Ruminating Reader: Reasoning with Gated Multi-Hop Attention

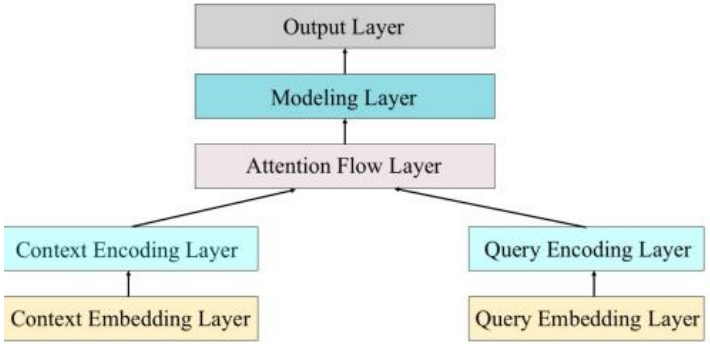
Yichen Gong and **Samuel R. Bowman**
New York University
New York, NY
{yichen.gong,bowman}@nyu.edu

*on arXiv 2017.04.24

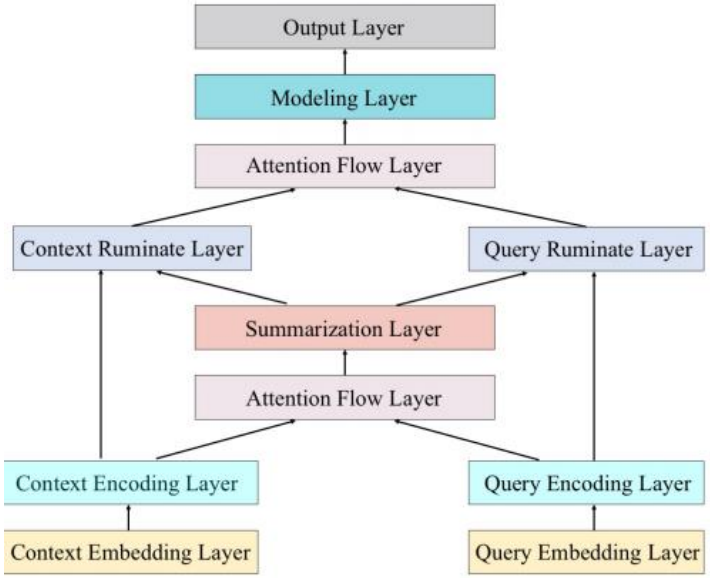
Motivation

- Machine Comprehension (MC): Context, Query -> Answer.
- Answer: a span of words within the context, <start, end>.
- Intuition: humans tend to read a text multiple times for QA.
- This paper: a second pass of attention to revise, fuse information through gating.

Architecture



(a) The high-level structure of BIDAF.



(b) The high-level structure of Ruminating Reader.

(a) Bidirectional Attention Flow for Machine Comprehension. In ICLR 2017.

Method

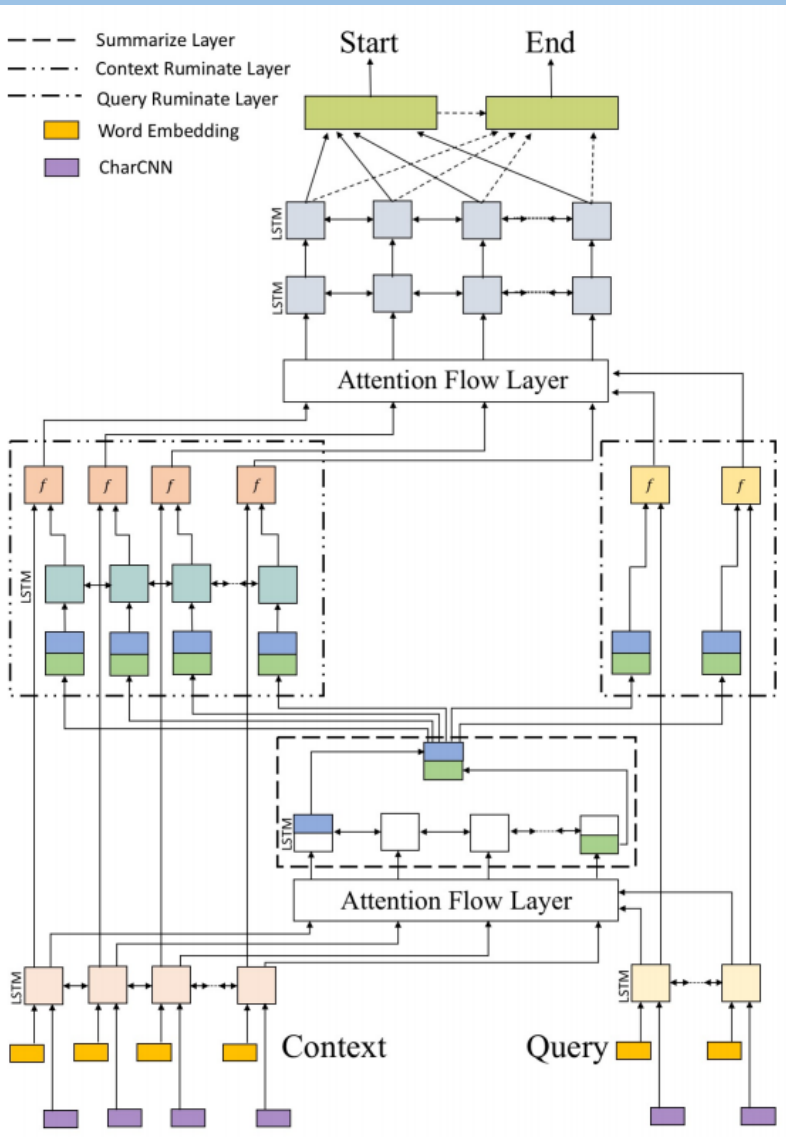


Figure 2: The model structure of our Ruminating Reader.

Bi-LSTM

Character Embedding

Word Embedding

Concact and Highway Net

$$\boldsymbol{C} \in \mathbb{R}^{2d \times C} \quad \boldsymbol{Q} \in \mathbb{R}^{2d \times Q}$$

Method

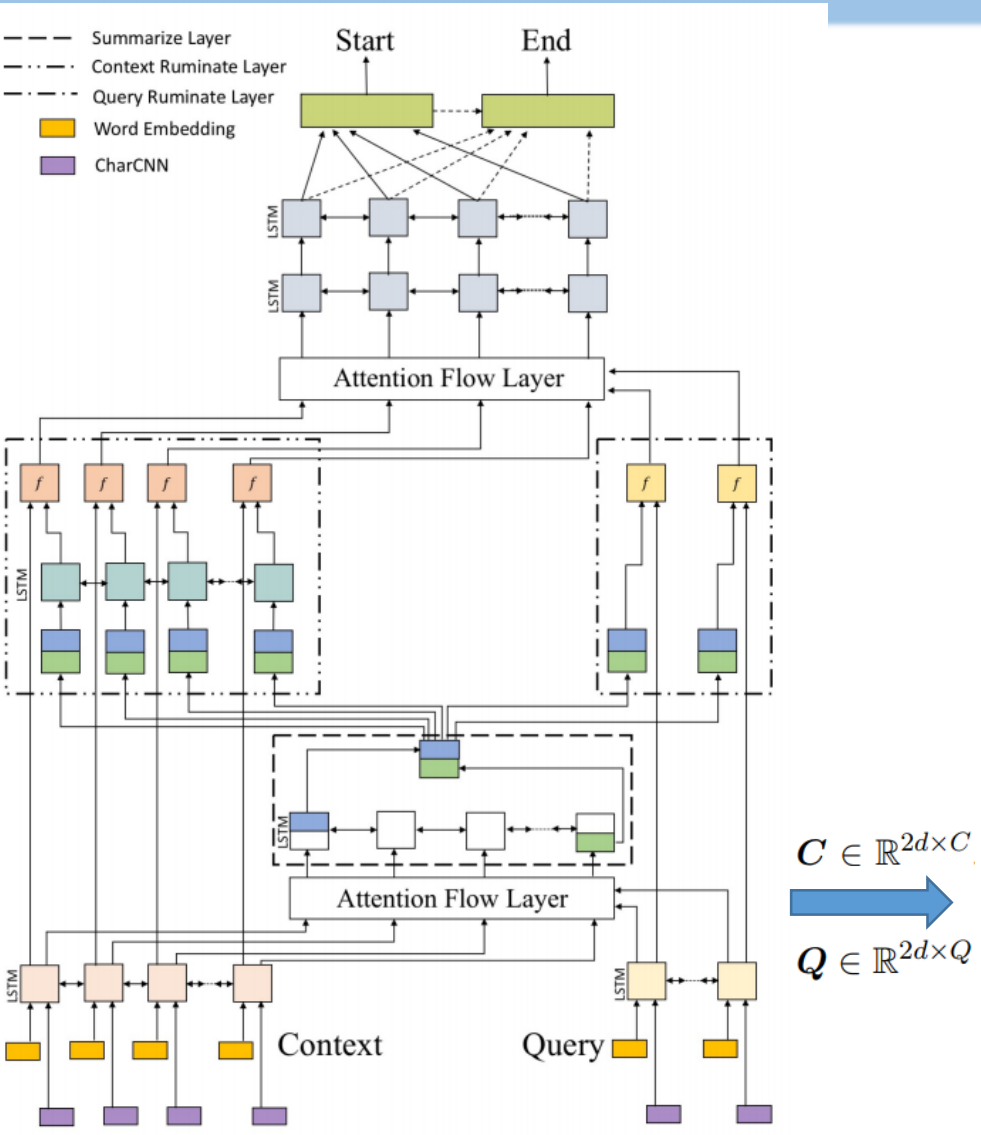


Figure 2: The model structure of our Ruminating Reader.

Interaction Matrix:

$$I_{cq} = w_{(I)}^{\top}[C_c; Q_q; C_c \circ Q_q]$$

Context-to-query Attention:

$$\tilde{Q}_c = \sum(a_{cq} Q_q),$$

$$a_c = softmax(I_c) \in \mathbb{R}^Q.$$

Query-to-context Attention:

$$\tilde{c} = \sum b_c C_c$$

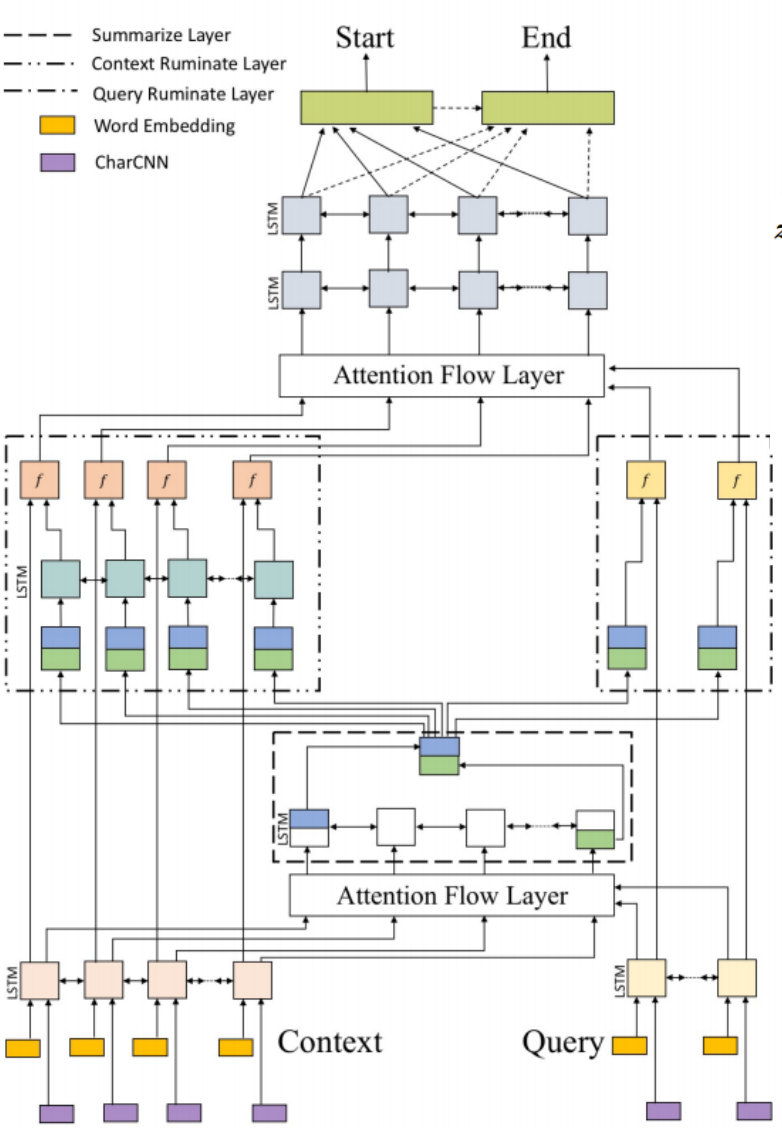
$$b = softmax(max_{col}(\bar{I}))$$

Query-aware context representation:

$$G_c = [C_c; \tilde{Q}_c; C_c \circ \tilde{Q}_c; C_c \circ \tilde{C}_c]$$

$$\in \mathbb{R}^{8d \times C}$$

Method



$$z_i = tanh(W_{Qz}^{1\top} S_{Qi} + W_{Qz}^{2\top} Q_i + b_{Qz}) \qquad z_i = tanh(W_{Cz}^{1\top} \tilde{S}_{Ci} + W_{Cz}^{2\top} C_i + b_{Cz})$$

$$f_i = \sigma(W_{Qf}^{1\top} S_{Qi} + W_{Qf}^{2\top} Q_i + b_{Qf}) \qquad f_i = \sigma(W_{Cf}^{1\top} \tilde{S}_{Ci} + W_{Cf}^{2\top} C_i + b_{Cf})$$

$$\tilde{Q}_i = f_i \circ Q_i + (1 - f_i) \circ z_i \qquad \tilde{C}_i = f_i \circ C_i + (1 - f_i) \circ z_i$$

Ruminate Layer: fuse summarization and encoding
Refine the encoding representation.

Summarization Layer: Bi-LSTM, final states

Figure 2: The model structure of our Ruminating Reader.

Method

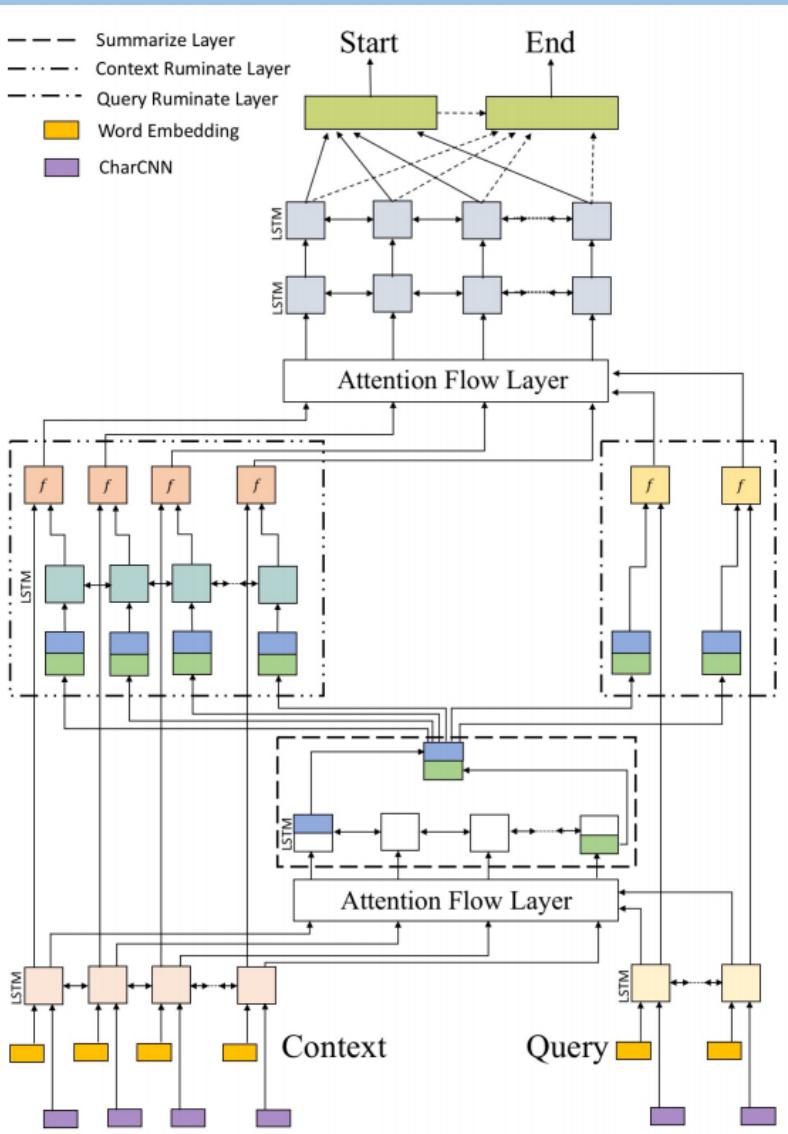


Figure 2: The model structure of our Ruminating Reader.

Softmax

Bi-LSTM

Second Hop Attention Flow Layer

Answer–Question Similarity Loss:

$$s = Argmax(\boldsymbol{p}^1) \tag{11}$$

$$e = Argmax(\boldsymbol{p}^2) \tag{12}$$

$$\vec{\boldsymbol{q}}_{BoW} = \frac{Sum_{row}(\boldsymbol{Q})}{Q} \tag{13}$$

$$AQSL(\theta) = cos(\boldsymbol{C}_s, \vec{\boldsymbol{q}}_{BoW}) + cos(\boldsymbol{C}_e, \vec{\boldsymbol{q}}_{BoW}) \tag{14}$$

Results

- Leaderboard of SQuAD:

Model	Test F1	Test EM
Logistic Regression[a]	51.0	40.4
Dynamic Chunk Reader[b]	70.956	62.499
Fine-grained Gating[c]	73.327	62.446
Match-LSTM[d]	73.743	64.744
Dynamic Coattention Network[e]	75.896	66.233
Bidirectional Attention Flow[f]	77.323	67.974
RaSoR[g]	77.696	69.642
Multi-perspective Matching[h]	77.771	68.877
FastQAExt[i]	78.857	70.849
Document Reader[j]	79.353	70.733
ReasoNet[k]	79.364	70.555
jNet[l]	79.821	70.607
Interactive AoA Reader[‡]	79.937	71.153
QFASE[‡]	79.989	71.898
r-net[‡]	**80.717**	**72.338**
Ruminating Reader	79.456	70.639

Table 2: The Official SQuAD leaderboard performance on test set for single model section from April 23, 2017, the time of submission. There are other unpublished systems shown on leaderboard, including Document Reader and r-net.

Ablation Study:

Ablation Variant	Dev F1	Dev EM
1. BIDAF	77.3	67.7
2. BIDAF w/ L2 Reg., AQSL, LS	77.7	68.6
3. RR w/o query ruminate layer	78.7	69.6
4. RR w/o context ruminate layer	78.9	70.0
5. RR w/ BiLSTM in QRL	**79.4**	70.4
6. RR w/o BiLSTM in CRL	74.0	64.2
7. RR w/o query input at s, f in QRL	78.8	70.1
8. RR w/o context input at s, f in CRL	78.9	70.3
9. RR w/o query input in QRL	63.3	54.1
10. RR w/o context input in CRL	27.0	9.4
11. RR w/o summ. input in QRL	79.2	70.1
12. RR w/o summ. input in CRL	79.2	70.3
Ruminating Reader	**79.5**	**70.6**

Table 3: Layer ablation results. The order of the listing corresponds to the description in Appendix A.1. CRL refers to context ruminate layer and QRL refers to query ruminate layer.

Visualization

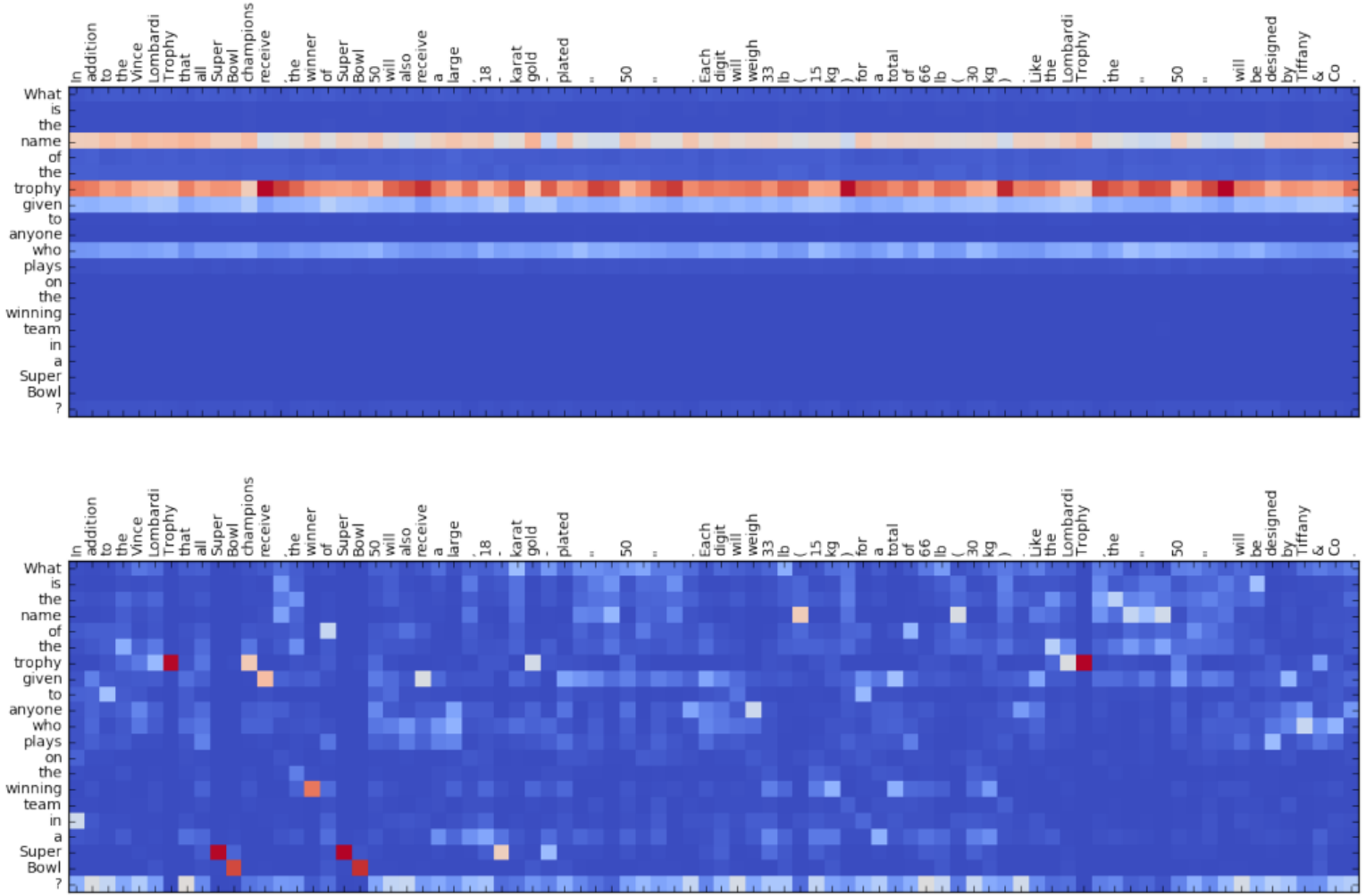
- 

Figure 3: The visualization of first hop (top) and second hop (bottom) attention interaction matrix. We use coolwarm colormap, where red is close to 1 and blue is close to 0. In the question "What is the name of the trophy given to anyone who plays on the winning team in a super Bowl?", the key words *name*, *trophy*, *given*, *who* are strongly attended to in the first hop.

Summarization

- Gating mechanism to fuse information.
- Motivation is not strong.