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 documented the first position on the SQuAD leaderboard for both single and ensemble models. We also achieve 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 get 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.

**Language Model & Encoder Layer**

We use a pre-trained word embedding model (glove 40B) and a char embedding model (ELMo) to lay the foundation for our whole framework.

**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**

- Co-attention
- Q2P Attention
- Q2P Attention

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**

\[ P' = g(P, Q) \cdot m(P, Q) + (1 - g(P, Q)) \cdot P \]

\[ m(P, Q) = \tanh(W_p[P, Q; P \circ Q; P - Q] + b) \]

**Self-attention & Fusion**

\[ L = \text{softmax}(D \cdot W^L \cdot D^T) \]

\[ \tilde{D} = L \cdot D \]

\[ D' = \text{Fused}(D, \tilde{D}) \]

**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-attention.

\[ \gamma = \text{softmax}(w^T \cdot Q^n) \]

\[ q = \sum_{j} \gamma_j \cdot Q''_j, \forall j \in [1, \ldots, m] \]

**Model & Output Layer**

\[ P_{\text{start}} = \text{softmax}(q \cdot W^T_{\text{start}} \cdot D^n) \]

\[ P_{\text{end}} = \text{softmax}(q \cdot W^T_{\text{end}} \cdot D^n) \]

\[ L(\theta) = \frac{1}{N} \sum_{i} \log p_s(y_i^f) + \log p_e(y_i^e) \]

**Experiments**

We compare our results with the state-of-the-art results on the SQuAD leaderboard.

**Fusion Kernel**

- Simple Concat
- Add Full Projection (FPU)
- Scalar-based Fusion (SFU)
- Vector-based Fusion (VUF)
- Matrix-based Fusion (MUF)

**SLQA-single**

- SLQA+
  - Manual Features
  - Language Embedding (ELMo)
  - Self-Matching
  - Multi-hop
  - Bi-linear Match
  - Fusion (sim-concat)
  - Fusion, Multi-hop
  - Fusion, Bi-linear Match

**SLQA+**

- SLQA+
  - EM / F1: 80.0 / 87.0

**SLQA-Single**

- SLQA+ (FL: 87.0)

**Fig 1. Ablation results on SQuAD**

**Fig 2. Learning curve of F1/EM on SQuAD**

**Fig 3. Published and unpublished results on TriviaQA Wikipedia Leaderboard**