ACL 2018

July 15 – 20

Melbourne, Australia

Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering

Wei Wang[1*], Ming Yan[1*] and Chen Wu[1*]

[1] Alibaba Group, 969 West Wenyi Road, Hangzhou 311121, China

{hebian.ww, ym119608, wuchen.wc}@alibaba-inc.com

Alibaba Group

阿里巴巴机器智能技术实验室

Alibaba Machine Intelligence Technology Lab

This paper describes a novel hierarchical attention network for reading comprehension style question answering. The main contribution of this paper lies in: (1) We propose a novel **hierarchical attention network** which combines co-attention and self-attention in a multi-step style. Attention and fusion are conducted horizontally and vertically across layers at different levels of granularity between question and paragraph. (2) We design a fine-grained **fusion approach** to blend attention vectors with the global representation for a better understanding of the question and passage. At the time of writing the paper (Jan. 12th 2018), our model **achieves the first position** on the SQuAD leaderboard for both single and ensemble models. We also achieves state-of-the-art results on TriviaQA dataset.

Introduction

As a brand new field in question answering community, reading comprehension is one of the key problems in artificial intelligence, which aims to read and comprehend a given text, and then answer questions based on it. This task is challenging which requires a comprehensive understanding of natural languages and the ability to do further inference and reasoning.

Benefiting from the availability of SQuAD benchmark dataset, rapid progress has been made these years. The large volume of the data available makes it possible to train end-to-end deep neural methods, among which the attention-based methods are most widely used.

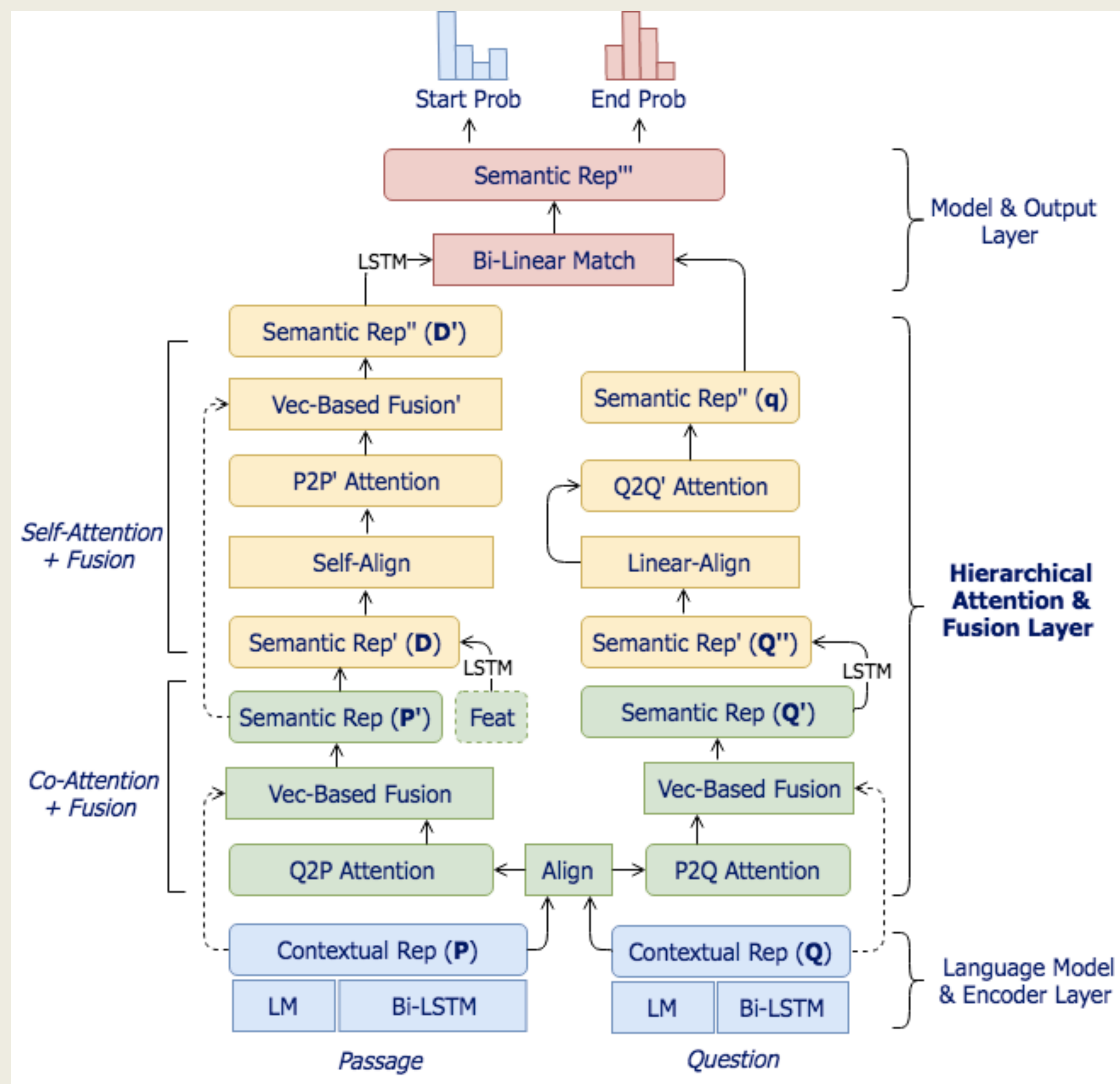
The idea of our approach derives from the normal human reading pattern. First, people scan through the whole passage to catch a glimpse of the main body of the passage. Then with the question in mind, people make connection between passage and question, and understand the main intent of the question related with the passage theme. A rough answer span is then located from the passage and the attention can be focused on to the located context. Finally, to prevent from forgetting the question, people come back to the question and select a best answer according to the previously located answer span.

Encode-Interaction-Pointer Framework

Framework: We follow the basic encode-interaction-pointer framework in MRC task. The proposed framework consists of four typical layers to learn different concepts of semantic representations :

- **Encoder Layer** as a language model, utilizes contextual cues from surrounding words to refine the embedding of the words.
- **Attention Layer** attempts to capture relations between question and passage.
- **Match Layer** employs a bi-linear match function to compute the relevance between the question and passage representation
- **Output Layer** uses a pointer network to search the answer span of question.

Hierarchical Attention Fusion Network



Language Model & Encoder Layer

We use a pre-trained word embedding model (glove 840B) and a char embedding model (**ELMo**) to lay the foundation for our whole framework.

$$\mathrm{u}_t^Q = \left[\mathrm{BiLSTM_Q}([\mathrm{e}_t^Q, \mathrm{c}_t^Q]), \mathrm{c}_t^Q\right]$$

$$\mathrm{u}_t^P = \left[\mathrm{BiLSTM_P}([\mathrm{e}_t^P, \mathrm{c}_t^P]), \mathrm{c}_t^P\right]$$

Hierarchical Attention & Fusion Layer

We propose a hierarchical attention structure by combining the co-attention and self-attention mechanism in a multi-hop style.

Co-attention & Fusion

- P2Q Attention
- Q2P Attention

$$\mathrm{S_{ij}} = \mathrm{Att}(\mathrm{u}_t^Q, \mathrm{u}_t^P) = \mathrm{ReLU}(\mathrm{W_{lin}^\top u_t^Q})^\top \cdot \mathrm{ReLU}(\mathrm{W_{lin}^\top u_t^P})$$

$$\beta_i = softmax(S_{i:})$$

$$\tilde{P}_{k:} = \sum_i \beta_{ik} \cdot P_{i:}, \forall i \in [1, ..., n]$$

Since we find that the original contextual representations are important in reflecting the semantics **at a more global level**, we also introduce different levels of **gating mechanism** to incorporate the projected representations with the original contextual representations .

Fusion with gating layer

$$P' = g(P, \tilde{Q}) \cdot m(P, \tilde{Q}) + (1 - g(P, \tilde{Q})) \cdot P$$

$$m(P, \tilde{Q}) = tanh(W_f[P; \tilde{Q}; P \circ \tilde{Q}; P - \tilde{Q}] + b_f)$$

Self-attention & Fusion

$$L = softmax(D \cdot W_l \cdot D^\top)$$

$$\tilde{D} = L \cdot D$$

$$D' = Fuse(D, \tilde{D})$$

Question side

since it is generally shorter in length and could be adequately represented with less information, we aggregate the resulting hidden units into one single question vector, with a linear self-alignment.

$$\gamma = softmax(\mathbf{w}_q^\top \cdot Q'')$$

$$\mathbf{q} = \sum_j \gamma_j \cdot Q''_{:j}, \forall j \in [1, ..., m]$$

Model & Output Layer

$$P_{start} = softmax(\mathbf{q} \cdot W_s^\top \cdot D'')$$

$$P_{end} = softmax(\mathbf{q} \cdot W_e^\top \cdot D'')$$

$$L(\theta) = -\frac{1}{N} \sum_i^N \log p_s(y_i^s) + \log p_e(y_i^e)$$

Experiments

	Dev Set	Test Set
Single model	**EM / F1**	**EM / F1**
LR Baseline (Rajpurkar et al., 2016)	40.0 / 51.0	40.4 / 51.0
Match-LSTM (Wang and Jiang, 2016)	64.1 / 73.9	64.7 / 73.7
DrQA (Chen et al., 2017a)	- / -	70.7 / 79.4
DCN+ (Xiong et al., 2017)	74.5 / 83.1	75.1 / 83.1
Interactive AoA Reader+ (Cui et al., 2016)	- / -	75.8 / 83.8
FusionNet (Huang et al., 2017)	- / -	76.0 / 83.9
SAN (Liu et al., 2017b)	76.2 / 84.0	76.8 / 84.4
AttentionReader+ (unpublished)	- / -	77.3 / 84.9
BiDAF + Self Attention + ELMo (Peters et al., 2018)	- / -	78.6 / 85.8
r-net+ (Wang et al., 2017)	- / -	79.9 / 86.5
SLQA+	**80.0 / 87.0**	**80.4 / 87.0**
Ensemble model		
FusionNet (Huang et al., 2017)	- / -	78.8 / 85.9
DCN+ (Xiong et al., 2017)	- / -	78.9 / 86.0
Interactive AoA Reader+ (Cui et al., 2016)	- / -	79.0 / 86.4
SAN (Liu et al., 2017b)	78.6 / 85.9	79.6 / 86.5
BiDAF + Self Attention + ELMo (Peters et al., 2018)	- / -	81.0 / 87.4
AttentionReader+ (unpublished)	- / -	81.8 / 88.2
r-net+ (Wang et al., 2017)	- / -	82.6 / 88.5
SLQA+	**82.0 / 88.4**	**82.4 / 88.6**
Human Performance	80.3 / 90.5	82.3 / 91.2

Fusion Kernel	EM / F1
Simple Concat	78.8 / 85.8
Add Full Projection (FPU)	79.1 / 86.1
Scalar-based Fusion (SFU)	79.5 / 86.5
Vector-based Fusion (VFU)	**80.0 / 87.0**
Matrix-based Fusion (MFU)	79.8 / 86.8

SLQA single model	EM / F1
SLQA+	**80.0 / 87.0**
-Manual Features	79.2 / 86.2
-Language Embedding (ELMo)	77.6 / 84.9
-Self Matching	79.5 / 86.4
-Multi-hop	79.1 / 86.1
-Bi-linear Match	65.4 / 72.0
-Fusion (simple concat)	78.8 / 85.8
-Fusion, -Multi-hop	77.5 / 84.8
-Fusion, -Bi-linear Match	63.1 / 69.6

Fig 1. Ablation results on SQuAD

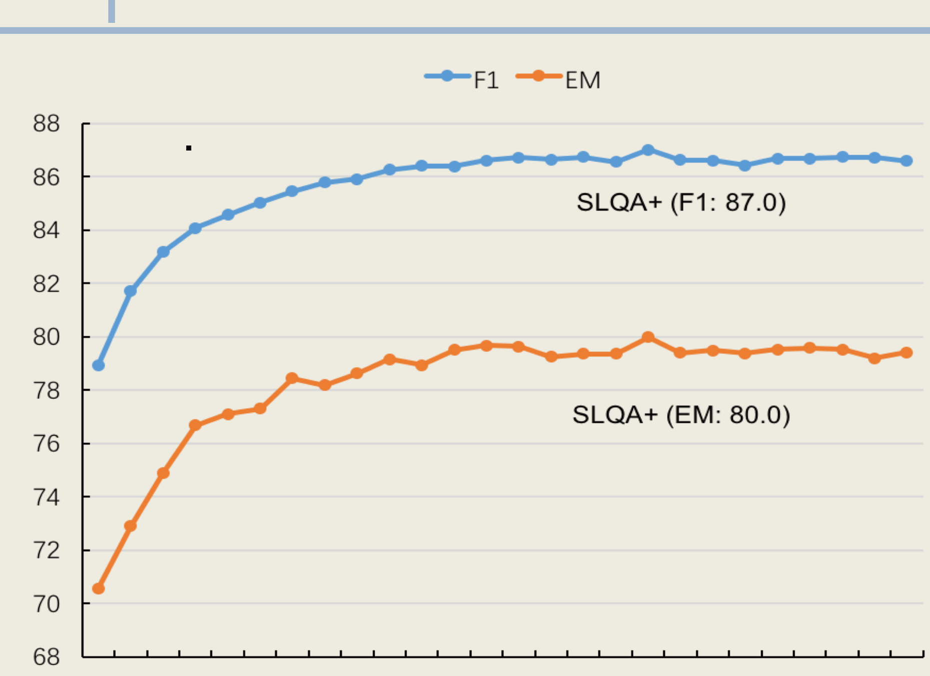


Fig 2. Learning curve of F1/EM on SQuAD

	Full	Verified
Model	**EM / F1**	**EM / F1**
BiDAF (Seo et al., 2016)	40.26 / 45.74	47.47 / 53.70
MEMEN (Pan et al., 2017)	43.16 / 46.90	49.28 / 55.83
M-Reader (Hu et al., 2017)	46.94 / 52.85	54.45 / 59.46
QANet (Yu et al., 2018)	51.10 / 56.60	53.30 / 59.20
document-qa (Clark and Gardner, 2017)	63.99 / 68.93	67.98 / 72.88
dirkweissenborn (unpublished)	64.60 / 69.90	72.77 / 77.44
SLQA-Single	**66.56 / 71.39**	**74.83 / 78.74**

Fig 3. Published and unpublished results on TriviaQA Wikipedia Leaderboard