SLQA 模型：Semantic learning for QA

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

Wei Wang[*], **Ming Yan**[*], **Chen Wu**[*]
Alibaba Group, 969 West Wenyi Road, Hangzhou 311121, China
{hebian.ww,ym119608,wuchen.wc}@alibaba-inc.com

Abstract

This paper describes a novel hierarchical attention network for reading comprehension style question answering, which aims to answer questions for a given narrative paragraph. In the proposed method, attention and fusion are conducted horizontally and vertically across layers at different levels of granularity between question and paragraph. Specifically, it first encode the question and paragraph with fine-grained language embeddings, to better capture the respective representations at semantic level. Then it proposes a multi-granularity fusion approach to fully fuse information from both global and attended representations. Finally, it introduces a hierarchical attention network to focuses on the answer span progressively with multi-level soft-alignment. Extensive experiments on the large-scale SQuAD and TriviaQA datasets validate the effectiveness of the proposed method. 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, AddSent and AddOne-Sent datasets.

叙述的

各自的

1 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. Restricted by the limited volume of the annotated dataset, early studies mainly rely on a pipeline of NLP models to complete this task, such as semantic parsing and linguistic annotation (Das et al., 2014). Not until the release of large-scale cloze-style dataset, such as Children's Book Test (Hill et al., 2015) and CNN/Daily Mail (Hermann et al., 2015), some preliminary end-to-end deep learning methods have begun to bloom and achieve superior results in reading comprehension task (Hermann et al., 2015; Chen et al., 2016; Cui et al., 2016).

However, these cloze-style datasets still have their limitations, where the goal is to predict the single missing word (often a named entity) in a passage. It requires less reasoning than previously thought and no need to comprehend the whole passage (Chen et al., 2016). Therefore, Stanford publish a new large-scale dataset SQuAD (Rajpurkar et al., 2016), in which all the question and answers are manually created through crowdsourcing. Different from cloze-style reading comprehension dataset, SQuAD constrains answers to all possible text spans within the reference passage, which requires more logical reasoning and content understanding.

完形填空式的阅读理解需要更少的推理，而且不需要理解整个段落.

需要更多的逻辑推理和内容理解

Benefiting from the availability of SQuAD benchmark dataset, rapid progress has been made these years. The work (Wang and Jiang, 2016) and (Seo et al., 2016) are among the first to investigate into this dataset, where Wang and Jiang propose an end-to-end architecture based on match-LSTM and pointer networks (Wang and Jiang, 2016), and Seo et al. introduce the bi-directional attention flow network which captures the question-document context at different levels of granularity (Seo et al., 2016). Chen et al. devise a simple and effective document reader, by introducing a bilinear match function and a few manual features (Chen et al., 2017a). Wang et al. propose a gated attention-based recurrent network where self-match attention mechanism is first incorporated (Wang et al., 2017). In (Liu et al., 2017b) and (Shen et al., 2017), the multi-turn memory networks are designed to simulate multi-step reasoning in machine reading comprehension.

本文的方法和思路是在仿真人的阅读模式.

The idea of our approach derives from the normal human reading pattern. ①First, people scan through the whole passage to catch a glimpse of the main body of the passage. ②Then 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. ③A rough answer span is then located from the passage and the attention can be focused on to the located context. ④Finally, to prevent from forgetting the question, people come back to the question and ⑤select a best answer according to the previously located answer span.

本文模型的三个部分

Inspired by this, we propose a hierarchical attention network which can gradually focus the attention on the right part of the answer boundary, while capturing the relation between the question and passage at different levels of granularity, as illustrated in Figure 1. Our model mainly consists of three joint layers: 1) encoder layer where pre-trained language models and recurrent neural networks are used to build representation for questions and passages separately; 2) attention layer in which hierarchical attention networks are designed to capture the relation between question and passage at different levels of granularity; 3) match layer where refined question and passage are matched under a pointer-network (Vinyals et al., 2015) answer boundary predictor.

encoder layer

In encoder layer, to better represent the questions and passages in multiple aspects, we combine two different embeddings to give the fundamental word representations. In addition to the typical glove word embeddings, we also utilize the ELMo embeddings (Peters et al., 2018) derived from a pre-trained language model, which shows superior performance in a wide range of NLP problems. Different from the original fusion way for intermediate layer representations, we design a representation-aware fusion method to compute the output ELMo embeddings and the context information is also incorporated by further passing through a bi-directional LSTM network.

The key in machine reading comprehension solution lies in how to incorporate the question context into the paragraph, in which attention mechanism is most widely used. Recently, many different attention functions and types have been designed (Xiong et al., 2016; Seo et al., 2016; Wang et al., 2017), which aims at properly aligning the question and passage. In our attention layer, we propose a hierarchical attention network by leveraging both the co-attention and self-attention mechanism, to gradually focus our attention on the best answer span. Different from the previous attention-based methods, we constantly complement the aligned representations with 不断地 global information from the previous layer, and an additional fusion layer is used to further refine the representations. In this way, our model can make some minor adjustment so that the attention will always be on the right place.

attention layer

match layer

Based on the refined question and passage representation, a bilinear match layer is finally used to identify the best answer span with respect to the question. Following the work of (Wang and Jiang, 2016), we predict the start and end boundary within a pointer-network output layer.

The proposed method achieves state-of-the-art results against strong baselines. Our single model achieves 79.2% EM and 86.6% F1 score on the hidden test set, while the ensemble model further boosts the performance to 82.4% EM and 88.6% F1 score. At the time of writing the paper (Jan. 12th 2018), our model SLQA+ (Semantic Learning for Question Answering) achieves the first position on the SQuAD leaderboard [1] for both single and ensemble models. Besides, we are also among the first to surpass human EM performance on this golden benchmark dataset.

2 Related Work

2.1 Machine Reading Comprehension

Traditional reading comprehension style question answering systems rely on a pipeline of NLP models, which make heavy use of linguistic annotation, structured world knowledge, semantic parsing and similar NLP pipeline outputs (Hermann et al., 2015). Recently, the rapid progress of machine reading comprehension has largely benefited from the availability of large-scale benchmark datasets and it is possible to train large end-to-end neural network models. Among them, CNN/Daily Mail (Hermann et al., 2015) and Children's Book Test (Hill et al., 2015) are the first

[1] https://rajpurkar.github.io/SQuAD-explorer/

large-scale datasets for reading comprehension task. However, these datasets are in cloze-style, in which the goal is to predict the missing word (often a named entity) in a passage. Moreover, Chen at al. have also shown that these cloze-style datasets requires less reasoning than previously thought (Chen et al., 2016). Different from the previous datasets, the SQuAD provides a more challenging benchmark dataset, where the goal is to extract an arbitrary answer span from the original passage.

2.2 Attention-based Neural Networks

The key in MRC task lies in how to incorporate the question context into the paragraph, in which attention mechanism is most widely used. In spite of a variety of model structures and attention types (Cui et al., 2016; Xiong et al., 2016; Seo et al., 2016; Wang et al., 2017; Clark and Gardner, 2017), a typical attention-based neural network model for MRC first encodes the symbolic representation of the question and passage in an embedding space, then identify answers with particular attention functions in that space. In terms of the question and passage attention or matching strategy, we roughly categorize these attention-based models into two large groups: one-way attention and two-way attention.

eg. docQA

In one-way attention model, question is first summarized into a single vector and then directly matched with the passage. Most of the end-to-end neural network methods on the cloze-style datasets are based on this model (Hermann et al., 2015; Kadlec et al., 2016; Chen et al., 2016; Dhingra et al., 2016). Hermann et al. are the first to apply the attention-based neural network methods to MRC task and introduce an attentive reader and an impatient reader (Hermann et al., 2015), by leveraging a two layer LSTM network. Chen et al. (Chen et al., 2016) further design a bilinear attention function based on the attentive reader, which shows superior performance on CNN/Daily Mail dataset. However, part of information may be lost when summarizing the question and a fine-grained attention on both the question and passage words should be more reasonable.

eg. SAN

Therefore, the two-way attention model unfolds both the question and passage into respective word embeddings, and compute the attention in a two-dimensional matrix. Most of the top-ranking methods on SQuAD leaderboard are based on this attention mechanism (Wang et al., 2017; Huang et al., 2017; Xiong et al., 2017; Liu et al., 2017b,a). (Cui et al., 2016) and (Xiong et al., 2016) introduce the co-attention mechanism to better couple 结合 the representations of the question and document. Seo et al. propose a bi-directional attention flow network to capture the relevance at different levels of granularity (Seo et al., 2016). (Wang et al., 2017) further introduce the self-attention mechanism to refine the representation by matching the passage against itself, to better capture the global passage information. Huang et al. introduce a fully-aware attention mechanism with a novel *history-of-word* concept (Huang et al., 2017).

We propose a hierarchical attention network by leveraging both co-attention and self-attention mechanisms in different layers, which can capture the relevance between the question and passage at different levels of granularity. Different from the above methods, we further devise a fusion function to combine both the aligned representation and the original representation from the previous layer within each attention. In this way, the model can always focus on the right part of the passage, while keeping the global passage topic in mind.

3 Machine Comprehension Model

3.1 Task Description

Typical machine comprehension systems take an evidence text and a question as input, and predict a span within the evidence that answers the question. Based on this definition, given a passage and a question, the machine needs to first read and understand the passage, and then finds the answer to the question. The passage is described as a sequence of word tokens $\mathrm{P} = \left\{\mathrm{w}_\mathrm{t}^\mathrm{P}\right\}_{\mathrm{t}=1}^{\mathrm{n}}$ 段落中的词 and the question is described as $\mathrm{Q} = \left\{\mathrm{w}_\mathrm{t}^\mathrm{Q}\right\}_{\mathrm{t}=1}^{\mathrm{m}}$, 问题中的词 where n is the number of words in the passage, and m is the number of words in the question. In general, $\mathrm{n} \gg \mathrm{m}$. The answer can have different types depending on the task. In the SQuAD dataset (Rajpurkar et al., 2016), the answer A is guaranteed to be a continuous span in the passage P. The object function for machine reading comprehension is to learn a function $\mathrm{f}(\mathrm{q}, \mathrm{p}) = \arg\max_{\mathrm{a} \in \mathrm{A}(\mathrm{p})} \mathrm{P}(\mathrm{a}|\mathrm{q}, \mathrm{p})$. The training data is a set of the question, passage and answer tuples $< \mathrm{Q}, \mathrm{P}, \mathrm{A} >$.

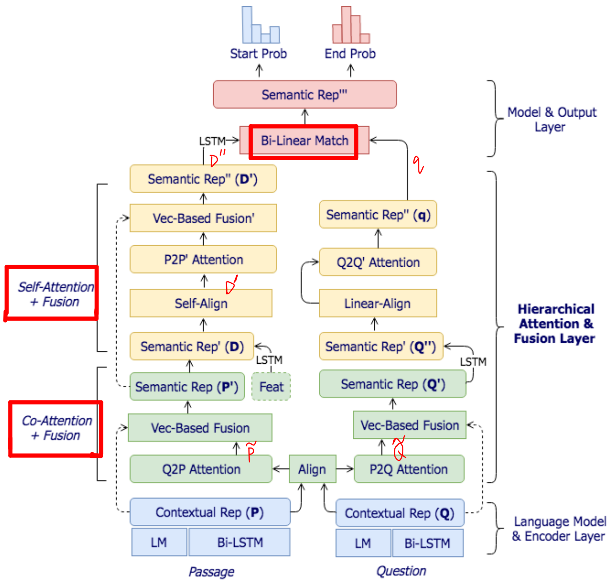


Figure 1: Hierarchical Attention Fusion Network.

3.2 Encode-Interaction-Pointer Framework

We will now describe our framework from the bottom up. As show in Figure 1, the proposed framework consists of four typical layers to learn different concepts of semantic representations:

把段落和问题从 token 转化成语义表达.

- **Encoder Layer** as a language model, utilizes contextual cues from surrounding words to refine the embedding of the words. It converts the passage and question from tokens to semantic representation;

捕捉段落和问题之间的关系.

- **Attention Layer** attempts to capture relations between question and passage. Besides the aligned context, the contextual embeddings are also merged by a fusion function. Moreover, the multi-level of this operation forms a ”working memory”;

在span level上计算段落和问题之间的相关性

- **Match Layer** employs a bi-linear match function to compute the relevance between the question and passage representation on a span level;

搜索答案范围

- **Output Layer** uses a pointer network to search the answer span of question.

The main contribution of this work is the attention layer, in order to capture the relationship between question and passage, a hierarchical strategy is used to progressively make the answer boundary clear with the refined attention mechanism. A fine-grained fusion function is also introduced to better align the contextual representations from different levels. The detailed description of the model is provided as follows.

重点在 attention layer.

3.3 Hierarchical Attention Fusion Network

Our design is based on a simple but natural intuition: performing fine-grained mechanism requires first to roughly see the potential answer domain and then progressively locate the most discriminative parts of the domain.

The overall framework of our Hierarchical Attention Fusion Network is shown in Figure 1. It consists of several parts: a basic co-attention layer with shallow semantic fusion, a self-attention layer with deep semantic fusion and a memory-wise bilinear alignment function. The proposed network has two distinctive characteristics: (i) A fine-grained fusion approach to blend attention vectors for a better understanding of the relationship between question and passage; (ii) A multi-granularity attention mechanism applied at the word and sentence-level, enabling it to properly attend to the most important content when constructing the question and passage representation. Experiments conducted on SQuAD and adversarial example datasets (Jia and Liang, 2017) demonstrate that the proposed framework outperform previous methods by a large margin. Details of different components will be described in the following sections.

浅层语义融合

深层语义融合

特点

混合

细粒度的 fusion approach

multi-granularity attention mechanism

3.4 Language Model & Encoder Layer

Encoder layer of the model transform the discrete word tokens of question and passage to a sequence of continuous vector representations. We use a pre-trained word embedding model and a char embedding model to lay the foundation for our model. For the word embedding model, we adopt the popular glove embeddings (Pennington et al., 2014) which are widely used in deep learning-based NLP domain. For the char embedding model, the ELMo language model (Peters et al., 2018) is used due to its superior performance in a wide range of NLP tasks. As a result, we obtain two types of encoding vectors, i.e., word embeddings $\{e_t^Q\}_{t=1}^m$, $\{e_t^P\}_{t=1}^n$ and char embeddings $\{c_t^Q\}_{t=1}^m$, $\{c_t^P\}_{t=1}^n$.

To further utilize contextual cues from surrounding words to refine the embedding of the words, we then put a shared Bi-LSTM network on top of the embeddings provided by the previous layers to model the temporal interactions between words. Before feeding into the Bi-LSTM

暂时的

contextual network, we concat the word embeddings and char embeddings for a full understanding of each word. The final output of our encoder layer is shown as below,

$$u_t^Q = \left[\mathrm{BiLSTM}_Q([e_t^Q, c_t^Q]), c_t^Q\right] \quad (1)$$

1224

$$u_t^P = \left[\mathrm{BiLSTM}_P([e_t^P, c_t^P]), c_t^P\right] \quad (2)$$

200

where we further concat the output of the contextual Bi-LSTM network with the pre-trained char embeddings for its good performance (Peters et al., 2018). This can be regarded as a residual connection between word representations in different levels.

3.5 Hierarchical Attention & Fusion Layer

The attention layer is responsible for linking and fusing information from the question and passage representation, which is the most critical in most MRC tasks. It aims to align the question and passage so that we can better locate on the most relevant passage span with respect to the question. We propose a hierarchical attention structure by combining the co-attention and self-attention mechanism in a multi-hop style. Besides, we think that the original representation and the aligned representation via attention can reflect the content semantics in different granularities. Therefore, we also apply a particular fusion function after each attention function, so that different levels of semantics can be better incorporated towards a better understanding.

3.5.1 Co-attention & Fusion

Given the question and passage representation u_t^Q and u_t^P, a soft-alignment matrix S has been built to calculate the shallow semantic similarity between question and passage as follows:

S: 计算浅层的语义相似度 $\in \mathbb{R}^{m \times n}$

$$S_{ij} = \mathrm{Att}(u_t^Q, u_t^P) = \mathrm{ReLU}(W_{\mathrm{lin}}^\top u_t^Q)^\top \cdot \mathrm{ReLU}(W_{\mathrm{lin}}^\top u_t^P) \quad (3)$$

where W_{lin} is a trainable weight matrix.

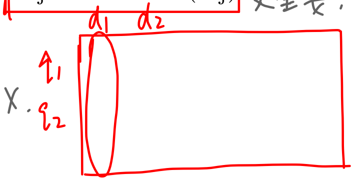
This decomposition avoids the quadratic complexity that is trivially parallelizable (Parikh et al., 2016). Now we use the unnormalized attention weights S_{ij} to compute the attentions between question and passage, which is further used to obtain the attended vectors in passage to question and question to passage direction, respectively.

P2Q Attention signifies which question words are most relevant to each passage word, given as below: 对每个passage word，哪个question word 更重要.

$$\alpha_j = \mathrm{softmax}(S_{:j}) \quad (4)$$

$\alpha \in \mathbb{R}^{m \times n}$

对S求列softmax.



where α_j represents the attention weights on the question words.

The aligned passage representation from question $Q = \left\{u_t^Q\right\}_{t=1}^m$ can thus be derived as, 1224

$Q = m \times d$

$$\tilde{Q}_{:t} = \sum_j \alpha_{tj} \cdot Q_{:j}, \forall j \in [1, ..., m] \quad (5)$$

$\tilde{Q} = \alpha^T \cdot Q = n \times d$

Q2P Attention signifies which passage words have the closest similarity to one of the question words and are hence critical for answering the question. 对每个query word，哪个passage word 更重要

We utilize the same way to calculate this attention as in the passage to question attention (P2Q), except for that in the opposite direction:

对S求行softmax.

$P: n \times d$

$$\beta_i = \mathrm{softmax}(S_{i:}) \in m \times n \quad (6)$$

$$\tilde{P}_{k:} = \sum_i \beta_{ik} \cdot P_{i:}, \forall i \in [1, ..., n] \quad (7)$$

$P: d \times n$

$\tilde{P} = \beta \cdot P = m \times d$

200

where $\tilde{P}$ indicates the weighted sum of the most important words in the passage with respect to the question.

With the aligned passage and question representations $\tilde{Q}$ and $\tilde{P}$ derived, a particular fusion unit has been designed to combine the original contextual representations and the corresponding attention vectors for question and passage separately:

$$\underset{d \times n}{P'} = \mathrm{Fuse}(\underset{d \times n}{P}, \underset{d \times n}{\tilde{Q}}) \quad (8)$$

$$\underset{d \times m}{Q'} = \mathrm{Fuse}(\underset{d \times m}{Q}, \underset{d \times m}{\tilde{P}}) \quad (9)$$

where Fuse$(\cdot, \cdot)$ is a typical fusion kernel.

The simplest way of fusion is a concatenation or addition of the two representations, followed by some linear or non-linear transformation. Recently, a heuristic matching trick with difference and element-wise product is found effective in combining different representations (Mou et al., 2016; Chen et al., 2017b):

$n \times 4d \quad 4d \times d$

$$m(P, \tilde{Q}) = \tanh(\underset{d \times 4d}{W_f} \underbrace{[P; \tilde{Q}; P \circ \tilde{Q}; P - \tilde{Q}]}_{4d \times n} + b_f) \quad (10)$$

$\in \mathbb{R}^{n \times d}$

where $\circ$ denotes the element-wise product, and W_f, b_f are trainable parameters. The output dimension is projected back to the same size as the original representation P or Q via the projected matrix W_f.

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 $\mathrm{m}(\cdot,\cdot)$ with the original contextual representations. As a result, the final fused representations of passage and question can be formulated as:

dxn dxm

$$\mathrm{P}' = \mathrm{g}(\mathrm{P},\tilde{\mathrm{Q}}) \cdot \mathrm{m}(\mathrm{P},\tilde{\mathrm{Q}}) + (1-\mathrm{g}(\mathrm{P},\tilde{\mathrm{Q}})) \cdot \mathrm{P} \quad (11)$$

$$\mathrm{Q}' = \mathrm{g}(\mathrm{Q},\tilde{\mathrm{P}}) \cdot \mathrm{m}(\mathrm{Q},\tilde{\mathrm{P}}) + (1-\mathrm{g}(\mathrm{Q},\tilde{\mathrm{P}})) \cdot \mathrm{Q} \quad (12)$$

where $\mathrm{g}(\cdot,\cdot)$ is a gating function. To capture the relation between the representations in different granularities, we also design a scalar-based, a vector-based and a matrix-based sigmoid gating function, which are compared in Section 4.5.

3.5.2 Self-attention & Fusion

Borrowing the idea from wide and deep network (Cheng et al., 2016), manual features have also been added to combine with the outputs of previous layer for a more comprehensive representation. In our model, these features are concatenated with the refined question-aware passage representation as below:

手工特征

$$\mathrm{D} = \mathrm{BiLSTM}([\mathrm{P}';\mathrm{feat}_{\mathrm{man}}]) \quad (13)$$

where $\mathrm{feat}_{\mathrm{man}}$ denotes the word-level manual passage features.

In this layer, we separately consider the semantic representations of question and passage, and further refine the obtained information from the co-attention layer. Since fusing information among context words allows contextual information to flow close to the correct answer, the self-attention layer is used to further align the question and passage representation against itself, so as to keep the global sequence information in memory. Benefiting from the advantage of self-alignment attention in addressing the long-distance dependence (Wang et al., 2017), we adopt a self-alignment fusion process in this level. To allow for more freedom of the aligning process, we introduce a bilinear self-alignment attention function on the passage representation:

$$\underbrace{\mathrm{L}}_{n\times n} = \mathrm{softmax}(\overset{n\times 200}{\mathrm{D}} \cdot \underbrace{\mathrm{W}_1}_{200\times 200} \cdot \overset{200\times n}{\mathrm{D}^\top}) \quad (14)$$

$$\underbrace{\tilde{\mathrm{D}}}_{n\times 200} = \underbrace{\mathrm{L}}_{n\times n} \cdot \underbrace{\mathrm{D}}_{n\times 200} \quad (15)$$

Another fusion function $\mathrm{Fuse}(\cdot,\cdot)$ is again adopted to combine the question-aware passage representation D and self-aware representation $\tilde{\mathrm{D}}$, as below:

$$\mathrm{D}' = \mathrm{Fuse}(\mathrm{D},\tilde{\mathrm{D}}) \quad (16)$$

nx200 nx200

Finally, a bidirectional LSTM is used to get the final contextual passage representation:

D''是最终的上下文段落表达

$$\mathrm{D}'' = \mathrm{BiLSTM}(\mathrm{D}') \quad (17)$$

As for question side, since it is generally shorter in length and could be adequately represented with less information, we follow the question encoding method used in (Chen et al., 2017a) and adopt a linear transformation to encode the question representation to a single vector.

First, another contextual bidirectional LSTM network is applied on top of the fused question representation: $\mathrm{Q}'' = \mathrm{BiLSTM}(\mathrm{Q}')$. Then we aggregate the resulting hidden units into one single question vector, with a linear self-alignment:

batch, m, 1
m, 200
200, 1

$$\boldsymbol{\gamma} = \mathrm{softmax}(\mathbf{w}_{\mathrm{q}}^\top \cdot \mathrm{Q}'') \in 1\times m \quad (18)$$

$$\mathbf{q} = \sum_{\mathrm{j}} \gamma_{\mathrm{j}} \cdot \mathrm{Q}''_{:\mathrm{j}}, \forall \mathrm{j} \in [1, ..., \mathrm{m}] = \gamma \cdot Q''^{T} \in 1\times d \quad (19)$$

batch, m, 1

where $\mathbf{w}_{\mathrm{q}}$ is a weight vector to learn, we self-align the refined question representation to a single vector according to the question self-attention weight, which can be further used to compute the matching with the passage words.

Q''
batch m, 200

3.6 Model & Output Layer

Instead of predicting the start and end positions based only on D″, a top-level bilinear match function is used to capture the semantic relation between question **q** and paragraph D″ in a matching style, which actually works as a multi-hop matching mechanism.

双线性匹配函数，可以视为多跳匹配机制。

Different from the co-attention layer that generates coarse candidate answers and the self-attention layer that focus the relevant context of passage to a certain intent of question, the top model layer uses a bilinear matching function to capture the interaction between outputs from previous layers and finally locate on the right answer span.

n·200 200x1
$D''\cdot xW$
$xW\cdot D''^{T}$
1x200 200xn
batch, 1xn

The start and end distribution of the passage words are calculated in a bilinear matching way as below,

$$\mathrm{P}_{\mathrm{start}} = \mathrm{softmax}(\overset{1\times 200}{\mathbf{q}} \cdot \overset{200\times 200}{\mathrm{W}_{\mathrm{s}}^\top} \cdot \overset{n\times 200}{\mathrm{D}''}) \quad (20)$$

$$\mathrm{P}_{\mathrm{end}} = \mathrm{softmax}(\mathbf{q} \cdot \mathrm{W}_{\mathrm{e}}^\top \cdot \mathrm{D}'') \quad (21)$$

1x200

where W_{s} and W_{e} are trainable matrices of the bilinear match function.

The output layer is application-specific, in MRC task, we use pointer networks to predict the

batch, n, 200

batch, n, 1
200, 3

batch, 1, 200

start and end position of the answer, since it requires the model to find the sub-phrase of the passage to answer the question.

In training process, with cross entropy as metric, the loss for start and end position is the sum of the negative log probabilities of the true start and end indices by the predicted distributions, averaged over all examples:

损失函数:

$$L(\theta) = -\frac{1}{N}\sum_{i}^{N} \log p_s(y_i^s) + \log p_e(y_i^e) \qquad (22)$$

where θ is the set of all trainable weights in the model, and p_s is the probability of start index, p_e is the probability of end index, respectively. y_i^s and y_i^e are the true start and end indices.

预测时: During prediction, we choose the answer span with the maximum value of $p_s \cdot p_e$ under a constraint that $s \leq e \leq s + 15$, which is selected via a dynamic programming algorithm in linear time.

4 Experiments

In this section, we first present the datasets used for evaluation. Then we compare our end-to-end Hierarchical Attention Fusion Networks with existing machine reading models. Finally, we conduct experiments to validate the effectiveness of our proposed components. We evaluate our model on the task of question answering using recently released SQuAD and TriviaQA Wikipedia (Joshi et al., 2017), which have gained a huge attention over the past year. An adversarial evaluation for the Stanford Question Answering SQuAD is also used to demonstrate the robust of our model under adversarial attacks (Jia and Liang, 2017).

4.1 Dataset

We focus on the SQuAD dataset to train and evaluate our model. SQuAD is a popular machine comprehension dataset consisting of 100,000+ questions created by crowd workers on 536 Wikipedia articles. Each context is a paragraph from an article and the answer to each question is guaranteed to be a span in the context. The answer to each question is always a span in the context. The model is given a credit if its answer matches one of the human chosen answers. Two metrics are used to evaluate the model performance: Exact Match (EM) and a softer metric F1 score, which measures the weighted average of the precision and recall rate at a character level.

Table 1: The performance of our SLQA model and competing approaches on SQuAD.

	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

TriviaQA is a newly available machine comprehension dataset consisting of over 650K context-query-answer triples. The contexts are automatically generated from either Wikipedia or Web search results. The length of contexts in TriviaQA (average 2895 words) is much more longer than the one in SQuAD (average 122 words).

4.2 Training Details

We use the AdaMax optimizer, with a mini-batch size of 32 and initial learning rate of 0.002. A dropout rate of 0.4 is used for all LSTM layers. To directly optimize our target against the evaluation metrics, we further fine-tune the model with some well-defined strategy. During fine-tuning, Focal Loss (Lin et al., 2017) and Reinforce Loss which take F1 score as reward are incorporated with Cross Entropy Loss. The training process takes roughly 20 hours on a single Nvidia Tesla M40 GPU. We also train an ensemble model consisting of 15 training runs with the identical framework and hyper-parameters. At test time, we choose the answer with the highest sum of confidence scores amongst the 15 runs for each question.

4.3 Main Results

The results of our model and competing approaches on the hidden test set are summarized in Table 1. The proposed SLQA+ ensemble model achieves an EM score of 82.4 and F1 score of 88.6, outperforming all previous approaches, which validates the effectiveness of our hierarchical attention and fusion network structure.

We also conduct experiments on the adversarial

Table 2: The F1 scores of different models on AddSent and AddOneSent datasets (S: Single Model, E: Ensemble).

Model	**AddSent**	**AddOneSent**
Logistic (Rajpurkar et al., 2016)	23.2	30.4
Match-S (Wang and Jiang, 2016)	27.3	39.0
Match-E (Wang and Jiang, 2016)	29.4	41.8
BiDAF-S (Seo et al., 2016)	34.3	45.7
BiDAF-E (Seo et al., 2016)	34.2	46.9
ReasoNet-S (Shen et al., 2017)	39.4	50.3
ReasoNet-E (Shen et al., 2017)	39.4	49.8
Mnemonic-S (Hu et al., 2017)	46.6	56.0
Mnemonic-E (Hu et al., 2017)	46.2	55.3
QANet-S (Yu et al., 2018)	45.2	55.7
FusionNet-E (Huang et al., 2017)	51.4	60.7
SLQA-S (our)	**52.1**	**62.7**
SLQA-E (our)	**54.8**	**64.2**

SQuAD dataset (Jia and Liang, 2017) to study the robustness of the proposed model. In the dataset, one or more sentences are appended to the original SQuAD context, aiming to mislead the trained models. We use exactly the same model as in our SQuAD dataset, the performance comparison result is shown in Table 2. It can be seen that the proposed model can still get superior results than all the other competing approaches.

4.4 Ablations

In order to evaluate the individual contribution of each model component, we run an ablation study. Table 3 shows the performance of our model and its ablations on SQuAD dev set. The bi-linear alignment plus fusion between passage and question is most critical to the performance on both metrics which results in a drop of nearly 15%. The reason may be that in top-level attention layer, the similar semantics between question and passage are strong evidence to locate the correct answer span. The ELMo accounts for about 5% of the performance degradation, which clearly shows the effectiveness of language model. We conjecture that language model layer efficiently encodes different types of syntactic and semantic information about words-in-context, and improves the task performance. To evaluate the performance of hierarchical architecture, we reduce the multi-hop fusion with the standard LSTM network. The result shows that multi-hop fusion outperforms the standard LSTM by nearly 5% on both metrics.

4.5 Fusion Functions 不同的Fusion函数

In this section, we experimentally demonstrate how different choices of the fusion kernel impact the performance of our model. The compared fusion kernels are described as follows:

Simple Concat: a simple concatenation of two channel inputs.

Table 3: Ablation tests of SLQA single model on the SQuAD dev set.

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

Table 4: Comparison of different fusion kernels on the SQuAD dev set.

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

Full Projection: the heuristic matching and projecting function as in Equ. 10. $m(x,y)=\tanh(W[x;y;x\circ y;x-y]+b)$

Scalar-based Fusion: the gating function is a trainable scalar parameter (a coarse fusion level):

$$g(P,\tilde{Q}) = g_p \tag{23}$$

where g_p is a trainable scalar parameter.

Vector-based Fusion: the gating function contains a weight vector to learn, which acts as a one-dimensional sigmoid gating,

性能最好.

$$\overset{\in 1\times n}{\boxed{g(P,\tilde{Q})}} = \sigma(\underbrace{\mathbf{w}_g^\top}_{1\times 4d}\cdot\underbrace{[P;\tilde{Q};P\circ\tilde{Q};P-\tilde{Q}]}_{4d\times n}+b_g) \tag{24}$$

where $\mathbf{w}_g$ is trainable weight vector, b_g is trainable bias, and σ is sigmoid function.

Matrix-based Fusion: the gating function contains a weight matrix to learn, which acts as a two-dimensional sigmoid gating,

$$g(P,\tilde{Q}) = \sigma(W_g^\top\cdot[P;\tilde{Q};P\circ\tilde{Q};P-\tilde{Q}]+b_g) \tag{25}$$

where W_g is a trainable weight matrix.

The comparison results of different fusion kernels can be found in Table 4. We can see that different fusion methods contribute differently to the final performances, and the vector-based fusion method performs best, with a moderate parameter size.

4.6 Attention Hierarchy and Function

In the proposed model, attention layer is the most important part of the framework. At the bottom of Table 5 we show the performances on SQuAD

Table 5: Comparison of different attention styles on the SQuAD dev set.

Attention Hierarchy	**EM / F1**
1-layer attention (only qp co-attention)	61.9 / 68.4
2-layer attention (add self-attention)	65.4 / 71.7
3-layer attention (add bilinear match)	**80.0 / 87.0**
Attention Function	**EM / F1**
dot product	62.9 / 69.3
linear attention	78.0 / 84.9
bilinear attention (linear + relu)	**80.0 / 87.0**
trilinear attention	78.9 / 85.8

Table 6: Published and unpublished results on the TriviaQA wikipedia leaderboard.

	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**

for four common attention functions. Empirically, we find bilinear attention which add ReLU after linearly transforming does significantly better than the others.

At the top of Table 5 we show the effect of varying the number of attention layers on the final performance. We see a steep and steady rise in accuracy as the number of layers is increased from N = 1 to 3.

4.7 Experiments on TriviaQA

To further examine the robustness of the proposed model, we also test the model performance on TriviaQA dataset. The test performance of different methods on the leaderboard (on Jan. 12th 2018) is shown in Table 6. From the results, we can see that the proposed model can also obtain state-of-the-art performance in the more complex TriviaQA dataset.

5 Conclusions

We introduce a novel hierarchical attention network, a state-of-the-art reading comprehension model which conducts attention and fusion horizontally and vertically across layers at different levels of granularity between question and paragraph. We show that our proposed method is very powerful and robust, which outperforms the previous state-of-the-art methods in various large-scale golden MRC datasets: SQuAD, TriviaQA, AddSent and AddOneSent.

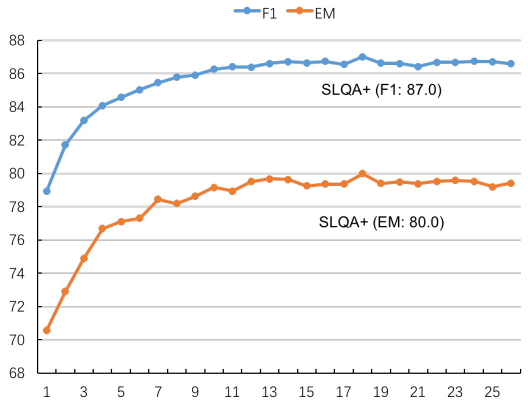


Figure 2: Learning curve of F1 / EM score on the SQuAD dev set

Acknowledgments

We thank the Stanford NLP Group and the University of Washington NLP Group for evaluating our results on the SQuAD and the TriviaQA test set.

References

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, and Diana Inkpen. 2017b. Natural language inference with external knowledge. *arXiv preprint arXiv:1711.04289*.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pages 7–10. ACM.

Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2016. Attention-over-attention neural networks for reading comprehension. *arXiv preprint arXiv:1607.04423*.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Minghao Hu, Yuxing Peng, and Xipeng Qiu. 2017. Reinforced mnemonic reader for machine comprehension. *CoRR, abs/1705.02798*.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *arXiv preprint arXiv:1711.07341*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*.

Rui Liu, Wei Wei, Weiguang Mao, and Maria Chikina. 2017a. Phase conductor on multi-layered attentions for machine comprehension. *arXiv preprint arXiv:1710.10504*.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2017b. Stochastic answer networks for machine reading comprehension. *arXiv preprint arXiv:1712.03556*.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 130.

Boyuan Pan, Hao Li, Zhou Zhao, Bin Cao, Deng Cai, and Xiaofei He. 2017. Memen: Multi-layer embedding with memory networks for machine comprehension. *arXiv preprint arXiv:1707.09098*.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1047–1055. ACM.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 189–198.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.