Review:Inner Attention based Recurrent Neural Networks for Answer Selection

ACL 16

An Weijie

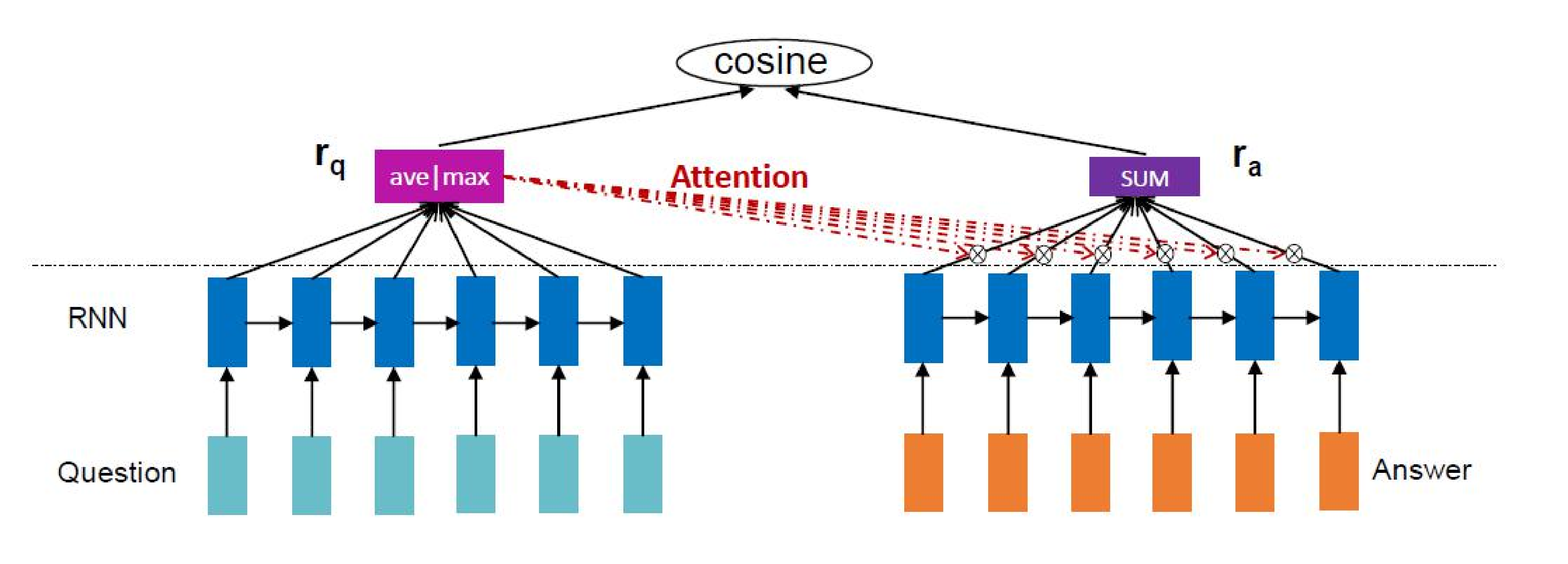
Motivation

- Because of the attention bias problem in traditional attention based RNN models, the author propose three inner attention based RNN models

attention bias problem

- In the RNN architecture, those hidden states near the end of the sentence are expected to capture more information
- the near-the-end hidden variables will be more attended, which may result in a biased attentive weight

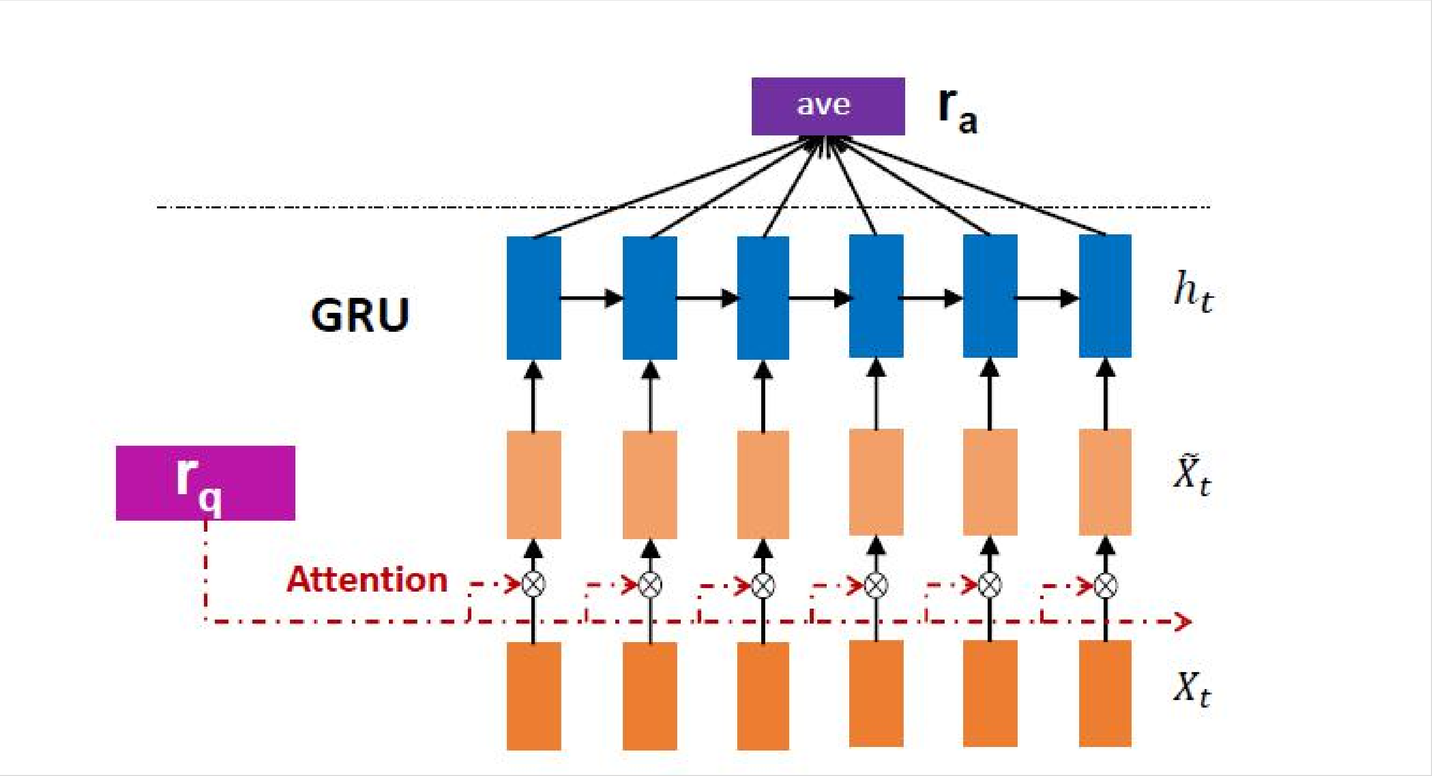
Traditional attention based RNN answerselection model



$$
\begin{aligned}
\mathbf{H}_a &= [\mathbf{h}_a(1), \mathbf{h}_a(2), ..., \mathbf{h}_a(m)] \\
s_t &\propto f_{attention}(\mathbf{r}_q, \mathbf{h}_a(t)) \\
\tilde{\mathbf{h}}_a(t) &= \mathbf{h}_a(t) s_t \\
\mathbf{r}_a &= \sum_{t=1}^{m} \tilde{\mathbf{h}}_a(t)
\end{aligned} \tag{2}
$$

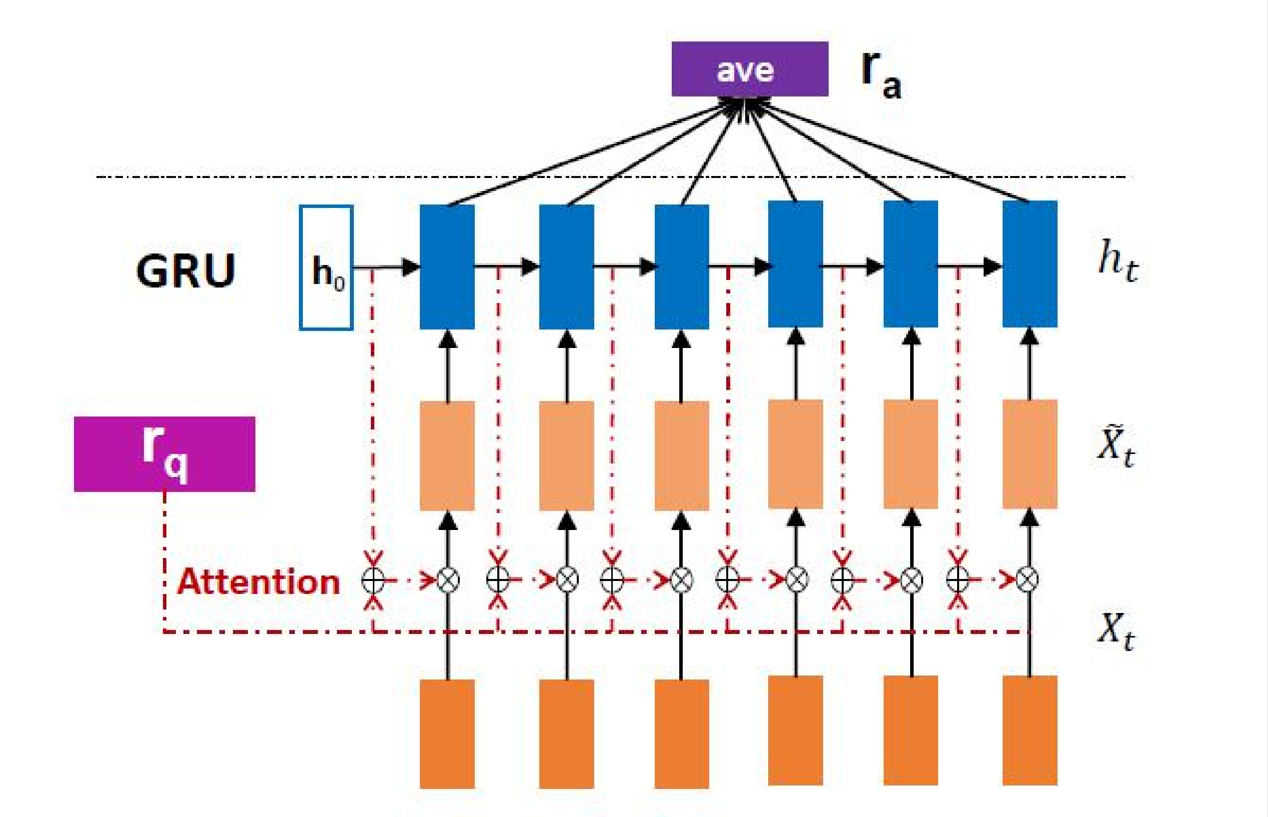
$$
\begin{aligned}
\mathbf{m}(t) &= tanh(\mathbf{W}_{hm}\mathbf{h}_a(t) + \mathbf{W}_{qm}\mathbf{r}_q) \\
f_{attention}(\mathbf{r}_q, \mathbf{h}_a(t)) &= exp(\mathbf{w}_{ms}^T \mathbf{m}(t))
\end{aligned} \tag{3}
$$

IARNN-WORD



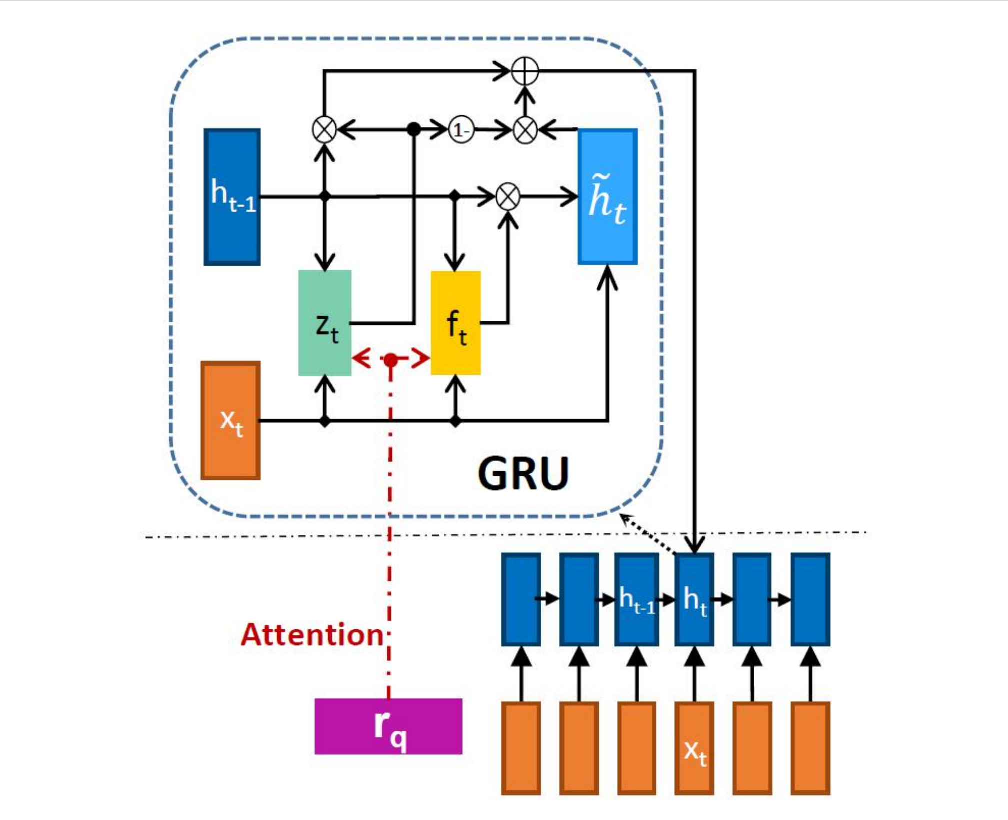
$$\alpha_t = \sigma(\mathbf{r}_q^T \mathbf{M}_{qi} \mathbf{x}_t)$$
$$\tilde{\mathbf{x}}_t = \alpha_t * \mathbf{x}_t \tag{4}$$

IARNN-CONTEXT



$$
\begin{aligned}
\mathbf{w}_C(t) &= \mathbf{M}_{hc}\mathbf{h}_{t-1} + \mathbf{M}_{qc}\mathbf{r}_q \\
\alpha_C^t &= \sigma(\mathbf{w}_C^T(t)\mathbf{x}_t) \\
\tilde{\mathbf{x}}_t &= \alpha_C^t * \mathbf{x}_t
\end{aligned}
\tag{6}
$$

IARNN-GATE



$$
\begin{aligned}
\mathbf{z}_t &= \sigma(\mathbf{W}_{xz}\mathbf{x}_t + \mathbf{W}_{hz}\mathbf{h}_{t-1} + \mathbf{M}_{qz}\mathbf{r}_q) \\
\mathbf{f}_t &= \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{M}_{qf}\mathbf{r}_q) \\
\tilde{\mathbf{h}}_t &= tanh(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}(\mathbf{f}_t \odot \mathbf{h}_{t-1})) \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t
\end{aligned} \tag{7}
$$

IARNN-OCCAM

- An application of Occam's Razor: Among the whole words set, we choose those with fewest number that can represent the sentence.

$$
\begin{aligned}
n_p^i &= \max\{\mathbf{w}_{qp}^T \mathbf{r}_q^i, \lambda_q\} \\
J_i^* &= J_i + n_p^i \sum_{t=1}^{mc} \alpha_t^i
\end{aligned} \tag{8}
$$

Quantify Traditional Attention based Model Bias Problem

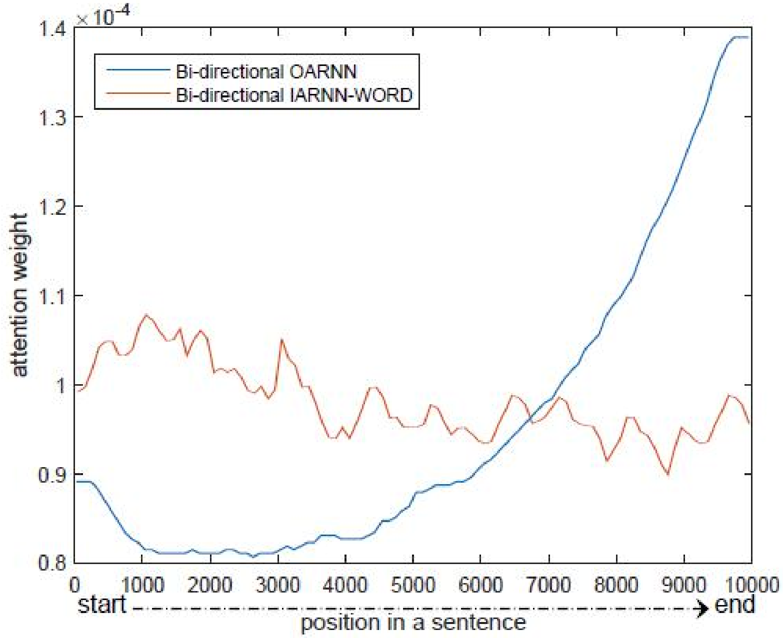


Figure 5: One directional OARNN attention distribution, the horizontal axis is position of word in a sentence that has been normalized from 1 to 10000.

Quantify Traditional Attention based Model Bias Problem

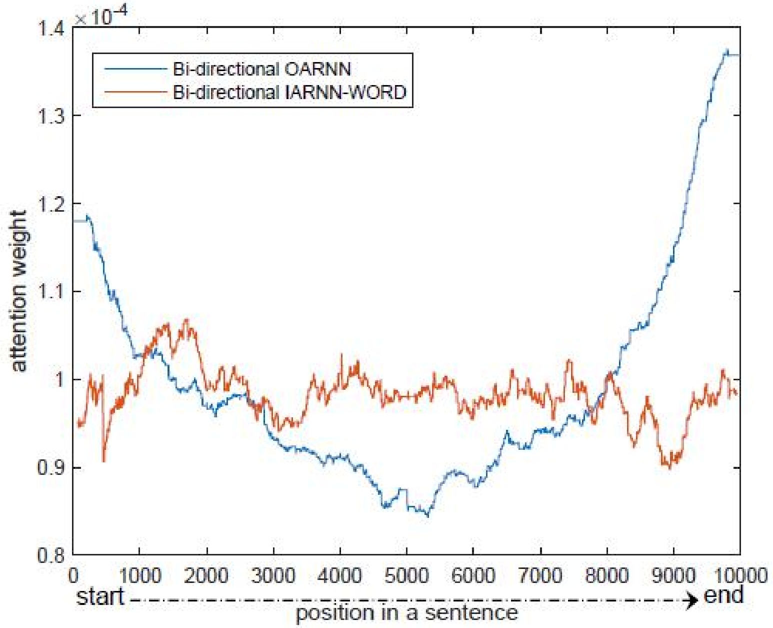


Figure 6: Bi-directional OARNN attention distribution, the horizontal axis is the postion of the word in a sentence that has been normalized from 1 to 10000.

Experiments

System	MAP	MRR
(Yang et al., 2015)	0.652	0.6652
(Yin et al., 2015)	0.6921	0.7108
(Santos et al., 2016)	0.6886	0.6957
GRU	0.6581	0.6691
OARNN	0.6881	0.7013
IARNN-word	0.7098	0.7234
IARNN-Occam(word)	0.7121	0.7318
IARNN-context	0.7182	0.7339
IARNN-Occam(context)	**0.7341**	**0.7418**
IARNN-Gate	0.7258	0.7394

Table 2: Performances on WikiQA

Experiments

System	Dev	Test1	Test2
(Feng et al., 2015)	65.4	65.3	61.0
(Santos et al., 2016)	66.8	67.8	60.3
GRU	59.4	53.2	58.1
OARNN	65.4	66.1	60.2
IARNN-word	67.2125	67.0651	61.5896
IARNN-Occam(word)	69.9130	69.5923	63.7317
IARNN-context	67.1025	66.7211	63.0656
IARNN-Occam(context)	69.1125	68.8651	**65.1396**
IARNN-Gate	**69.9812**	**70.1128**	62.7965

Table 3: Experiment result in InsuranceQA, (Feng et al., 2015) is a CNN architecture without attention mechanism.

Experiments

System	MAP	MRR
(Wang and Nyberg, 2015) †	0.7134	0.7913
(Wang and Ittycheriah, 2015) †	0.7460	0.8200
(Santos et al., 2016) †	**0.7530**	**0.8511**
GRU	0.6487	0.6991
OARNN	0.6887	0.7491
IARNN-word	0.7098	0.7757
IARNN-Occam(word)	0.7162	0.7916
IARNN-context	0.7232	0.8069
IARNN-Occam(context)	0.7272	0.8191
IARNN-Gate	<u>0.7369</u>	<u>0.8208</u>

Table 4: Result of different systems in Trec-QA.(Wang and Ittycheriah, 2015) propose a question similarity model to extract features from word alignment between two questions which is suitable to FAQ based QA. It needs to mention that the system marked with † are learned on TREC-QA original full training data.

Conclution

- Analyze the deficiency of traditional attention based RNN models quantitatively and qualitatively

- present three new RNN models that add attention information before RNN hidden representation, which shows advantage in representing sentence and achieves new state-of-art results in answer selection task