FusionNet: Fusing via Fully-Aware Attention with Application to Machine Comprehension

Hsin-Yuan Huang[1,2], **Chenguang Zhu**[2], **Yelong Shen**[2], **Weizhu Chen**[2]

[1]National Taiwan University, [2]Microsoft AI+Research

momohuang@gmail.com, {chezhu, yeshen, wzchen}@microsoft.com

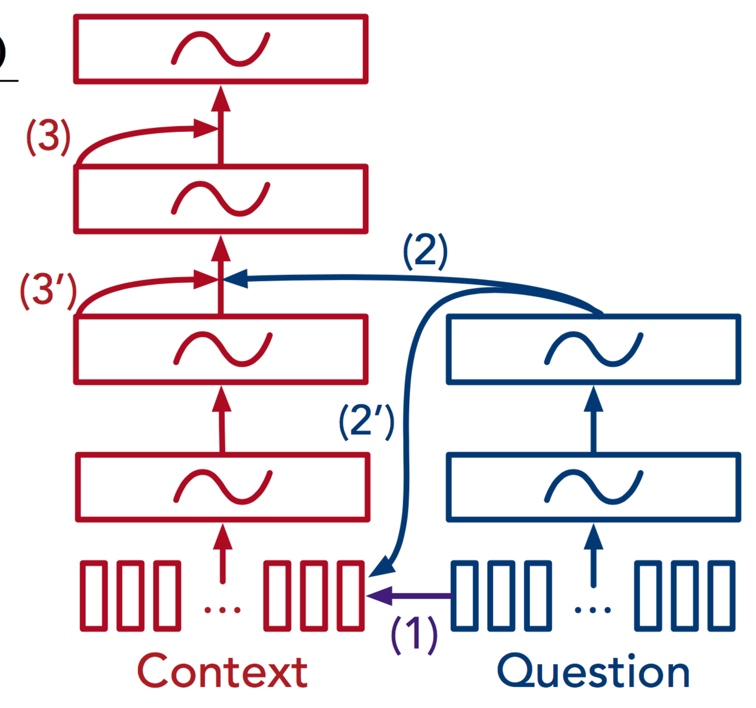
Introduction

- Teaching machines to read, process and comprehend text then answer questions is one of key problems in artificial intelligence.
- Given the context, $\mathbf{C} = \{w_1^C, \dots, w_m^C\}$, and the question: $\mathbf{Q} = \{w_1^Q, \dots, w_n^Q\}$, we want to find the answer $\mathbf{Ans}$. In the SQuAD task [1], the answer $\mathbf{Ans}$ is always $\{w_i^C, \dots, w_{i+k}^C\}$.

Conceptual Architecture for Existing Models

- Input vectors: Embedding vectors for each word in the context and the question.
- Integration components: The rectangular box. Typically LSTM or the like.
- Fusion process: Given two sets of vectors, A and B, enhance or modify *every single vector* in A with information from B (denoted as $A \leftarrow B$).
- A common trait is that none of them employs all levels of representation jointly.

Architectures	(1)	(2)	(2')	(3)	(3')
Match-LSTM (Wang & Jiang, 2016)		✓			
DCN (Xiong et al., 2017)		✓			✓
FastQA (Weissenborn et al., 2017)	✓				
FastQAExt (Weissenborn et al., 2017)	✓	✓		✓	
BiDAF (Seo et al., 2017)		✓			✓
RaSoR (Lee et al., 2016)	✓		✓		
DrQA (Chen et al., 2017a)	✓				
MPCM (Wang et al., 2016)	✓	✓			
Mnemonic Reader (Hu et al., 2017)	✓	✓		✓	
R-net (Wang et al., 2017)		✓		✓	



Fully-Aware Attention on History of Word

Context: The Alpine Rhine is part of the Rhine, a famous European river. The **Alpine Rhine** begins in the most western part of the Swiss canton of Graubünden, and later **forms the border** between Switzerland to the West and **Liechtenstein** and later Austria to the East. On the other hand, the **Danube** separates Romania and Bulgaria.

Question: What is the other country the Rhine separates Switzerland to?

Answer: Liechtenstein

"History of Word" Concept

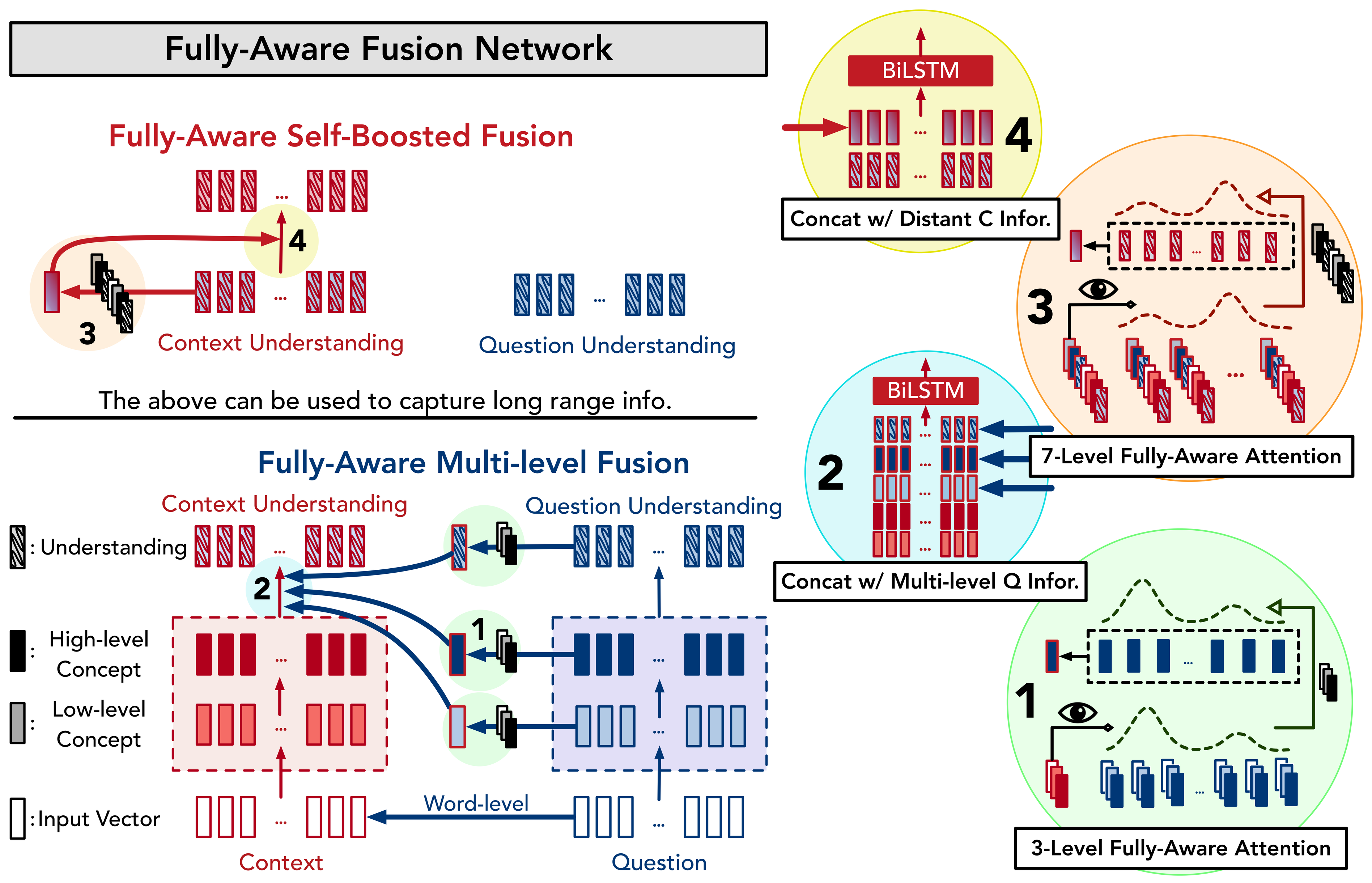
Input Word	Low level	High level
Alpine Rhine →	European river →	Separating River
Forms the border →	Border countries →	Separates
Liechtenstein →	Country →	Country, separate
Danube →	European river →	Separating River

- In neural architectures, we define history of the i-th word, HoW_i, to be the concatenation of all the representations generated for this word.
- Upgrade Normal Attention to Fully-Aware Attention:

$$S(\boldsymbol{h}_i^A, \boldsymbol{h}_j^B) \implies S(\mathrm{HoW}_i^A, \mathrm{HoW}_j^B), \quad S: \text{attention score calculation}$$

- Since HoW_i is much longer than $\boldsymbol{h}_i$, a proper choice of S is crucial for good performance. We will explore this in the experiments.

Fully-Aware Fusion Network



- Fully-Aware Multi-level Fusion will fuse all level of representation in $\mathbf{Q}$ to $\mathbf{C}$ through the proposed *fully-aware attention*.
- Fully-Aware Self-Boosted Fusion will fuse $\mathbf{C}$ to itself to incorporate long-distance information through the proposed *fully-aware attention*.
- For the SQuAD task, the answer finding layer is based on pointer network similar to previous models [2].

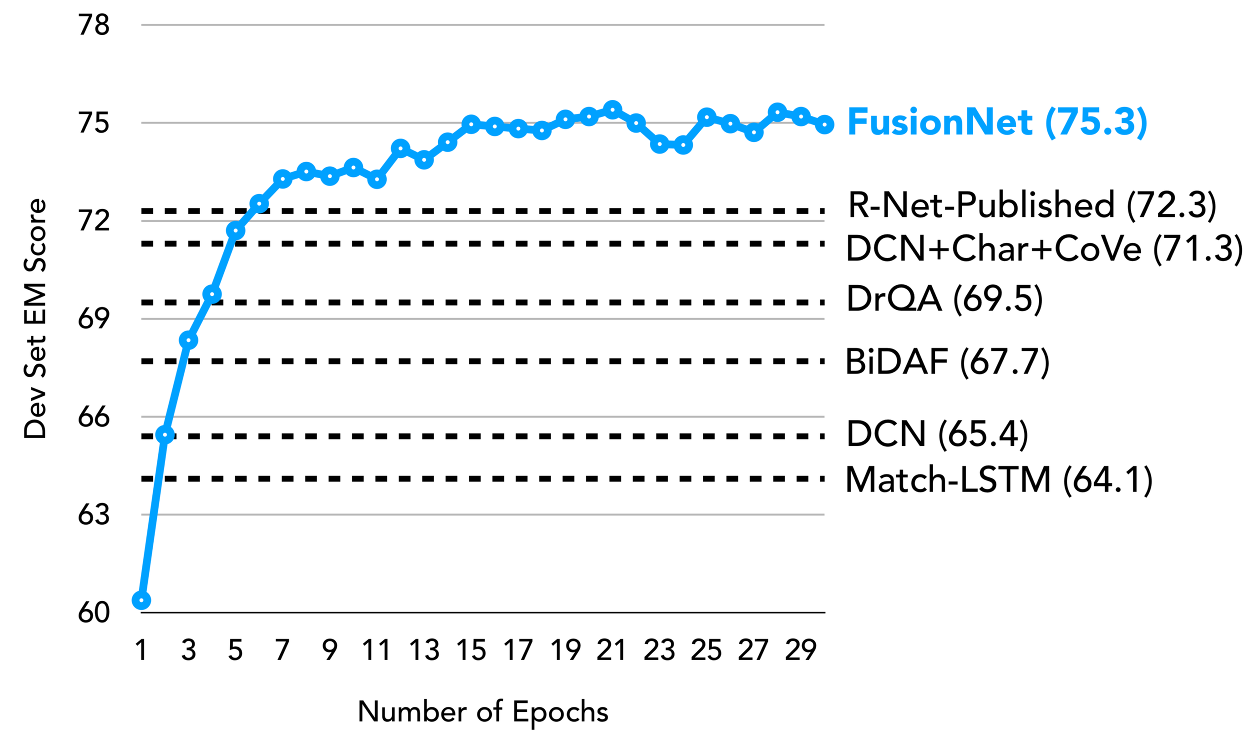
Experiments

- In this section, we focus on the SQuAD dataset [1] and the adversarial SQuAD dataset [3]. We use the standard exact match (EM) and F1 score for evaluation.
- Ablation studies on SQuAD dev set are shown in the following tables, including comparison of attention score $S(\mathbf{x}, \mathbf{y})$ (left) and comparison of model architecture (right).

Attention Function	EM / F1
Additive (MLP)	71.8 / 80.1
Multiplicative	72.1 / 80.6
Scaled Multiplicative	72.4 / 80.7
Scaled Multiplicative + ReLU	72.6 / 80.8
Symmetric Form	73.1 / 81.5
Symmetric Form + ReLU	**75.3 / 83.6**
Previous SotA (Hu et al., 2017)	72.1 / 81.6

Configuration C, Q Fusion	Configuration Self C	EM / F1
High-Level	None	64.6 / 73.2
FA High-Level	None	73.3 / 81.4
FA All-Level	None	72.3 / 80.7
FA Multi-Level	None	74.6 / 82.7
FA Multi-Level	Normal	74.4 / 82.6
FA Multi-Level	FA	**75.3 / 83.6**
Previous SotA (Hu et al., 2017)		72.1 / 81.6

- Performance comparison between FusionNet and existing models on SQuAD dev set (left) and adversarial SQuAD (right).



AddOneSent	EM / F1
LR Baseline	22.3 / 30.4
Match-LSTM (E)	34.8 / 41.8
BiDAF (E)	40.7 / 46.9
SEDT (E)	40.0 / 46.5
Mnemonic Reader (S)	48.5 / 56.0
Mnemonic Reader (E)	48.7 / 55.3
ReasoNet (E)	43.6 / 49.8
FusionNet (E)	**54.7 / 60.7**

AddSent	EM / F1
LR Baseline	17.0 / 23.2
Match-LSTM (E)	24.3 / 34.2
BiDAF (E)	29.6 / 34.2
SEDT (E)	30.0 / 35.0
Mnemonic Reader (S)	39.8 / 46.6
Mnemonic Reader (E)	40.7 / 46.2
ReasoNet (E)	34.6 / 39.4
FusionNet (E)	**46.2 / 51.4**

- By enhancing normal attention with fully-aware attention, we have also improved a state-of-the-art model for natural language inference. An open-source implementation can be found at `https://github.com/momohuang/FusionNet-NLI`.

References

[1] Rajpurkar et al. Squad: 100,000+ questions for machine comprehension of text. *EMNLP*, 2016.
[2] Wang et al. Gated self-matching networks for reading comprehension and question answering. In *ACL*, 2017.
[3] Jia et al. Adversarial examples for evaluating reading comprehension systems. *EMNLP*, 2017.
[4] Wang et al. Machine comprehension using match-lstm and answer pointer. *ICLR*, 2017.
[5] Chen et al. Reading wikipedia to answer open-domain questions. *ACL*, 2017.