Hierarchical Tensor Fusion Network for Deception Handling Negotiation Dialog Model

Nguyen The Tung[1], Koichiro Yoshino[1,2,3], Sakriani Sakti[1,3], and Satoshi Nakamura[1,3]

[1]Augmented Human Communication – Nara Institute of Science and Technology

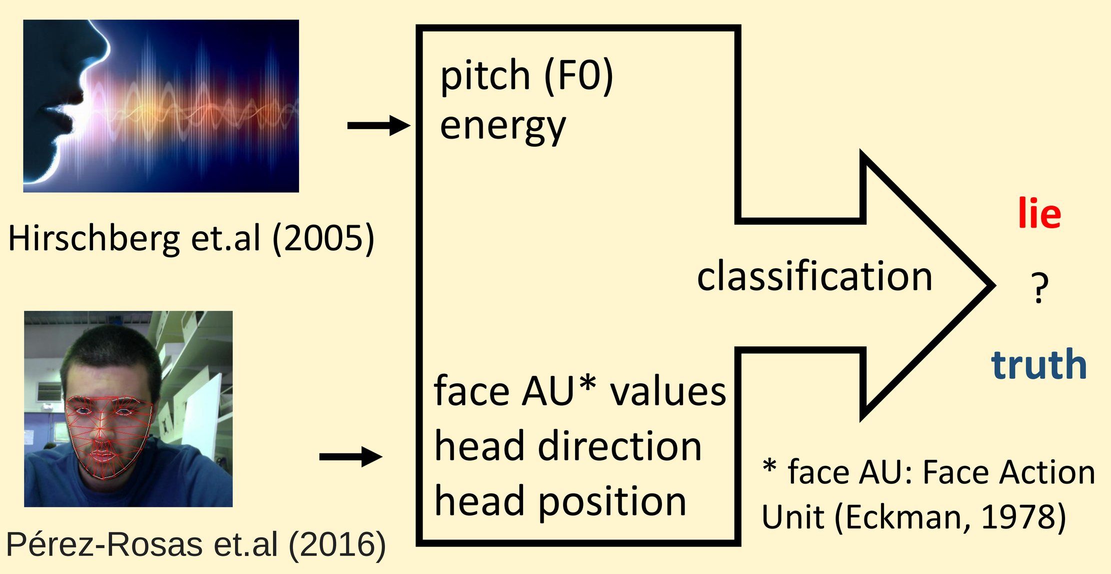
[2]PRESTO, JST, Japan

[3]Center for Advanced Intelligence Project (AIP), RIKEN, Japan

1. Overview

Background: An effective negotiation system needs to **know whether the other party (user) is lying or not** to choose the most appropriate response.

Deception detection:
- Classification human's spoken utterances into **lie** or **truth**.
- Current state-of-the-art models use **multimodal approach**



Problems: Current multimodal fusion methods cannot take full advantage of the rich multimodal information.
- Do not differentiate the abstraction level of information
- Complex and inefficient learning of features interaction

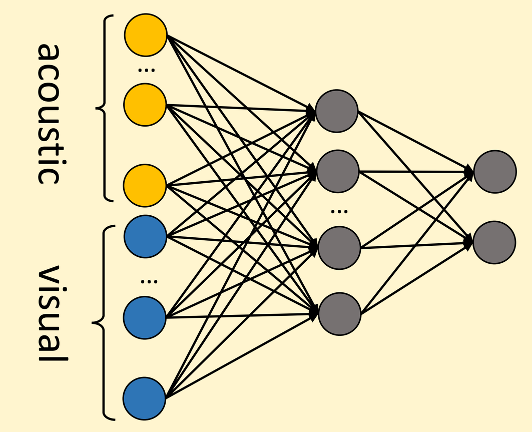
Our solution: Hierarchical tensor fusion network (Hierarchical TFN)
- Combination of hierarchical fusion (Tian et.al 2015) and tensor fusion (Zadeh et.al 2017)
- Balance the abstraction level and learning features interaction efficiently.

Results:
- **Proposed fusion method outperforms the others by more than 4%.**
- **Achieves highest DA selection accuracy** when using labels from Hierarchical-TFN-based deception detector.

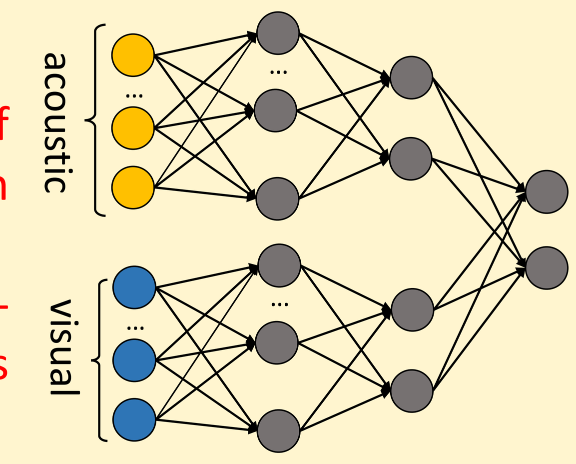
2. Problems: basic fusion methods

Multimodal fusion methods used in current multimodal deception detection works.

Early fusion:
- No distinction of modality abstraction level ☹
- Entangle the learning of intra-modality and inter-modality interactions ☹

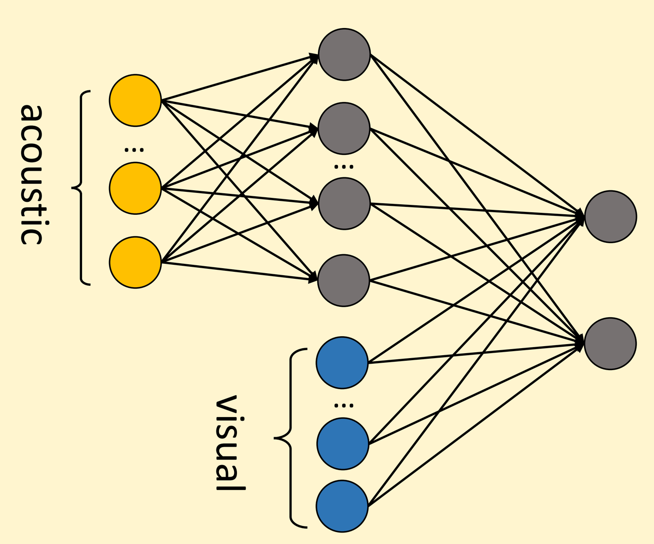


Late fusion:
- No distinction of modality abstraction level ☹
- Cannot learns inter-modality interactions ☹

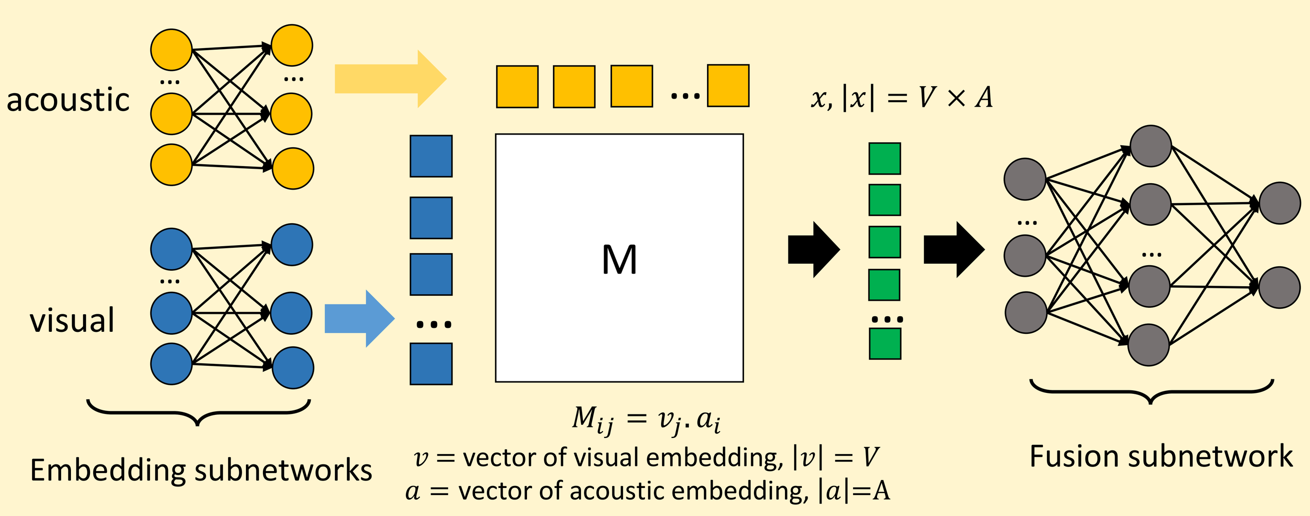


2. Problems: Advanced fusion methods

Hierarchical fusion:
- Can balance modality abstraction level ☺
- Entangle the learning of intra-modality and inter-modality interactions ☹



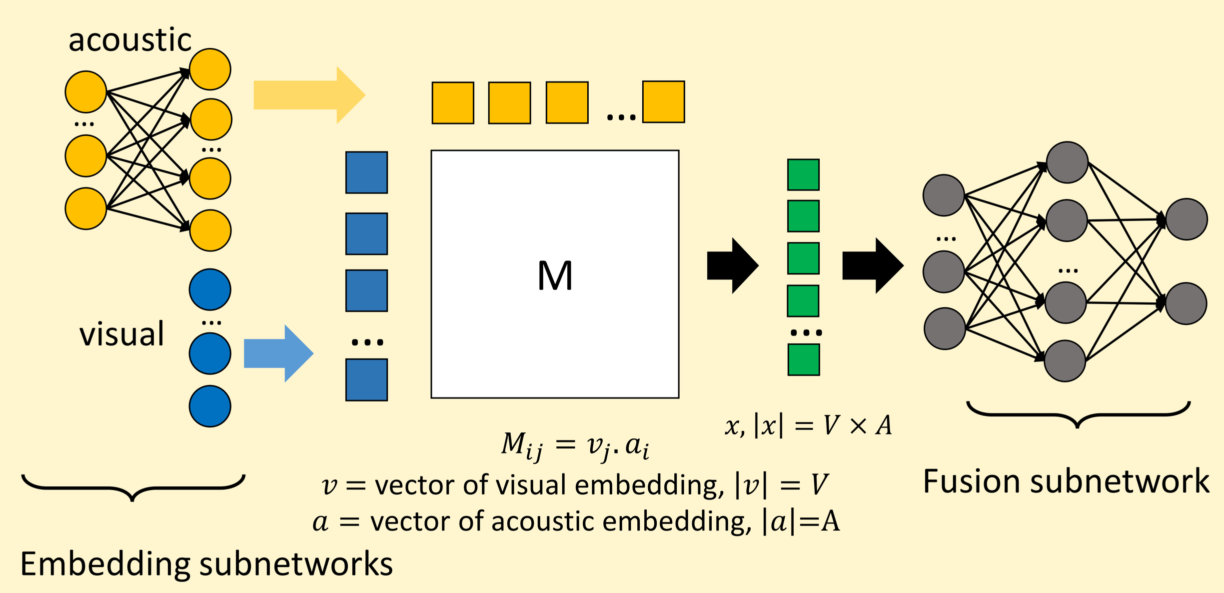
Tensor fusion (TFN):



- Separate learning of intra-modality interactions (**embedding subnetwork**) and inter-modality interactions (**fusion subnetwork**) ☺
- Cannot balance modality abstraction level ☹

3. Proposed fusion method

Hierarchical tensor fusion (Hierarchical TFN):



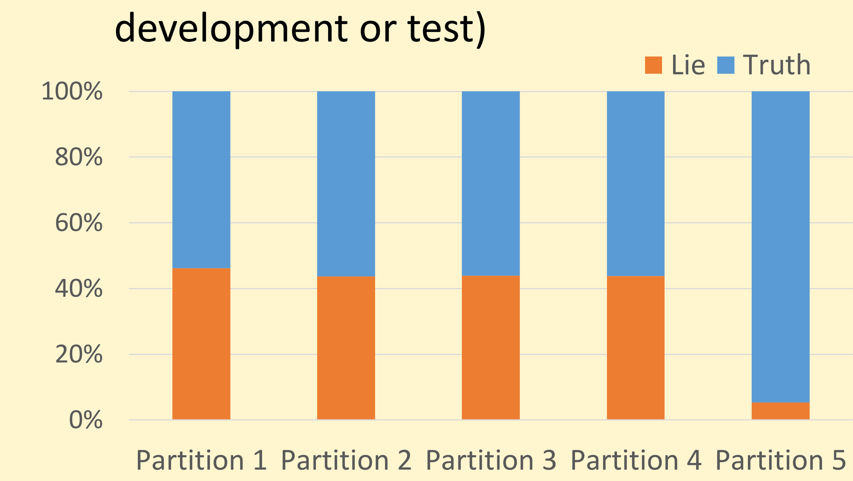
Advantages of hierarchical tensor fusion:
- ✓ Balance the abstraction level of different modalities.
- ✓ Separate learning of intra-modality and inter-modality interactions.
- ✓ Forcing the network to learn useful intra-modality interactions from certain modalities.
- ✓ Prevent learning of unimportant interactions, reduce unnecessary parameters and make network structure simpler.

4. Experiment #1 Deception detection

Dataset:
- Real-life trial (Rosas et.al 2015): recordings from court trials, 245 (105/140; deceptive/truthful)
- Simulated health consultation (Tung et.al 2018): 1021 (177/844)
- Total: 1266 (282/984)

Features extraction:
- Visual: Face Action Units, using (Baltrusaitis et.al 2016)
- Acoustic: IS_09 emotion acoustic features set, (Eyben et.al 2010)

Experiment setup:
- 4-fold cross-validation
- Utterances from same recording belong to same set (train, development or test)



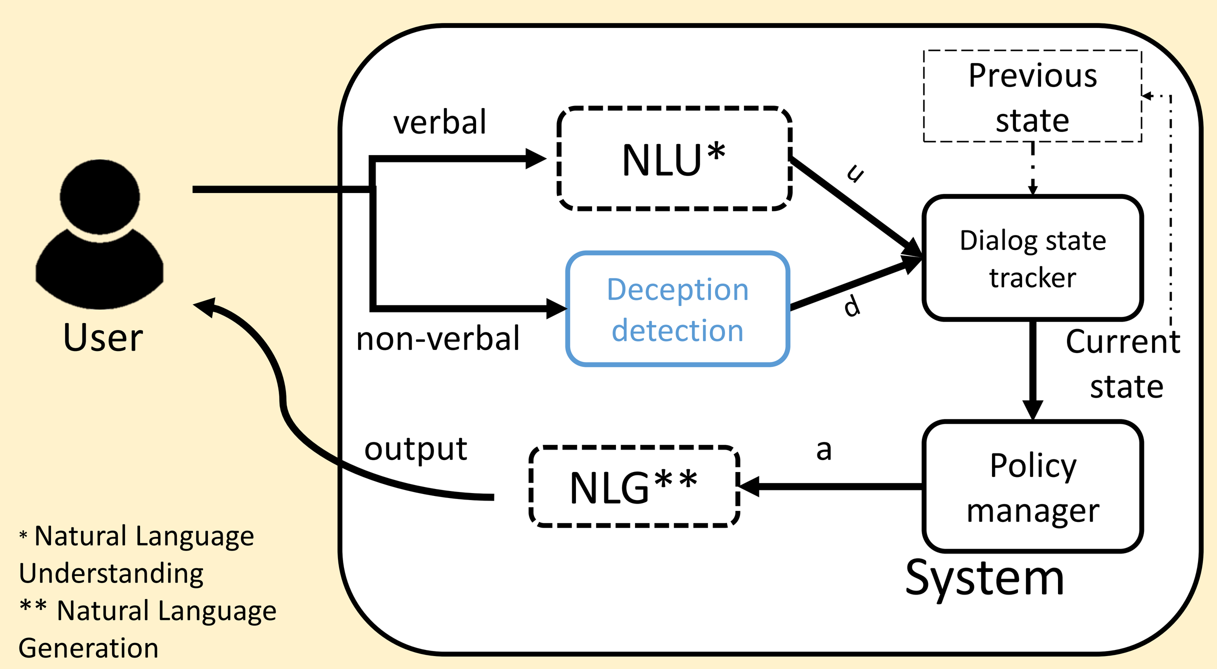
Experimental results

Model	Accuracy	Precision	Recall	F1-score
Single acoustic	53.78%	0.475	0.500	0.487
Single visual	49.28%	0.409	0.353	0.388
Multi early	53.42%	0.460	0.357	0.402
Multi late	54.68%	0.479	0.381	0.425
Multi hierarchical	53.78%	0.473	0.471	0.472
Multi TFN	50.36%	0.421	0.353	0.384
Multi hierarchical TFN	**58.63%**	**0.530**	**0.500**	**0.515**

- Precision, recall, and F1-score are measured for deceptive label (positive).
- Single visual model performance is much worse than single visual acoustic
- The Hierarchical TFN outperforms all other methods significantly.

4. Experiment #2: Negotiation System's dialog management

Negotiation system's dialog management



Dialog modeling:
- Model the dialog management process using **Partially Observable Markov Decision Process (POMDP).**
- Dialog state: $s = (u, d)$ - u: user's dialog act, d: user's deception.
- State transition: $P(u^{t+1}, d^{t+1} | u^t, d^t, \hat{a}^t) = \underbrace{P(u^{t+1} | d^{t+1}, u^t, d^t, \hat{a}^t)}_{intention\ model} \underbrace{P(d^{t+1} | d^t, \hat{a}^t)}_{deception\ model}$
- Train the dialog management using **reinforcement learning**:

$$Q(s^t, a^t) = (1 - \alpha) Q(s^t, a^t) + \alpha \left(r^t + \gamma \max_{a^{t+1}} Q(s^{t+1}, a^{t+1}) \right)$$

Experimental results

Deception labels used for dialog management	System DA selection accuracy
Chance rate deception	65.69%
Gold-label deception	80.31%
Single visual prediction	70.15%
Single acoustic prediction	66.22%
Multi early prediction	66.48%
Multi late prediction	68.58%
Multi hierarchy prediction	69.10%
Multi TFN prediction	69.66%
Multi Hierarchical TFN prediction	**71.20%**

- Human expert selects best reaction in each dialog turn (based on annotated user's action and user's deception)
- Compare system's choice with human choice for each dialog turn.
- Highest accuracy of DA selection achieved when using labels predicted by Hierarchical TFN deception detection model.

5. Discussion

- Collect/augment more multimodal deception data for evaluation on a larger scale
- Applied this fusion methods for other multimodal processing tasks: emotion or sentiment analysis