10,039 video segments of dyadic interaction from 302 videos
- emotion categories: happy, sad, angry, neutral …

Experiments

Multimodal features

- Language: pre-trained Glove word embedding
- Visual: facial actions units from Facet
- Acoustic: MFCCs from COVAREP
- Features are aligned by P2FA

Metrics

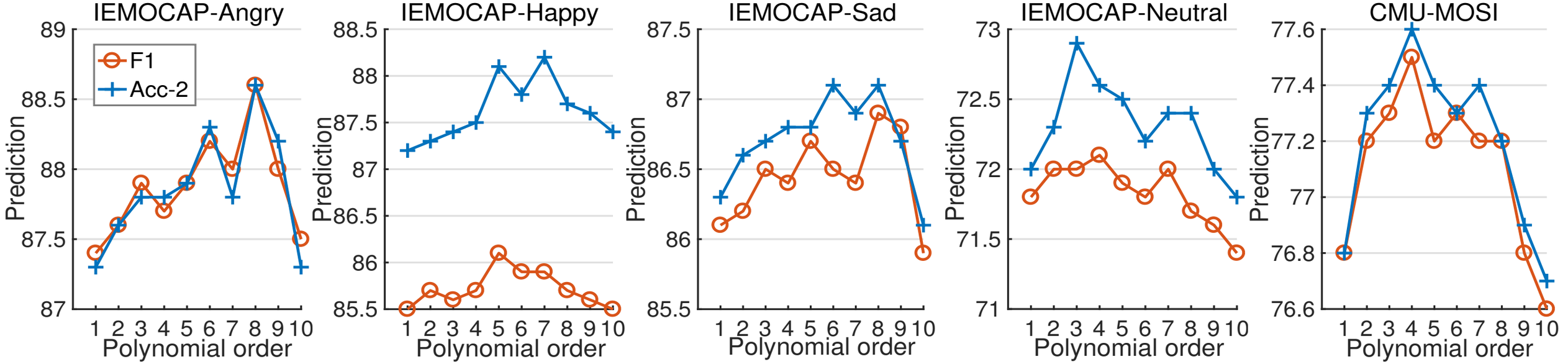
- Binary classification: Accuracy (Acc2), F1
- Multi-class classification: Accuracy (Acc7), F1
- Regression: MAE, Correlation

Overall Performance Comparison

Models	CMU-MOSI					IEMOCAP			
	MAE	Corr	Acc-2	F1	Acc-7	F1-Happy	F1-Sad	F1-Angry	F1-Neutral
SVM	1.864	0.057	50.2	50.1	17.5	81.5	78.8	82.4	64.9
DF	1.143	0.518	72.3	72.1	26.8	81.0	81.2	65.4	44.0
BC-LSTM	1.079	0.581	73.9	73.9	28.7	81.7	81.7	84.2	64.1
MV-LSTM	1.019	0.601	73.9	74.0	33.2	81.3	74.0	84.3	66.7
MARN	0.968	0.625	77.1	77.0	34.7	83.6	81.2	84.2	65.9
MFN	0.965	0.632	77.4	77.3	34.1	84.0	82.1	83.7	69.2
TFN	0.970	0.633	73.9	73.4	32.1	83.6	82.8	84.2	65.4
LMF	**0.912**	0.668	76.4	75.7	32.8	85.8	85.9	89.0	71.7
HPFN-L1, P=[4] (audio)	1.404	0.223	57.3	57.4	19.0	79.4	81.8	84.9	63.6
HPFN-L1, P=[4] (video)	1.409	0.221	57.0	57.1	20.6	83.2	73.2	72.3	58.5
HPFN-L1, P=[4] (text)	0.975	0.634	76.4	76.4	35.1	85.3	83.0	85.6	70.8
HPFN-L1, P=[4]	0.956	0.660	77.6	77.5	36.2	85.7	86.4	87.7	72.1
HPFN-L1, P=[8]	0.956	0.658	77.2	77.2	**38.3**	85.7	86.9	88.6	71.7
HPFN-L2, P=[2, 2]	0.931	**0.679**	**77.7**	**77.7**	37.9	**86.4**	**86.6**	**89.1**	**73.3**

Effect of the order of polynomial fusion

- The best accuracies of IEMCAP and MOSI are all achieved with P-order > 3 (greater than number of modalities M = 3)
- Signify the necessity and effectiveness of exploring high-order interactions in the multimodal fusion



Effect of architecture designs

- Effect of the depth and dense connectivity

Models	IEMOCAP				CMU-MOSI				
	F1-Happy	F1-Sad	F1-Angry	F1-Neutral	MAE	Corr	Acc-2	F1	Acc-7
HPFN-L1, P=[2]	85.7	86.2	87.6	72.0	0.970	0.636	77.2	77.2	37.3
HPFN-L2, P=[2, 2]	86.2	86.3	89.0	72.4	0.965	0.636	77.4	77.3	**37.8**
HPFN-L2-S1, P=[2, 2]	86.2	86.8	88.6	72.9	0.953	0.657	**78.0**	**77.8**	37.3
HPFN-L2-S2, P=[2, 2]	**86.4**	86.6	**89.1**	**73.3**	0.953	0.651	77.6	77.5	36.7
HPFN-L3, P=[2, 2, 1]	85.8	**87.0**	87.8	72.7	**0.941**	**0.666**	76.7	76.6	37.3
HPFN-L4, P=[2, 2, 2, 1]	84.5	86.0	87.5	71.8	0.986	0.637	76.7	76.6	34.7

- Effect of the modelling mixed temporal-modality features

Models	CMU-MOSI				
	MAE	Corr	Acc-2	F1	Acc-7
HPFN-L2, P=[2, 2] (non-temporal)	0.965	0.636	77.4	77.3	37.8
HPFN-L2, P=[2, 2] (temporal-overlapped, audio)	1.407	0.229	57.4	56.2	20.1
HPFN-L2, P=[2, 2] (temporal-overlapped, video)	1.358	0.183	61.2	61.3	20.3
HPFN-L2, P=[2, 2] (temporal-overlapped, text)	0.933	0.677	76.7	76.6	35.4
HPFN-L2, P=[2, 2] (temporal-overlapped)	**0.931**	**0.679**	**77.7**	**77.7**	**37.9**
HPFN-L2, P=[2, 2] (weight-shared)	0.943	**0.679**	77.6	77.5	37.0