# Multi-turn Response Selection

Jimblin

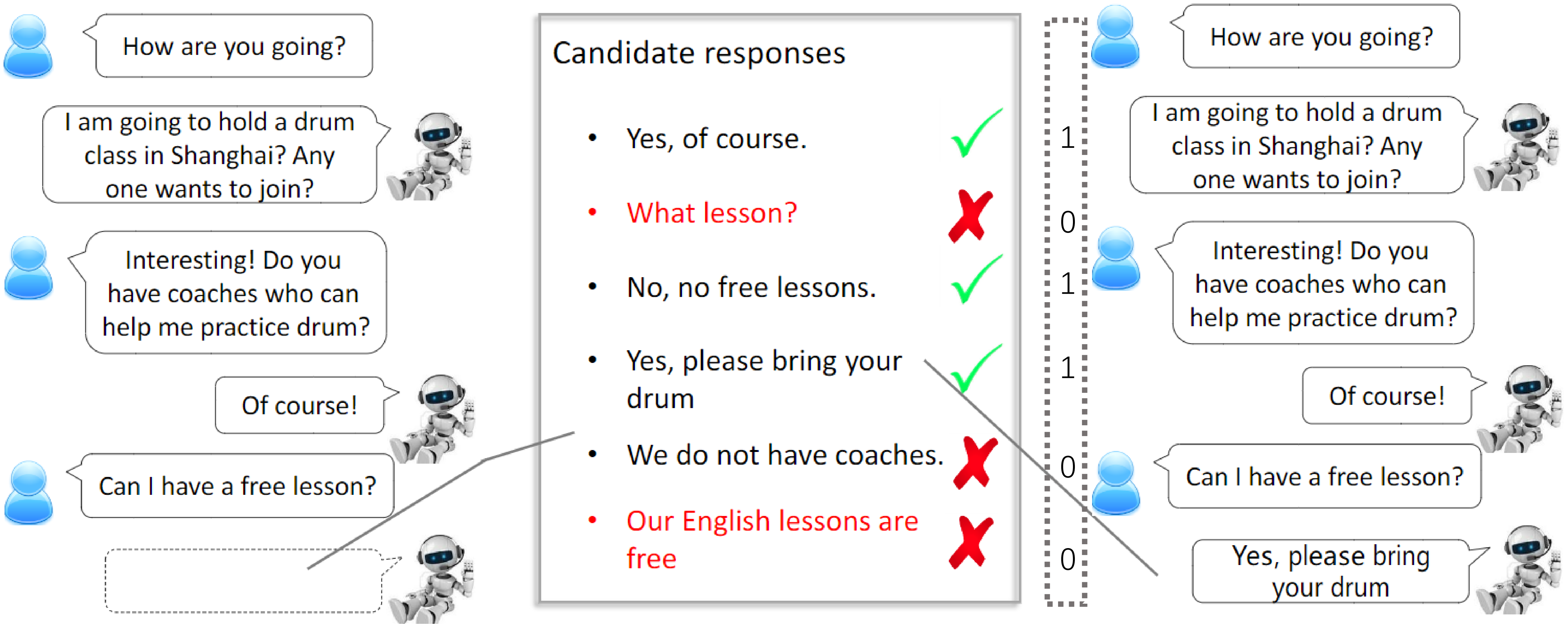
Tencent AI Lab, NLP Center

# Overview

- Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network（*ACL2018*）
- One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues （*ACL2019*）
- Constructing Interpretive Spatio-Temporal Features for Multi-Turn Response Selection（*ACL2019*）

# Problem Formalization

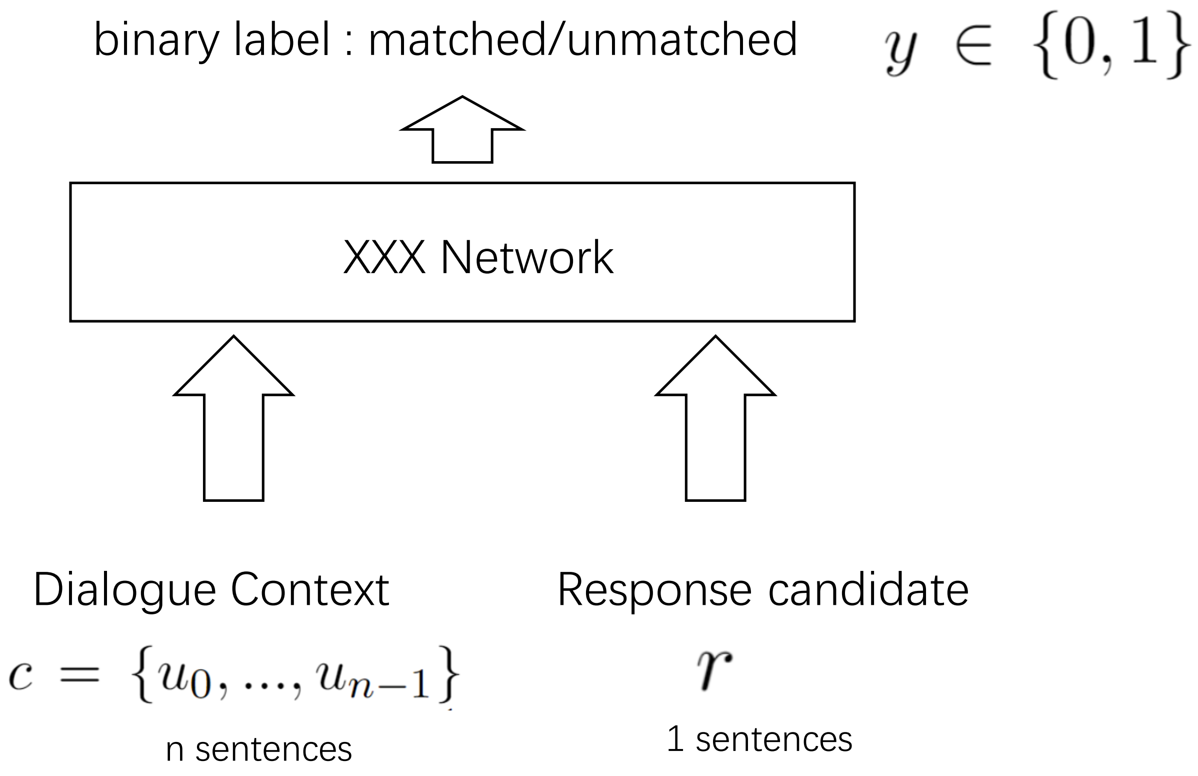
Fig from Deep Chit-Chat: Deep: Learning for ChatBots. Wei Wu and Rui Yan. EMNLP 2018




Candidates in red are good without the context!

Context $c = \{u_0, \ldots, u_{n-1}\}$ Response candidate $r$ Label $y \in \{0, 1\}$

# Problem Formalization




If there are m response candidates, the network need to

$$\text{data set } \mathcal{D} = \{(c, r, y)_Z\}_{Z=1}^{N}$$

# Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network

**Xiangyang Zhou**[*], **Lu Li**[*], **Daxiang Dong**, **Yi Liu**, **Ying Chen**,
**Wayne Xin Zhao**[†], **Dianhai Yu** and **Hua Wu**
Baidu Inc., Beijing, China
{zhouxiangyang, lilu12, dongdaxiang, liuyi05, chenying04, v_zhaoxin, yudianhai, wu_hua}@baidu.com

# Motivation & Contribution

- Existing models only consider the textual relevance, which suffers from matching response that latently depends on previous turns.
- RNN-based network is  too costly to use for capturing semantic representations.

- The authors jointly introduce self-attention and cross-attention in one uniform neural matching network.

# Methodology

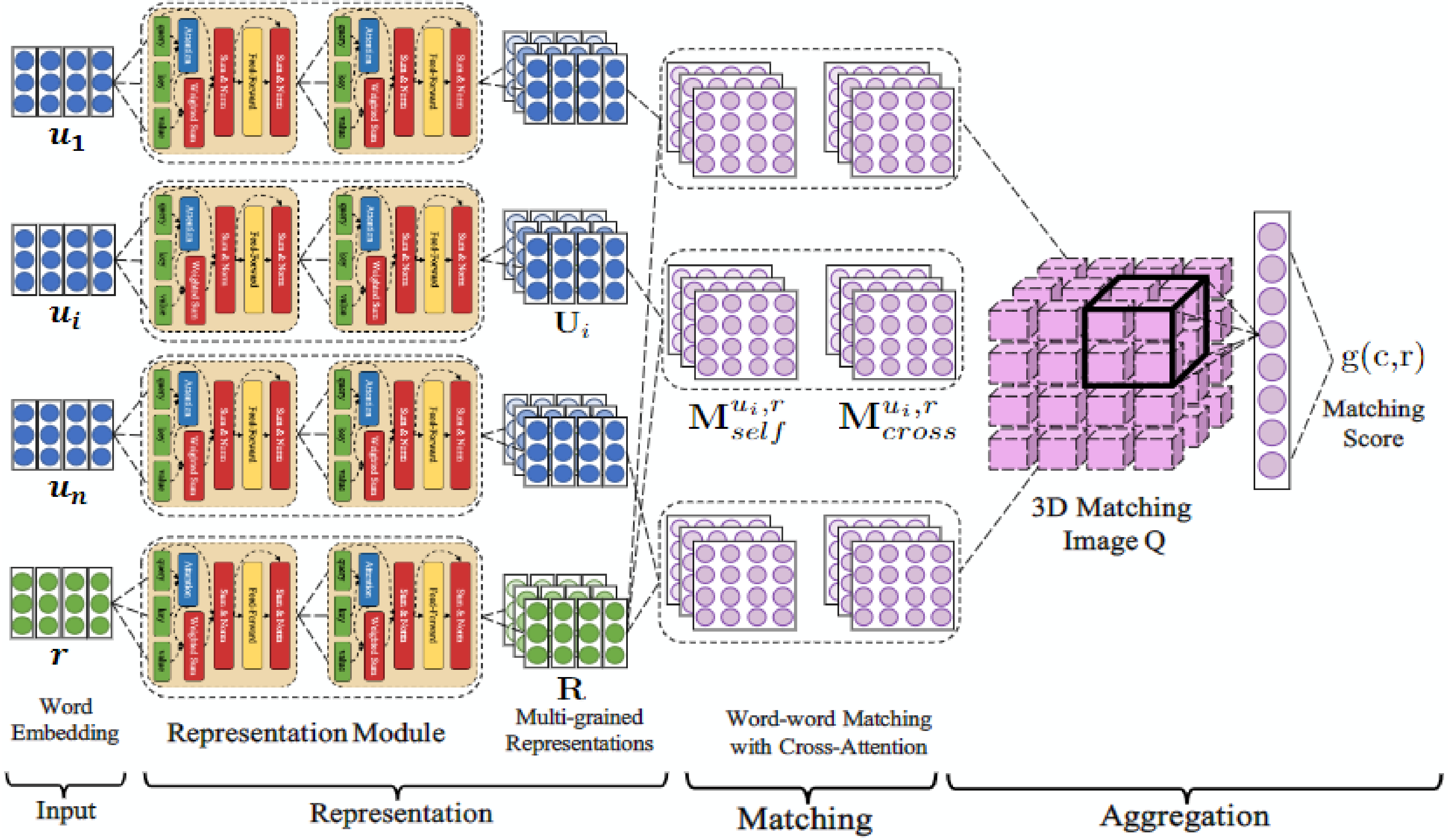



Figure 2: Overview of Deep Attention Matching Network.

# Methodology

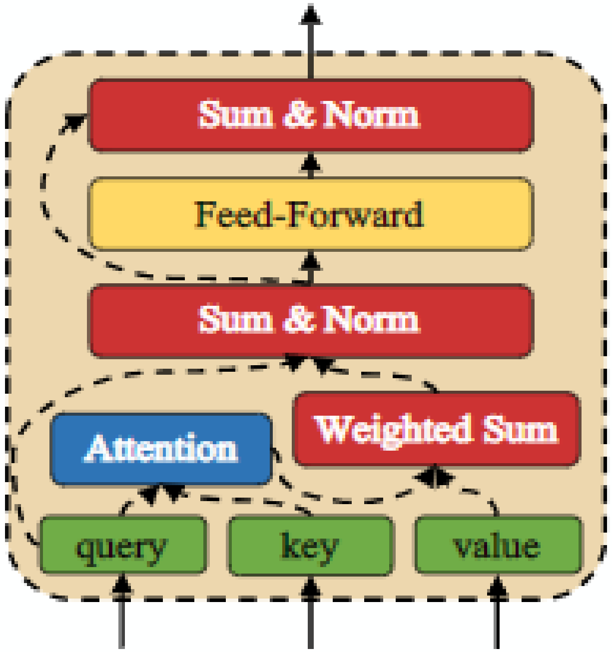



Figure 3: Attentive Module.

$$Att(\mathcal{Q},\mathcal{K}) = \left[softmax\left(\frac{\mathcal{Q}[i]\cdot\mathcal{K}^T}{\sqrt{d}}\right)\right]_{i=0}^{n_{\mathcal{Q}}-1} \quad (1)$$

$$\mathcal{V}_{att} = Att(\mathcal{Q},\mathcal{K})\cdot\mathcal{V} \in \mathbb{R}^{n_{\mathcal{Q}}\times d} \quad (2)$$

$$\widetilde{\mathbf{U}}_i^l = \mathbf{AttentiveModule}(\mathbf{U}_i^l, \mathbf{R}^l, \mathbf{R}^l) \quad (8)$$

$$\widetilde{\mathbf{R}}^l = \mathbf{AttentiveModule}(\mathbf{R}^l, \mathbf{U}_i^l, \mathbf{U}_i^l) \quad (9)$$

$$\mathbf{M}_{cross}^{u_i,r,l} = \{\widetilde{\mathbf{U}}_i^l[k]^T \cdot \widetilde{\mathbf{R}}^l[t]\}_{n_{u_i}\times n_r} \quad (10)$$

# Experiment

| | Ubuntu Corpus | | | | Douban Conversation Corpus | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $R_2$@1 | $R_{10}$@1 | $R_{10}$@2 | $R_{10}$@5 | MAP | MRR | P@1 | $R_{10}$@1 | $R_{10}$@2 | $R_{10}$@5 |
| $\text{DualEncoder}_{lstm}$ | 0.901 | 0.638 | 0.784 | 0.949 | 0.485 | 0.527 | 0.320 | 0.187 | 0.343 | 0.720 |
| $\text{DualEncoder}_{bilstm}$ | 0.895 | 0.630 | 0.780 | 0.944 | 0.479 | 0.514 | 0.313 | 0.184 | 0.330 | 0.716 |
| MV-LSTM | 0.906 | 0.653 | 0.804 | 0.946 | 0.498 | 0.538 | 0.348 | 0.202 | 0.351 | 0.710 |
| Match-LSTM | 0.904 | 0.653 | 0.799 | 0.944 | 0.500 | 0.537 | 0.345 | 0.202 | 0.348 | 0.720 |
| Multiview | 0.908 | 0.662 | 0.801 | 0.951 | 0.505 | 0.543 | 0.342 | 0.202 | 0.350 | 0.729 |
| DL2R | 0.899 | 0.626 | 0.783 | 0.944 | 0.488 | 0.527 | 0.330 | 0.193 | 0.342 | 0.705 |
| $\text{SMN}_{dynamic}$ | 0.926 | 0.726 | 0.847 | 0.961 | 0.529 | 0.569 | 0.397 | 0.233 | 0.396 | 0.724 |
| DAM | **0.938** | **0.767** | **0.874** | **0.969** | **0.550** | **0.601** | **0.427** | **0.254** | **0.410** | **0.757** |
| $\text{DAM}_{first}$ | 0.927 | 0.736 | 0.854 | 0.962 | 0.528 | 0.579 | 0.400 | 0.229 | 0.396 | 0.741 |
| $\text{DAM}_{last}$ | 0.932 | 0.752 | 0.861 | 0.965 | 0.539 | 0.583 | 0.408 | 0.242 | 0.407 | 0.748 |
| $\text{DAM}_{self}$ | 0.931 | 0.741 | 0.859 | 0.964 | 0.527 | 0.574 | 0.382 | 0.221 | 0.403 | 0.750 |
| $\text{DAM}_{cross}$ | 0.932 | 0.749 | 0.863 | 0.966 | 0.535 | 0.585 | 0.400 | 0.234 | 0.411 | 0.733 |

Table 1: Experimental results of DAM and other comparison approaches on Ubuntu Corpus V1 and Douban Conversation Corpus.

# Conclusion

- Using stacked self-attention to harvest multi-grained semantic representations.
- Utilizing cross-attention to match with dependency information.

- DAM is a fast network which achieves the SOTA result.

# One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues

**Chongyang Tao**[1], **Wei Wu**[2], **Can Xu**[2], **Wenpeng Hu**[1], **Dongyan Zhao**[1,3] and **Rui Yan**[1,3*]
[1]Institute of Computer Science and Technology, Peking University, Beijing, China
[2]Microsoft Corporation, Beijing, China
[3]Center for Data Science, Peking University, Beijing, China
[1,3]{chongyangtao,wenpeng.hu,zhaody,ruiyan}@pku.edu.cn
[2]{wuwei,caxu}@microsoft.com

# one-time interaction not enough

- Existing methods are executed ***in a rather shallow manner***.
- Matching between an utterance and a response candidate is determined ***only by one step*** of interaction on each type or each layer of representations.
- If a model extracts some matching information from utterance-response pairs in one step of interaction, then by stacking ***multiple interaction blocks*** , the matching network can capture ***the semantic relationship*** between a context and a response candidate in a more comprehensive way.

# Motivation & Contribution

- By stacking *multiple interaction blocks* , the matching network can capture the semantic relationship between a context and a response candidate in a more comprehensive way.

- This paper performs matching by stacking multiple interaction blocks, and thus extends the shallow interaction in state-of-the-art methods to a deep form.

# Methodology

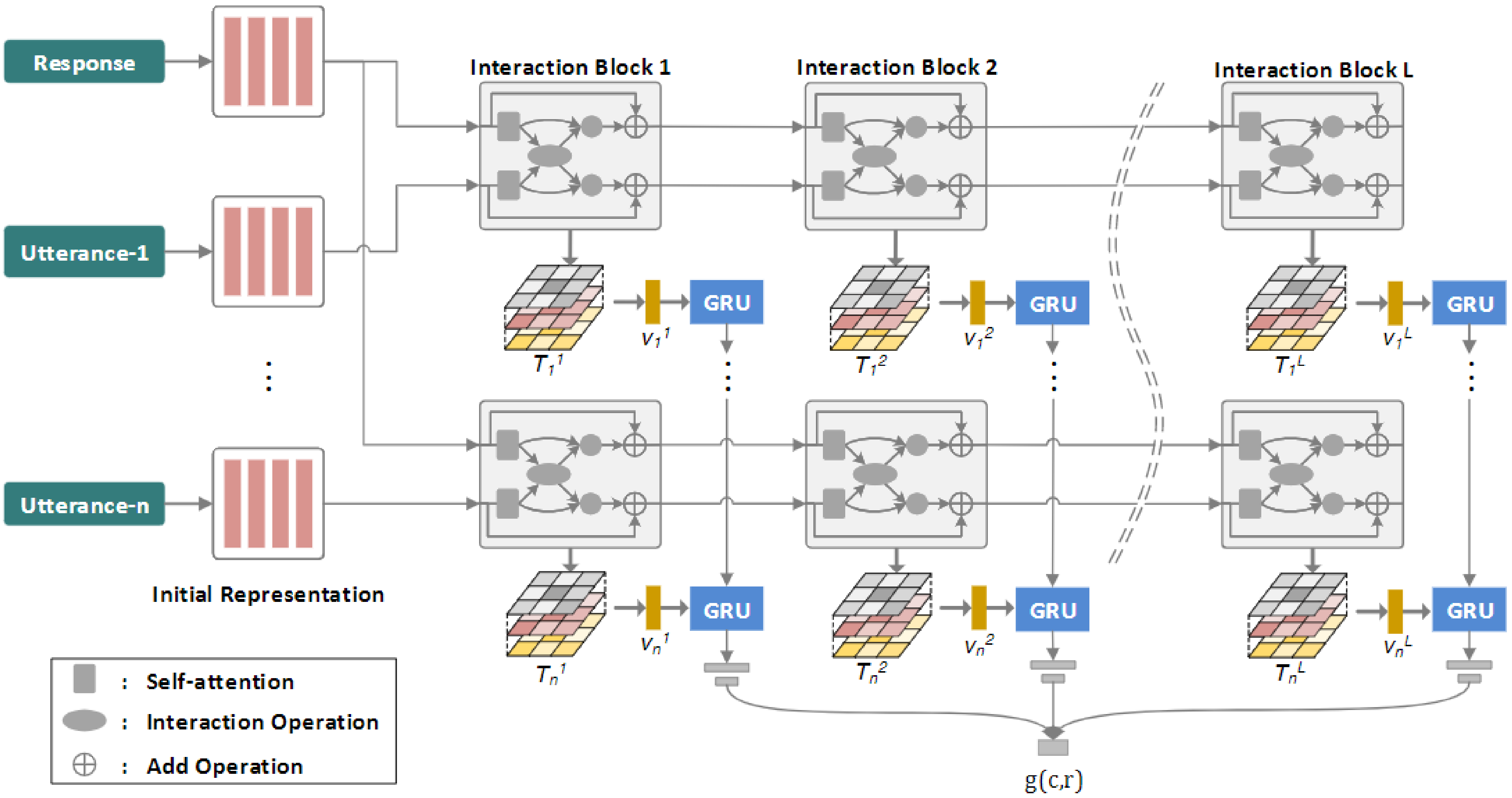



Figure 1: Architecture of interaction-over-interaction network.

# Interaction Block

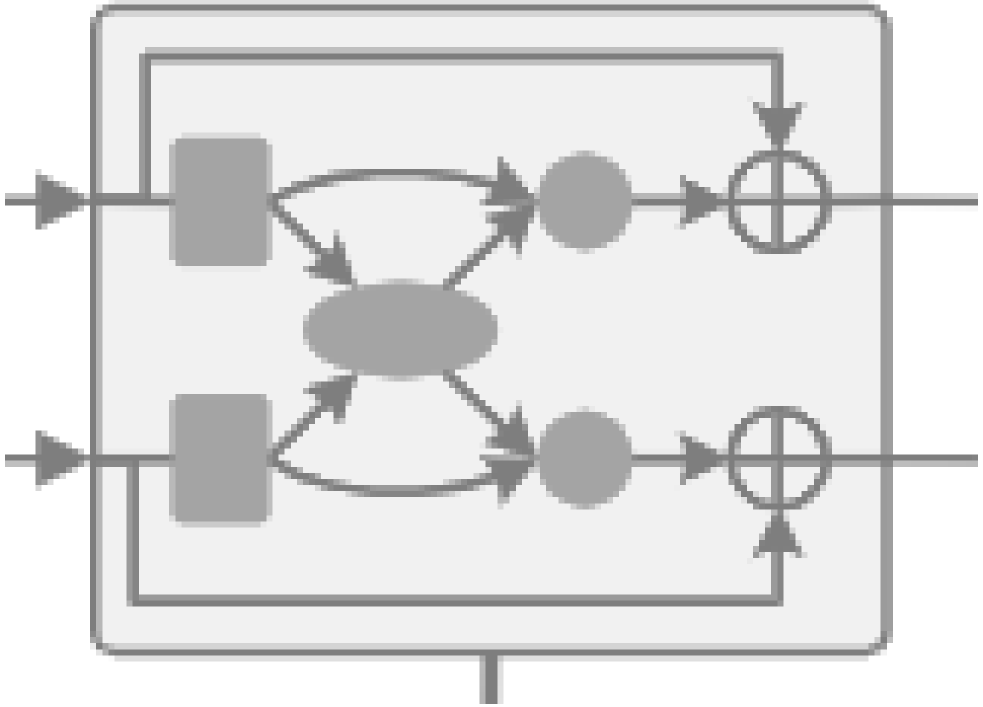
Self Attention Mechanism $f_{ATT}(\mathbf{Q}, \mathbf{K})$

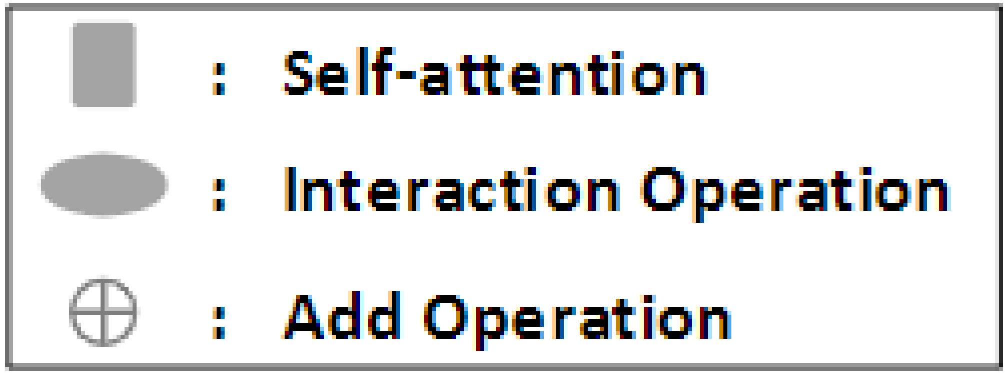
$$\hat{\mathbf{Q}} = S(\mathbf{Q}, \mathbf{K}) \cdot \mathbf{K}, \quad (1)$$

$$S(\mathbf{Q}, \mathbf{K}) = \mathrm{softmax}(f(\mathbf{Q}\mathbf{W})\mathbf{D}f(\mathbf{K}\mathbf{W})^{\top}). \quad (2)$$

$$\mathrm{ReLU}(\tilde{\mathbf{Q}}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \quad (3)$$

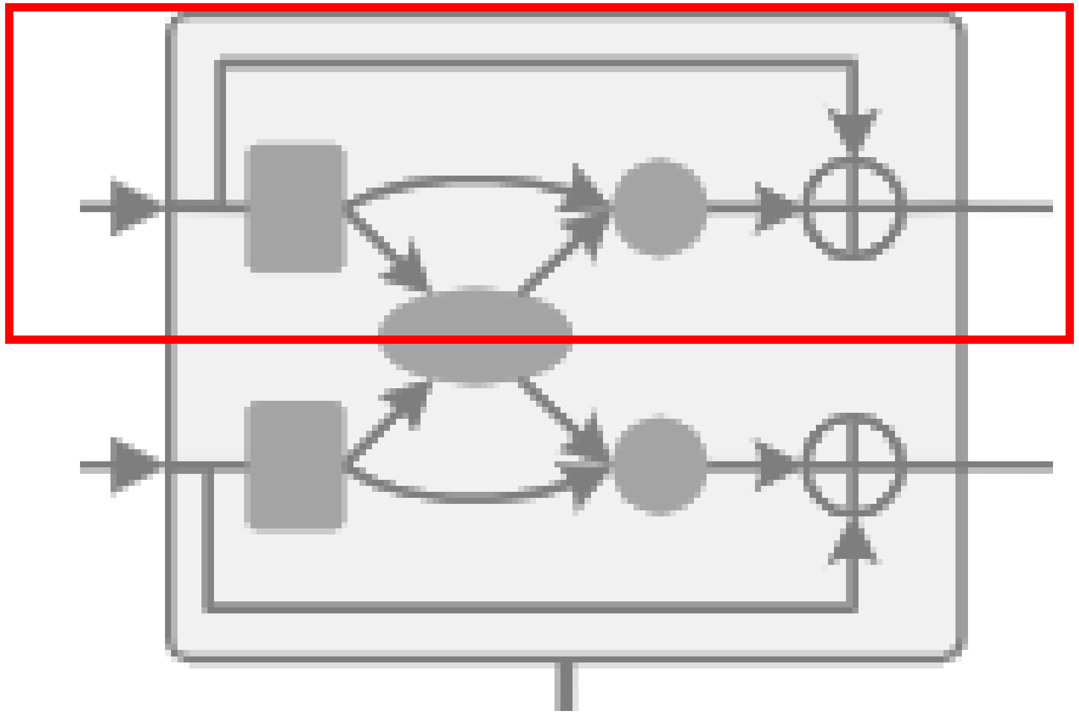
**Interaction Block 1**





# Interaction Block

Interaction Block 1



Utterance

Interaction Operation $f_{ATT}(\mathbf{Q}, \mathbf{K})$

$$\hat{\mathbf{U}}^k = f_{\mathrm{ATT}}(\mathbf{U}^{k-1}, \mathbf{U}^{k-1}), \tag{4}$$

$$\overline{\mathbf{U}}^k = f_{\mathrm{ATT}}(\mathbf{U}^{k-1}, \mathbf{R}^{k-1}), \tag{6}$$

$$\tilde{\mathbf{U}}^k = \mathbf{U}^{k-1} \odot \overline{\mathbf{U}}^k, \tag{8}$$

$$\mathbf{e}_{u,i}^k = \mathrm{ReLU}(\mathbf{w}_p \begin{bmatrix} \mathbf{e}_{u,i}^{k-1} \\ \hat{\mathbf{e}}_{u,i}^k \\ \overline{\mathbf{e}}_{u,i}^k \\ \tilde{\mathbf{e}}_{u,i}^k \end{bmatrix} + \mathbf{b}_p) + \mathbf{e}_{u,i}^{k-1}, \tag{10}$$

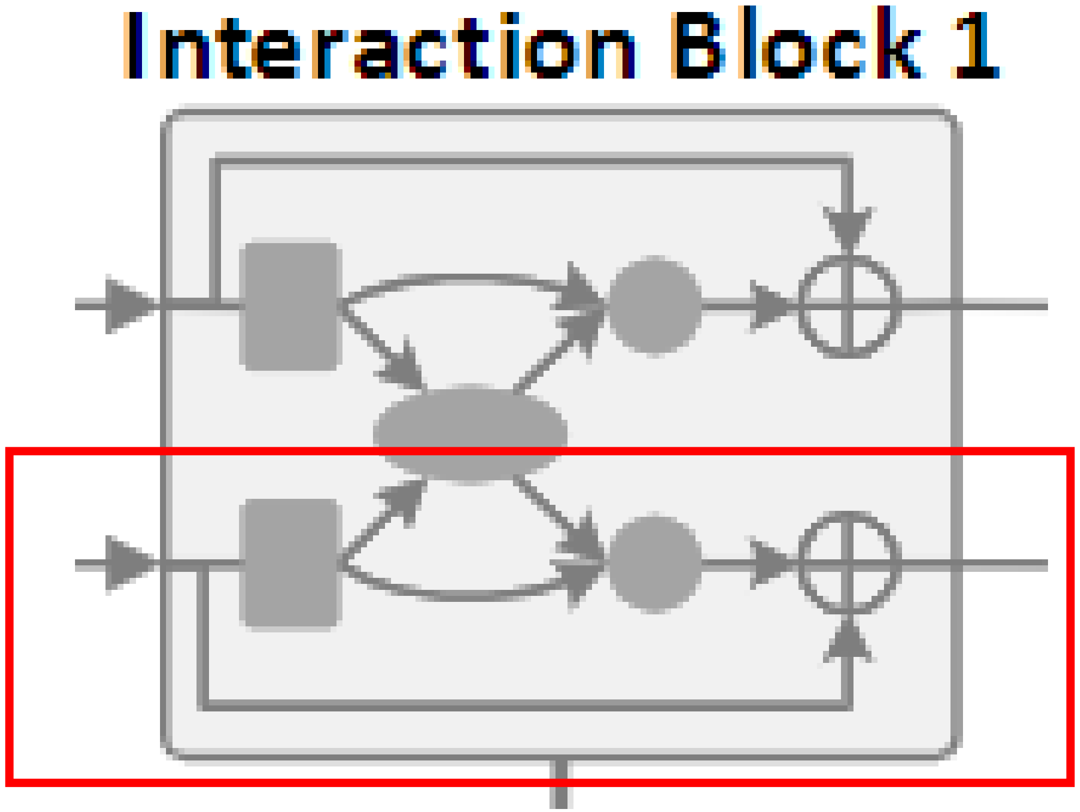
# Interaction Block

Interaction Operation $f_{ATT}(\mathbf{Q}, \mathbf{K})$

$$\hat{\mathbf{R}}^k = f_{\text{ATT}}(\mathbf{R}^{k-1}, \mathbf{R}^{k-1}). \quad (5)$$

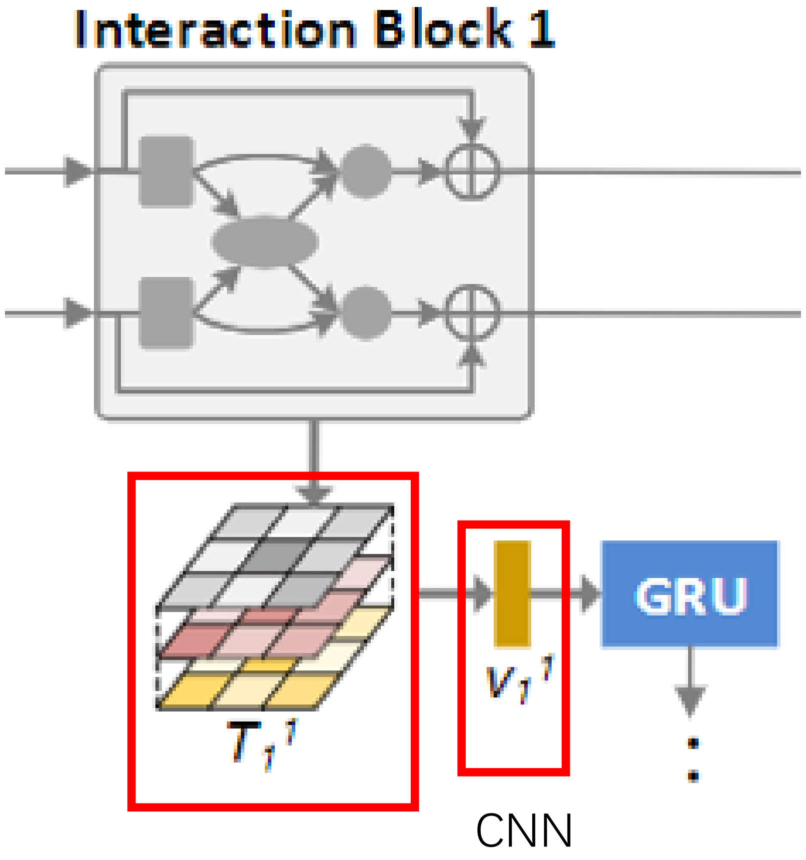
$$\overline{\mathbf{R}}^k = f_{\text{ATT}}(\mathbf{R}^{k-1}, \mathbf{U}^{k-1}). \quad (7)$$

$$\tilde{\mathbf{R}}^k = \mathbf{R}^{k-1} \odot \overline{\mathbf{R}}^k, \quad (9)$$




$$\mathbf{e}_{r,i}^k = \text{ReLU}(\mathbf{w}_p \begin{bmatrix} \mathbf{e}_{r,i}^{k-1} \\ \hat{\mathbf{e}}_{r,i}^k \\ \overline{\mathbf{e}}_{r,i}^k \\ \tilde{\mathbf{e}}_{r,i}^k \end{bmatrix} + \mathbf{b}_p) + \mathbf{e}_{r,i}^{k-1}, \quad (11)$$

# Matching Aggregation



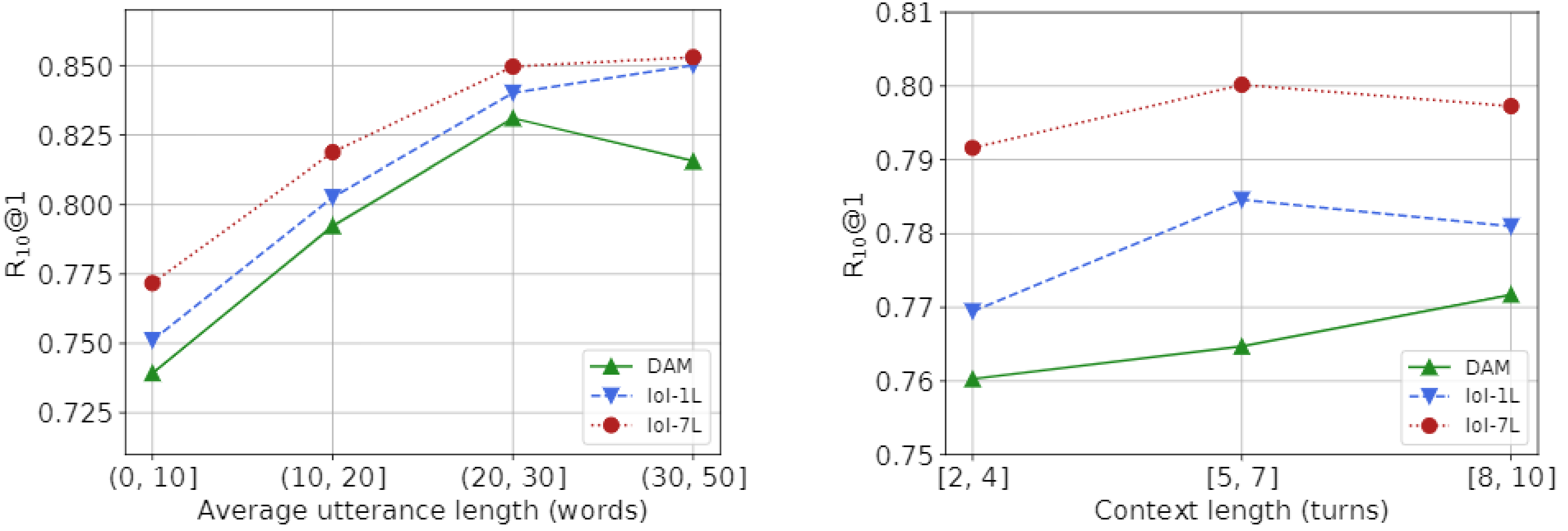

$$\mathbf{M}_{i,1}^{k} = \frac{\mathbf{U}_i^{k-1} \cdot (\mathbf{R}^{k-1})^\top}{\sqrt{d}},$$
$$\mathbf{M}_{i,2}^{k} = \frac{\hat{\mathbf{U}}_i^{k} \cdot (\hat{\mathbf{R}}^{k})^\top}{\sqrt{d}}, \tag{12}$$
$$\mathbf{M}_{i,3}^{k} = \frac{\overline{\mathbf{U}}_i^{k} \cdot (\overline{\mathbf{R}}^{k})^\top}{\sqrt{d}},$$

$$\mathbf{T}_i^k = \mathbf{M}_{i,1}^{k} \oplus \mathbf{M}_{i,2}^{k} \oplus \mathbf{M}_{i,3}^{k}, \tag{13}$$

# Experiment

| Models \ Metrics | Ubuntu Corpus | | | | Douban Corpus | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $R_2@1$ | $R_{10}@1$ | $R_{10}@2$ | $R_{10}@5$ | MAP | MRR | P@1 | $R_{10}@1$ | $R_{10}@2$ | $R_{10}@5$ |
| RNN (Lowe et al., 2015) | 0.768 | 0.403 | 0.547 | 0.819 | 0.390 | 0.422 | 0.208 | 0.118 | 0.223 | 0.589 |
| CNN (Lowe et al., 2015) | 0.848 | 0.549 | 0.684 | 0.896 | 0.417 | 0.440 | 0.226 | 0.121 | 0.252 | 0.647 |
| LSTM (Lowe et al., 2015) | 0.901 | 0.638 | 0.784 | 0.949 | 0.485 | 0.527 | 0.320 | 0.187 | 0.343 | 0.720 |
| BiLSTM (Kadlec et al., 2015) | 0.895 | 0.630 | 0.780 | 0.944 | 0.479 | 0.514 | 0.313 | 0.184 | 0.330 | 0.716 |
| DL2R (Yan et al., 2016) | 0.899 | 0.626 | 0.783 | 0.944 | 0.488 | 0.527 | 0.330 | 0.193 | 0.342 | 0.705 |
| MV-LSTM (Wan et al., 2016) | 0.906 | 0.653 | 0.804 | 0.946 | 0.498 | 0.538 | 0.348 | 0.202 | 0.351 | 0.710 |
| Match-LSTM (Wang and Jiang, 2016) | 0.904 | 0.653 | 0.799 | 0.944 | 0.500 | 0.537 | 0.345 | 0.202 | 0.348 | 0.720 |
| Multi-View (Zhou et al., 2016) | 0.908 | 0.662 | 0.801 | 0.951 | 0.505 | 0.543 | 0.342 | 0.202 | 0.350 | 0.729 |
| SMN (Wu et al., 2017) | 0.926 | 0.726 | 0.847 | 0.961 | 0.529 | 0.569 | 0.397 | 0.233 | 0.396 | 0.724 |
| DUA(Zhang et al., 2018b) | - | 0.752 | 0.868 | 0.962 | 0.551 | 0.599 | 0.421 | 0.243 | 0.421 | 0.780 |
| DAM (Zhou et al., 2018b) | 0.938 | 0.767 | 0.874 | 0.969 | 0.550 | 0.601 | 0.427 | 0.254 | 0.410 | 0.757 |
| IoI-global | **0.941** | **0.778** | **0.879** | 0.970 | **0.566** | 0.608 | 0.433 | 0.263 | **0.436** | **0.781** |
| IoI-local | **0.947** | **0.796** | **0.894** | **0.974** | **0.573** | **0.621** | 0.444 | **0.269** | **0.451** | **0.786** |

# Experiment




(a) $R_{10}$@1 vs. Average utterance length

(b) $R_{10}$@1 vs. Number of turns

Figure 3: Performance of IoI across contexts with different lengths on the Ubuntu data.

# Summary

- ***Strength***
  - We present an interaction-over-interaction network (IoI) that lets utterance-response interaction in context-response matching go deep.
  - A good example of stacked network using self-attention, cnn and gru. (Like GoogleNet)
- ***Weakness***
  - The params of IoI is futher larger than baseline model, so the improvement is uncertain.

# Constructing Interpretive Spatio-Temporal Features for Multi-Turn Responses Selection

**Junyu Lu[†], Chenbin Zhang[†], Zeying Xie, Guang Ling, Chao Zhou, Zenglin Xu[†]**
† SMILE Lab, University of Electronic Science and Technology of China, Sichuan, China
{cs.junyu, aleczhang13, swpdtz, zacharyling}@gmail.com,
tom.chaozhou@foxmail.com, zenglin@gmail.com

# Motivation

- In multi-turn dialogues, the next sentence is generally based on what was presented before and tends to match a recent local context.
- This is because the topic in a conversation may change over time, and the effective matching between the dialogue may only appear in a local time period.
- This phenomena generally appear in video processing. Therefore, Each turn of dialogue can be regarded as a frame of a video.

# Contribution

- The first work that representations of the dialogue context and candidate answers are learned through from dual encoders, and deep 3D ConvNets.
- The Spatio-Temporal Matching block (STM) models local semantic relation between each turn of dialog and candidates by soft-attention mechanism.

# Methodology

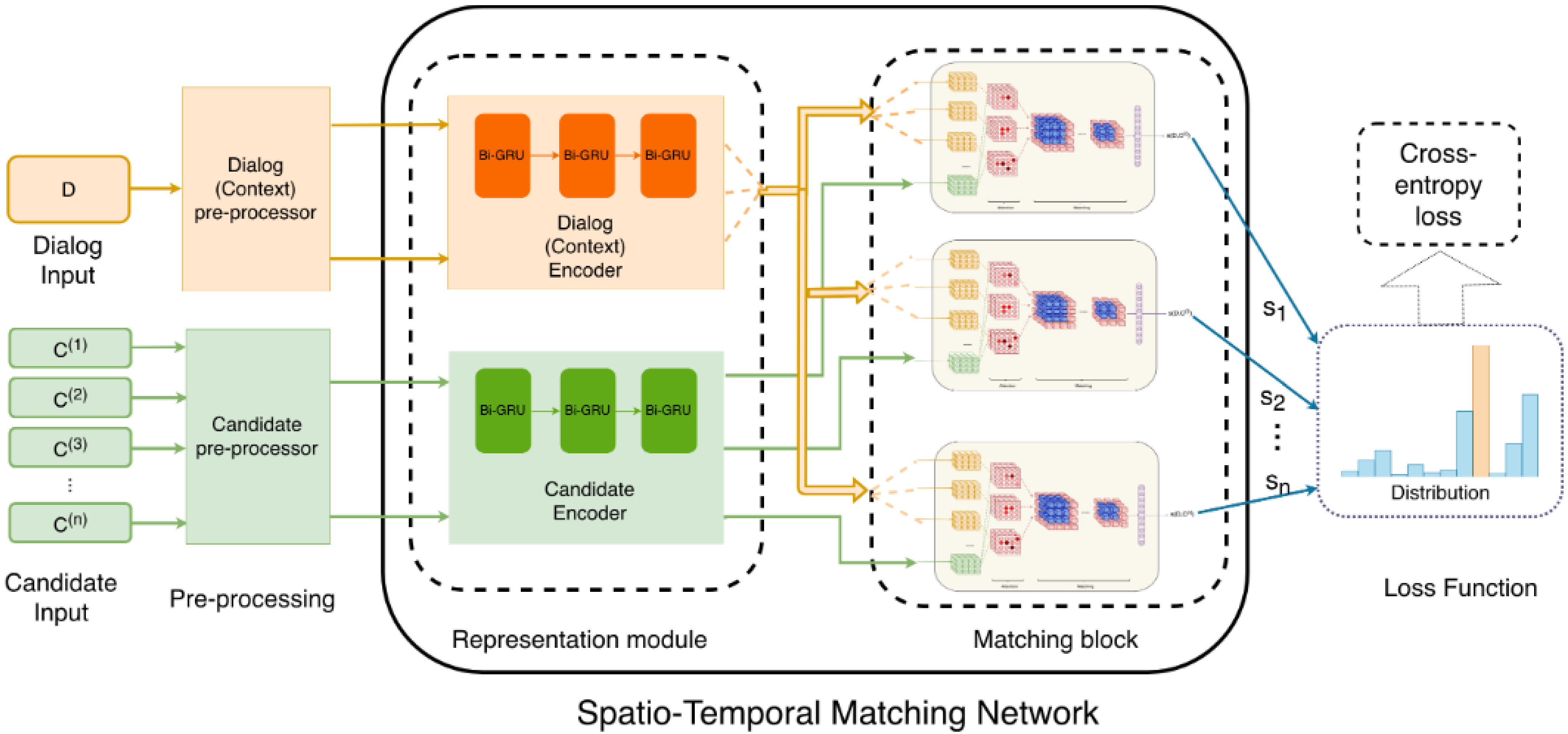



Figure 2: The proposed spatio-temporal matching framework for response selection.

# Spatio-Temporal Matching block

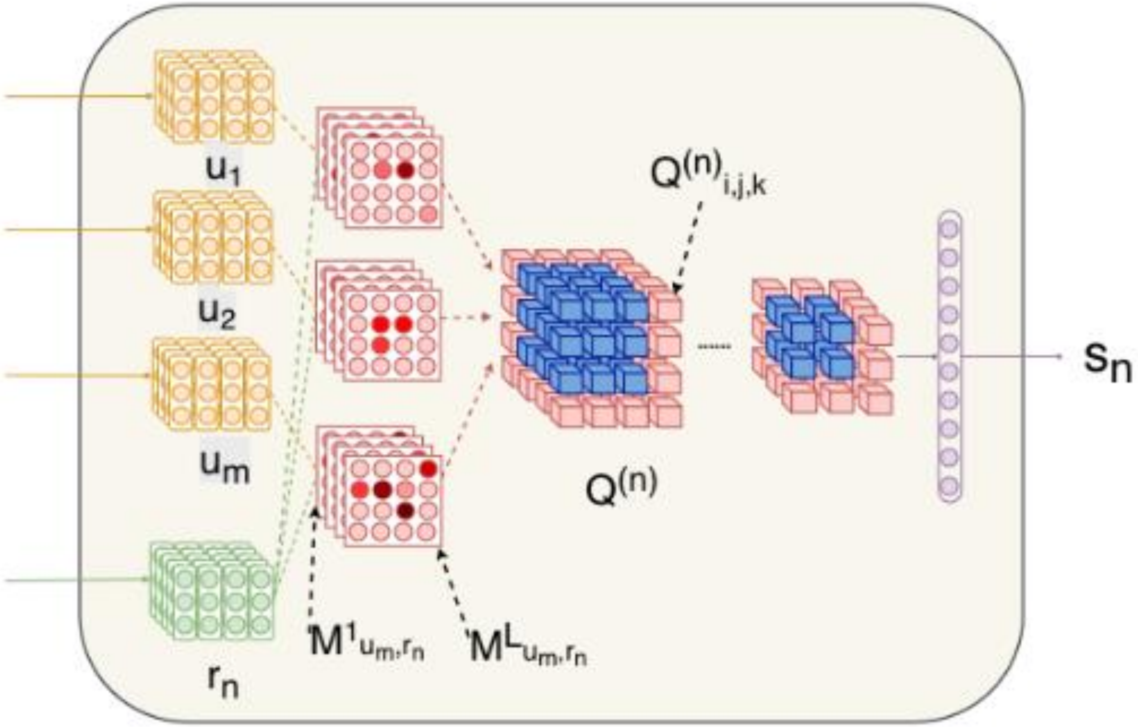



Figure 3: A close-up of the matching block

$$\mathbf{M}^{l}_{\mu_m,\gamma_n} = \frac{(\mu^l_m)^T \gamma^l_n}{\sqrt{d}}, \qquad (1)$$

$$\mathbf{Q^{(n)}} = \{Q^{(n)}_{i,j,k}\}_{m \times n_\mu \times n_\gamma}, \qquad (2)$$

$$Q^{(n)}_{i,j,k} = \{M^l_{\mu_i,\gamma_n}[j,k]\}^L_{l=0}, \qquad (3)$$

$$s_n = \mathbf{W} f_{conv}(\mathbf{Q^{(n)}}) + \mathbf{b}, \qquad (4)$$

# Experiment

| Model | $R_{100}$@1 | $R_{100}$@10 | MRR |
|---|---|---|---|
| Baseline | 0.083 | 0.359 | - |
| DAM | 0.347 | 0.663 | 0.356 |
| DAM+Fine-tune | 0.364 | 0.664 | 0.443 |
| DME | 0.383 | 0.725 | 0.498 |
| DME-SMN | 0.455 | 0.761 | 0.558 |
| STM(Transform) | **0.490** | **0.764** | **0.588** |
| STM(GRU) | **0.503** | **0.783** | **0.597** |
| STM(Ensemble) | **0.521** | **0.797** | **0.616*** |
| STM(BERT) | **0.548*** | **0.827*** | **0.614** |

Table 1: Experiment Result on the Ubuntu Corpus.

| Model | Advising 1 | | Advising 2 | |
|---|---|---|---|---|
| | $R_{100}$@10 | MRR | $R_{100}$@10 | MRR |
| Baseline | 0.296 | - | - | - |
| DAM | 0.603 | 0.312 | 0.374 | 0.174 |
| DAM+Fine-tune | 0.622 | 0.333 | 0.416 | 0.192 |
| DME | 0.420 | 0.215 | 0.304 | 0.142 |
| DME-SMN | 0.570 | 0.335 | 0.388 | 0.183 |
| STM(Transform) | 0.590 | 0.320 | 0.404 | 0.182 |
| STM(GRU) | **0.654** | **0.380** | **0.466** | **0.220** |
| STM(Ensemble) | **0.662*** | **0.385*** | **0.502*** | **0.232*** |

Table 2: Experiment Results on the Advising Dataset.

# Summary

- ***Strength***
  - The model applies spatio-temporal matching block to measure the matching degree of a pair of context and candidate.
  - 3D CNN is used to model the multi-turn at the first time.

- ***Weakness***
  - A simple attempt of 3D CNN, the result isn’t good enough.

END!