Original Semantics-Oriented Attention and Deep Fusion Network for Sentence Matching

Mingtong Liu, Yujie Zhang, Jinan Xu, Yufeng Chen
School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
{16112075,yjzhang,jaxu,chenyf}@bjtu.edu.cn

Abstract

Sentence matching is a key issue in natural language inference and paraphrase identification. Despite the recent progress on multi-layered neural network with cross sentence attention, one sentence learns attention to the intermediate representations of another sentence, which are propagated from preceding layers and therefore are uncertain and unstable for matching, particularly at the risk of error propagation. In this paper, we present an original semantics-oriented attention and deep fusion network (OSOA-DFN) for sentence matching. Unlike existing models, each attention layer of OSOA-DFN is oriented to the original semantic representation of another sentence, which captures the relevant information from a fixed matching target. The multiple attention layers allow one sentence to repeatedly read the important information of another sentence for better matching. We then additionally design deep fusion to propagate the attention information at each matching layer. At last, we introduce a self-attention mechanism to capture global context to enhance attention-aware representation within each sentence. Experiment results on three sentence matching benchmark datasets SNLI, SciTail and Quora show that OSOA-DFN has the ability to model sentence matching more precisely.

1 Introduction

Natural language sentence matching is a key technique of comparing two sentences and identifying the semantic relationship between them, which is usually viewed as a classification problem (Wang et al., 2017). The technique has applications in natural language inference to judge whether a hypothesis sentence can be inferred from a premise sentence (Bowman et al., 2015) and in paraphrase identification to determine whether two sentences express the equivalent meaning or not (Yin et al., 2015). The core issue for sentence matching is to model the relatedness between two sentences (Rocktäschel et al., 2015; Parikh et al., 2016; Wang et al., 2017; Duan et al., 2018).

Recently, neural network-based models for sentence matching have attracted more attention for their powerful ability to learn sentence representation (Bowman et al., 2015; Wang et al., 2017; Duan et al., 2018). There are mainly two types of frameworks: sentence encoding based framework and attention-based framework. For the first type of framework, a simple and effective model is proposed by using two sentence vectors (Bowman et al., 2015), but the interaction between two sentences is neglected. For the second type of framework, attention mechanism is introduced to model word-level interaction between two sentences and a higher accuracy is achieved (Rocktäschel et al., 2015; Parikh et al., 2016; Wang et al., 2017). Particularly, multi-layered deep matching network with attention shows that deeper models outperform shallower models (Duan et al., 2018).

However, the existing attention mechanism still has some limitations. When one sentence learns attention to another sentence, the attention is performed between two parallel layers and oriented to the intermediate representations from the preceding layer of another one. As a result, semantics to be paid attention are uncertain and unstable for matching because semantics are changed at different layers. On the other hand, the intermediate representations tend to be affected by error propagation in multi-layered attentions, in which if the first attention aligns the wrong position, the second attention will now have the incorrect information as input for alignment.

In order to address these problems, we propose an original semantics-oriented attention and deep fusion network (OSOA-DFN) for sentence matching. OSOA-DFN mainly consists of three sub-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2652–2661, Hong Kong, China, November 3–7, 2019.

components: (1) original semantics-oriented cross sentence attention; (2) deep fusion; and (3) self-attention mechanism. The cross sentence attention is oriented to the original semantic representations of another sentence, so as to be able to capture inherent semantics by relying on the fixed matching target. The multiple cross attention operations allow one sentence to repeatedly read the important information of another sentence for better interaction. We then design a deep fusion in addition to usual fusion to augment the propagation of attention information at each matching layer. The self-attention mechanism is also introduced at the last to capture global context to enhance attention-aware representation within each sentence. Experiment results demonstrate that OSOA-DFN has the ability to model sentence matching more precisely on the SNLI, the SciTail, and the Quora datasets.

Our contributions can be summarized as follows:

- We pay attention to the original semantic representations for cross sentence interaction and the matching target of attention for a certain sentence is therefore ensured to be fixed in spite of multiple layers. The multiple cross attention operations allow one sentence to repeatedly read the important information of another sentence for better interaction.
- We design a deep fusion in addition to usual fusion to augment the propagation of attention information for matching, and introduce a self-attention mechanism at the last to capture global context to enhance attention-aware representation within each sentence.
- We evaluate our model on three challenging datasets and show that the proposed model has the ability to model sentence matching more precisely and significantly improves the performance.

2 General Neural Attention-Based Model for Sentence Matching

Formally, we can define the sentence matching as follows. Given two sentences $P = [p_1, \cdots, p_i, \cdots, p_m]$ and $Q = [q_1, \cdots, q_j, \cdots, q_n]$, the goal is to predict a label $y^* \in \mathcal{Y}$, where $\mathcal{Y}$ = {entailment, contradiction, neutral} in natural language inference and $\mathcal{Y}$ = {0,1} in paraphrase identification, indicating the logic semantic relationship between two sentences P and Q (Wang et al., 2017).

$$y^* = \arg\max_{y \in \mathcal{Y}} P_{\mathrm{r}}(\mathrm{y}|P, Q) \quad (1)$$

Generally, the architecture of neural attention-based models for sentence matching includes three components (Wang et al., 2017; Duan et al., 2018): (1) **input encoding layer** encodes each sentence into semantic representation; (2) **attention-based matching layer** models word-level alignment between two sentences and produces attention-aware representation for each sentence; and (3) **prediction layer** predicts the semantic relation between two sentences. Figure 1(a) illustrates the general model.

2.1 Input Encoding Layer

For the given sentence pairs $P = [p_1, \cdots, p_i, \cdots, p_m]$ and $Q = [q_1, \cdots, q_j, \cdots, q_n]$, where p_i and q_j indicate the i-th and j-th word in P and Q respectively, the input encoding layer first converts words of P and Q into vectors $[\mathbf{e}_{p_1}, \cdots, \mathbf{e}_{p_i}, \cdots, \mathbf{e}_{p_m}]$ and $[\mathbf{e}_{q_1}, \cdots, \mathbf{e}_{q_j}, \cdots, \mathbf{e}_{q_n}]$ by looking up M respectively, where $M \in \mathbf{R}^{d \times |V|}$ is the embedding table. d is the dimension of embeddings and $|V|$ is the size of the vocabulary.

In order to encode contextual information into word representations, we use a BiLSTM neural network (Hochreiter and Schmidhuber, 1997) to encode two sentences P and Q. The sequential BiLSTM calculates a new hidden state conditioned on the previous states to incorporate contextual information, and several previous works have shown its effectiveness for sentence matching (Rocktäschel et al., 2015; Wang et al., 2017; Duan et al., 2018).

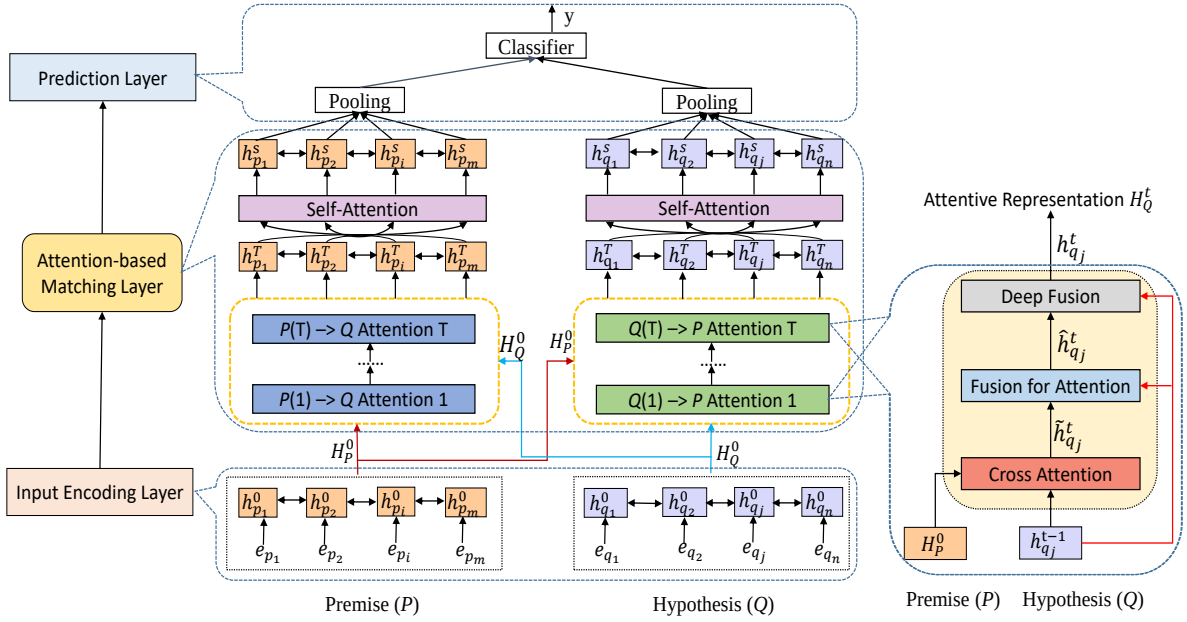
$$\mathbf{h}^0_{p_i} = \mathrm{BiLSTM}(\mathbf{e}_{p_i}, \overrightarrow{\mathbf{h}}^0_{p_{i-1}}, \overleftarrow{\mathbf{h}}^0_{p_{i+1}}) \quad (2)$$

$$\mathbf{h}^0_{q_j} = \mathrm{BiLSTM}(\mathbf{e}_{q_j}, \overrightarrow{\mathbf{h}}^0_{q_{j-1}}, \overleftarrow{\mathbf{h}}^0_{q_{j+1}}) \quad (3)$$

Then the two sentences are converted to $\mathbf{H}^0_P = [\mathbf{h}^0_{p_1}, \cdots, \mathbf{h}^0_{p_i}, \cdots, \mathbf{h}^0_{p_m}]$ and $\mathbf{H}^0_Q = [\mathbf{h}^0_{q_1}, \cdots, \mathbf{h}^0_{q_j}, \cdots, \mathbf{h}^0_{q_n}]$. Hereafter, we call $\mathbf{H}^0_P$ and $\mathbf{H}^0_Q$ as original semantic representations of sentences P and Q respectively. In this paper, we will use them as the targets of cross sentence attention.

2.2 Attention-Based Matching Layer

Generally, this layer employs the attention mechanism to model the interaction information between



(a) The General Model (b) The Architecture of Our Proposed OSOA-DFN (c) A Cross Attention Unit $Q(t)\rightarrow P$

Figure 1: (a) is the general model for sentence matching. (b) is an overview architecture of our proposed OSOA-DFN. (c) is an original semantics-oriented cross attention unit that learns the interaction information from the original semantic representation of another sentence.

two sentences. It can be formulated as:

$$\mathbf{V}_P = \text{Match}(\mathbf{H}_P^0, \mathbf{H}_Q^0), \mathbf{V}_Q = \text{Match}(\mathbf{H}_Q^0, \mathbf{H}_P^0) \quad (4)$$

where Match$(\cdot)$ is a neural attention-based matching function, $\mathbf{V}_P = [\mathbf{v}_{p_1}, \cdots, \mathbf{v}_{p_i}, \cdots, \mathbf{v}_{p_m}]$ and $\mathbf{V}_Q = [\mathbf{v}_{q_1}, \cdots, \mathbf{v}_{q_j}, \cdots, \mathbf{v}_{q_n}]$ are new attention-aware representations for P and Q, respectively. This layer is the core layer for sentence matching. Match$(\cdot)$ is mainly focused by researches and some effective frameworks are proposed (Rocktäschel et al., 2015; Wang et al., 2017; Duan et al., 2018). In this paper, we also focus on this layer, and propose an original semantics-oriented attention and deep fusion network. The details will be described in Section 3.

2.3 Prediction Layer

A pooling layer is used to convert the resulting representations of all position in P and Q into a fixed-length vector and feed it into a classifier to determine the semantic relationship between the two sentences.

A mean pooling is usually adopted on each sentence for capturing all of the information and also a max pooling for highlighting the significant properties. In this paper, we get a fixed dimensional representation $\mathbf{V}$ by concatenating them together as (Chen et al., 2017; Duan et al., 2018).

$$\mathbf{V}_{P_{mean}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{v}_{p_i}, \mathbf{V}_{P_{max}} = \max_{i=1}^{m}\mathbf{v}_{p_i} \quad (5)$$

$$\mathbf{V}_{Q_{mean}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_{q_j}, \mathbf{V}_{Q_{max}} = \max_{j=1}^{n}\mathbf{v}_{q_j} \quad (6)$$

$$\mathbf{V} = [\mathbf{V}_{P_{mean}}; \mathbf{V}_{P_{max}}; \mathbf{V}_{Q_{mean}}; \mathbf{V}_{Q_{max}}] \quad (7)$$

Finally, we pass representation $\mathbf{V}$ into a multilayer perceptron (MLP) classifier to calculate the probability $P_r(\cdot)$ of each label.

$$P_r(\cdot|P,Q) = \text{softmax}(\mathbf{W}_2\text{ReLU}(\mathbf{W}_1\mathbf{V}+\mathbf{b}_1)+\mathbf{b}_2) \quad (8)$$

where, $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$ are learnable parameters.

3 Original Semantics-Oriented Attention and Deep Fusion Network

In this paper, we mainly focus on the structure of attention-based matching layer. Inspired by the recent successful deep models (He et al., 2016; Duan et al., 2018), we propose an original semantics-oriented attention and deep fusion network (OSOA-DFN) for sentence matching, as shown in Figure 1(b). OSOA-DFN is mainly composed of: (1) original semantics-oriented cross sentence attention; (2) deep fusion; and (3) self-attention mechanism. (1) and (2) are combined to

form one unit of cross attention, as shown in Figure 1(c), and there are T units in attention-based matching layer. Finally, one layer of self-attention is introduced after the T units of cross attention.

3.1 Original Semantics-Oriented Cross Sentence Attention

Cross sentence attention is utilized to model the relevance between two sentences. In the t-th attention layer, we use $P(t)\rightarrow Q$ to annotate that the sentence P learns attention to the sentence Q to extract the relevant information from Q.

Given the representations of P and Q: $\mathbf{H}_P^{t-1} = [\mathbf{h}_{p_1}^{t-1}, \cdots, \mathbf{h}_{p_i}^{t-1}, \cdots, \mathbf{h}_{p_m}^{t-1}]$ and $\mathbf{H}_Q^0 = [\mathbf{h}_{q_1}^0, \cdots, \mathbf{h}_{q_j}^0, \cdots, \mathbf{h}_{q_n}^0]$, each cross attention $P(t)\rightarrow Q$ will use the original semantics $\mathbf{H}_Q^0$ of Q for interaction, where $t = \{1, \cdots, T\}$ and $t = 1$ represents P using the original representation $\mathbf{H}_P^0$. We first compute the unnormalized attention weights as the similarity of $P(t)$ and Q, the alignment matrix $\mathbf{A}^t \in \mathbf{R}^{m\times n}$ is defined as follows:

$$\mathbf{A}_{ij}^t = \mathbf{h}_{p_i}^{t-1^T}\mathbf{W}^t\mathbf{h}_{q_j}^0 + \langle \mathbf{U}_p^t, \mathbf{h}_{p_i}^{t-1}\rangle + \langle \mathbf{U}_q^t, \mathbf{h}_{q_j}^0\rangle \quad (9)$$

where $\mathbf{W}^t \in \mathbf{R}^{h\times h}$, $\mathbf{U}_p^t$, $\mathbf{U}_q^t \in \mathbf{R}^h$ are learnable parameters, and $\langle\cdot,\cdot\rangle$ denotes the inner production operation. p_i and p_j are the the i-th and j-th word in the P and Q respectively. Next, the semantics of sentence Q related to $\mathbf{h}_{p_i}^{t-1}$ is extracted to compute $\widetilde{\mathbf{h}}_{p_i}^t$ according to $\mathbf{A}^t$, as shown in Equation (10).

$$\widetilde{\mathbf{h}}_{p_i}^t = \sum_{j=1}^{n} \frac{exp(\mathbf{A}_{ij}^t)}{\sum_{k=1}^{n} exp(\mathbf{A}_{ik}^t)}\mathbf{h}_{q_j}^0 \quad (10)$$

Intuitively, $\widetilde{\mathbf{h}}_{p_i}^t$ is a representation by using attentive information in $\mathbf{H}_Q^0$ that is softly aligned to $\mathbf{h}_{p_i}^{t-1}$, and the semantics of $\mathbf{H}_Q^0$ is more probably selected if it is more related to $\mathbf{h}_{p_i}^{t-1}$.

3.2 Deep Fusion

To further enrich the interaction, we first perform an usual fusion and then design a deep fusion for each cross attention to augment the propagation of attention information.

The usual fusion (Wang and Jiang, 2016a; Duan et al., 2018) can be formulated as the Equations (11) - (13).

$$\overline{\mathbf{h}}_{p_i}^t = [\mathbf{h}_{p_i}^{t-1}; \widetilde{\mathbf{h}}_{p_i}^t; |\, \mathbf{h}_{p_i}^{t-1} - \widetilde{\mathbf{h}}_{p_i}^t \,|; \mathbf{h}_{p_i}^{t-1} \odot \widetilde{\mathbf{h}}_{p_i}^t] \quad (11)$$

$$\widetilde{\mathbf{h}}_{p_i}^t = \mathrm{ReLU}(\mathbf{W}_{\mathrm{h}}^{\mathrm{t}}\overline{\mathbf{h}}_{\mathrm{p_i}}^{\mathrm{t}} + \mathbf{b}_{\mathrm{h}}^{\mathrm{t}}) \quad (12)$$

$$\widehat{\mathbf{h}}_{p_i}^t = \mathrm{BiLSTM}(\widetilde{\mathbf{h}}_{p_i}^t, \overrightarrow{\mathbf{h}}_{p_{i-1}}^t, \overleftarrow{\mathbf{h}}_{p_{i+1}}^t) \quad (13)$$

where $[\cdot;\cdot;\cdot;\cdot]$ refers to the concatenation operation. In matching operation, the concatenation can retain all the information (Wang and Jiang, 2016a; Chen et al., 2017). We use a neural nonlinear transformation ReLU (Glorot et al., 2011) as local comparison function. This operation helps the model to better fuse the attention information and also reduce the complexity of vector representation. Since the understanding of some word-level alignments may rely on the contextual matching information, we then apply a BiLSTM to incorporate the sequential matching information, which further gathers interactive features between two sentences.

We design the deep fusion layer as follows. A gated connection layer is used to learn adaptively controlling how much information to be stored and carried to the next attention layer. It can be formulated as Equations (14) - (17):

$$\mathbf{r}_{p_i}^t = \sigma(\mathbf{W}_r^t[\mathbf{h}_{p_i}^{t-1}; \widehat{\mathbf{h}}_{p_i}^t; \mathbf{h}_{p_i}^{t-1} \odot \widehat{\mathbf{h}}_{p_i}^t] + \mathbf{b}_r^t) \quad (14)$$

$$\mathbf{z}_{p_i}^t = \sigma(\mathbf{W}_z^t[\mathbf{h}_{p_i}^{t-1}; \widehat{\mathbf{h}}_{p_i}^t; \mathbf{h}_{p_i}^{t-1} \odot \widehat{\mathbf{h}}_{p_i}^t] + \mathbf{b}_z^t) \quad (15)$$

$$\widetilde{\mathbf{c}}_{p_i}^t = \tanh(\mathbf{W}_c^t[\mathbf{r}_{p_i}^t \odot \mathbf{h}_{p_i}^{t-1}; \widehat{\mathbf{h}}_{p_i}^t] + \mathbf{b}_c^t) \quad (16)$$

$$\mathbf{h}_{p_i}^t = \mathbf{z}_{p_i}^t \odot \mathbf{h}_{p_i}^{t-1} + (1 - \mathbf{z}_{p_i}^t) \odot \widetilde{\mathbf{c}}_{p_i}^t \quad (17)$$

where $\mathbf{W}_*^t$ and $\mathbf{b}_*^t$ are the learnable parameters, $\mathbf{h}_{p_i}^t$ is the result of current layer, $\mathbf{h}_{p_i}^{t-1}$ is the result from preceding layer, σ is a sigmoid function, the value of $\mathbf{r}_{p_i}^t$ and $\mathbf{z}_{p_i}^t$ is between 0 and 1. Intuitively, the model can learn to set the $\mathbf{r}_{p_i}^t$ and $\mathbf{z}_{p_i}^t$ close to 1, thus the more attention information from the preceding layers will be propagated to the following attention layers for matching, and close to 0 implying that the information of preceding layers is less propagated.

After the t-th layer of the original semantics-oriented cross attention, each word p_i in sentence P is newly represented by $\mathbf{h}_{p_i}^t$. Similarly, we conduct cross attention for $Q(t)\rightarrow P$, implying that the sentence Q learns attention to the sentence P, which will be oriented to the original semantic representation $\mathbf{H}_P^0$ of P to derive the attention-aware representation $\mathbf{h}_{q_j}^t$ for each word q_j of Q.

3.3 Self-Attention Mechanism

We additionally introduce a self-attention mechanism after cross sentence attention. It captures long-distance context information to learn word representation within each sentence and further enhances the attention-aware representation.

For sentence P, its attentive representation $\mathbf{H}_p^T = [\mathbf{h}_{p_1}^T, \cdots, \mathbf{h}_{p_i}^T, \cdots, \mathbf{h}_{p_m}^T]$ is computed after T layers of original semantics-oriented cross sentence attention. We first compute a self-attention matrix $\mathbf{S}^s \in \mathbf{R}^{m \times m}$ as Equation (9).

$$\mathbf{S}_{ij}^s = \langle \mathbf{h}_{p_i}^T, \mathbf{h}_{p_j}^T \rangle \tag{18}$$

where, $\mathbf{S}_{ij}^s$ indicates the relevance between the i-th word and j-th word in P. Then, the self-attention vector for each word in P is computed as follows:

$$\widetilde{\mathbf{h}}_{p_i}^s = \sum_{j=1}^{m} \frac{exp(\mathbf{S}_{ij}^s)}{\sum_{k=1}^{m} exp(\mathbf{S}_{ik}^s)} \mathbf{h}_{p_j}^T \tag{19}$$

Intuitively, $\widetilde{\mathbf{h}}_{p_i}^s$ augments each word representation with global context of the sentence P.

After that, an usual fusion augmented by deep fusion, as described in Section 3.2, is also introduced to further enhance the self-attention information within each sentence as follows:

$$\overline{\mathbf{h}}_{p_i}^s = [\mathbf{h}_{p_i}^T; \widetilde{\mathbf{h}}_{p_i}^s; | \mathbf{h}_{p_i}^T - \widetilde{\mathbf{h}}_{p_i}^s |; \mathbf{h}_{p_i}^T \odot \widetilde{\mathbf{h}}_{p_i}^s] \tag{20}$$

$$\widetilde{\mathbf{h}}_{p_i}^s = \mathrm{ReLU}(\mathbf{W}_{\mathrm{h}}^{\mathrm{s}} \overline{\mathbf{h}}_{\mathrm{p_i}}^{\mathrm{s}} + \mathbf{b}_{\mathrm{h}}^{\mathrm{s}}) \tag{21}$$

$$\widehat{\mathbf{h}}_{p_i}^s = \mathrm{BiLSTM}(\widetilde{\mathbf{h}}_{p_i}^s, \overrightarrow{\mathbf{h}}_{p_{i-1}}^s, \overleftarrow{\mathbf{h}}_{p_{i+1}}^s) \tag{22}$$

The deep fusion after the self-attention layer is computed as Equations (14) and (17), which fuses the $\mathbf{h}_{p_i}^T$ from original semantics-oriented cross attention and the $\widehat{\mathbf{h}}_{p_i}^s$ from self-attention to get the final attention-ware representation.

Similarly, we conduct self-attention and deep fusion operations to the sentence Q to derive the attention-aware representation $\mathbf{h}_{q_j}^s$ for each word q_j of Q. Then, two sentences are converted to $\mathbf{H}_P^s = [\mathbf{h}_{p_1}^s, \cdots, \mathbf{h}_{p_i}^s, \cdots, \mathbf{h}_{p_m}^s]$ and $\mathbf{H}_Q^s = [\mathbf{h}_{q_1}^s, \cdots, \mathbf{h}_{q_j}^s, \cdots, \mathbf{h}_{q_n}^s]$. Finally, $\mathbf{H}_P^s$ and $\mathbf{H}_Q^s$ are passed into the prediction layer as input $\mathbf{V}_P$ and $\mathbf{V}_Q$ for deciding their semantic relationship.

4 Training

For model training, we employ cross-entropy as the loss function since the goal is the make the correct classification. Considering the model complexity, we also add l_2-norm of all learnable parameters to the final loss function. Finally, the object is to minimize the following objective function $\mathcal{J}(\theta)$, which can be formulated as:

$$\mathcal{J}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log P(y^{(i)} | P^{(i)}, Q^{(i)}; \theta) + \frac{1}{2} \lambda \|\theta\|_2^2 \tag{23}$$

Dataset	Train	Dev	Test	Avg.L		Vocab
SNLI	549K	9.8K	9.8K	14	8	36K
SciTail	23K	1.3K	2.1K	17	12	24K
Quora	384K	10K	10K	12	12	107K

Table 1: Statistics of datasets: SNLI, SciTail and Quora. Avg.L refers to average length of a pair of sentences.

where θ denotes all the learnable parameters of our model, N is the number of instances in the training set, $(P^{(i)}, Q^{(i)})$ are the sentence pairs, and $y^{(i)}$ denotes the corresponding annotated label for the i-th instance.

Word Embedding Following (Tay et al., 2017), to represent each input word, we concatenate three types of vectors: a pre-trained vector, a learnable vector for each word type, and a learnable vector for the POS tag of the word. We use NLTK [1] to acquire POS tags. Finally, we apply a nonlinear transformation ReLU to the concatenated vector to get the final word embedding.

5 Experiments

5.1 Dataset

We evaluate our model on natural language inference and paraphrase identification tasks with three datasets: the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015), the SciTail dataset (Khot et al., 2018), and the Quora Questions Pairs dataset (Quora).

SNLI is a natural language inference dataset (Bowman et al., 2015). The original data set contains 570,152 sentence pairs, each labeled with one of the following relationships: $\mathcal{Y}$ = {entailment, contradiction, neutral}. We follow the same data split as in (Bowman et al., 2015).

SciTail is a binary entailment classification task and $\mathcal{Y}$ = {entailment, neutral}. We have the same data split as in (Khot et al., 2018). Notably, the premise and the corresponding hypothesis have high lexical similarity both for entailed and non-entailed pairs, which makes the task particularly difficult.

Quora consists of over 400,000 question pairs and $\mathcal{Y}$ = {0, 1} indicating whether two questions are paraphrases of each other. We have the same data split as in (Wang et al., 2017).

[1] http://www.nltk.org/

Models	Train	Test
LstmAtt(Rocktäschel et al., 2015)	85.3	83.5
mLSTM (Wang and Jiang, 2016b)	92.0	86.1
LSTMN (Cheng et al., 2016)	88.5	86.3
DecompAtt (Parikh et al., 2016)	89.5	86.8
re-read (Sha et al., 2016)	90.7	87.5
btree-LSTM (Paria et al., 2016)	88.6	87.6
SAN (Im and Cho, 2017)	89.6	86.3
BiMPM (Wang et al., 2017)	90.9	87.5
ESIM (Chen et al., 2017)	92.6	88.0
DIIN (Gong et al., 2017)	91.2	88.0
AF-DMN (Duan et al., 2018)	94.5	88.6
OSOA-DFN (single)	92.3	**88.8**
BiMPM (ensemble)	93.2	88.8
ESIM (ensemble)	93.5	88.6
DIIN (ensemble)	92.3	88.9
AF-DMN (ensemble)	94.9	89.0
OSOA-DFN (ensemble)	93.5	**89.3**

Table 2: Comparative results with previous models on SNLI dataset.

The detailed statistical information of the three datasets is shown in Table 1.

5.2 Implementation Details

We set word embeddings and all of the hidden states of BiLSTMs and MLPs to 300 dimensions. Pre-trained word vectors are 300-dimensional $Glove\ 840B$ (Pennington et al., 2014) and without updating during training. The learnable word vectors and POS vectors have 30 dimensions. For all datasets, there are 3 cross sentence attention layers and 1 self-attention layer. The batch size is set to 64 for SNLI and Quora, 32 for SciTail. We use the Adam method (Kingma and Ba, 2014) for model training. We set the initial learning rate to 5e-4 with a decay ratio of 0.95 for each epoch, and l_2 regularizer strength to 6e-5. To prevent overfitting, we use dropout regularization (Srivastava et al., 2014) with a drop rate of 0.2 for all MLPs.

5.3 Ensemble

The ensemble strategy has been proved to effectively improve model accuracy. Following (Duan et al., 2018), our ensemble model averages the probability distributions from three individual single OSOA-DFNs, and each of them has the same architecture but different parameter initialization.

5.4 Comparison on Natural Language Inference

SNLI We compare our model with the following previous models on SNLI dataset, and show the results in Table 2. LstmAtt (Rocktäschel et al., 2015) extend the general LSTM model with attention mechanism. mLSTM (Wang and Jiang, 2016b) exploit LSTM with memory. DecompAtt (Parikh et al., 2016) propose a decomposable word-by-word matching model with attention, and use pre-trained word vector without relying on any word-order information. SAN (Im and Cho, 2017) is a distance-based self-attention network. BiMPM (Wang et al., 2017) design a bilateral multi-perspective matching model from both directions. ESIM (Chen et al., 2017) incorporate the chain LSTM and tree LSTM. Recently, AF-DMN (Duan et al., 2018) adopt attention-fused deep matching network by using multiple stacked cross attention and self-attention layers.

In Table 2, our single OSOA-DFN achieves 88.8% test accuracy. Moreover, we also report the ensemble result, and the test accuracy is 89.3%. Comparative results show that our model outperforms the previous models on single and ensemble scenarios on SNLI dataset. ELMO (Peters et al., 2018) and BERT (Devlin et al., 2018) have been well known as pre-trained language model for acquiring contextual word vectors. However, our model has less computing complexity (340M parameters in BERT while 10M in our model), but obtained competitive performance. We will conduct the comparison with them in the future. In this paper, we evaluated the contribution of original semantics-oriented cross attention and deep fusion to our model.

Models	Dev	Test
Majority class	63.3	60.3
Ngram	65.0	70.6
DecompAtt	75.4	72.3
ESIM	70.5	70.6
DGEM	79.6	77.3
DEISTE	82.4	82.1
CAFE	-	83.3
AF-DMN (re-imp)	87.2	84.4
OSOA-DFN	**88.9**	**86.8**

Table 3: Comparative results with previous models on SciTail dataset.

SciTail We compare our model with the following previous models on SciTail dataset, and show the results in Table 3. The first five models in Table 3 are all implemented in (Khot et al., 2018). DGEM is a graph based attention model using lingual syntactic structures for improved performance (Khot et al., 2018). CAFE (Tay et al., 2017) improve previous comparison operations by compressing alignment vectors into scalar valued fea-

Models	Test
Siamese-CNN	79.60
Multi-Perspective-CNN	81.38
Siamese-LSTM	82.58
Multi-Perspective-LSTM	83.21
L.D.C	85.55
BiMPM	88.17
AF-DMN	88.72
OSOA-DFN	**89.03**

Table 4: Comparative results with previous models on Quora dataset.

Models	Dev	Test
OSOA-DFN (ori-attention)	88.9	86.8
OSOA-DFN (inter-attention)	87.1	84.2

Table 5: Effect of original semantics-oriented cross sentence attention on SciTail dataset.

Num	Dev	Test
1	86.9	84.0
2	88.6	85.1
3	88.9	86.8
4	89.1	87.2
5	89.2	87.4

Table 6: Effect of cross sentence attention layers on SciTail dataset.

Models	Dev	Test
OSOA-DFN	**88.9**	**86.8**
- Deep fusion	85.8	84.7
- Self-attention	88.1	84.8
- BiLSTM fusion	87.2	83.2

Table 7: Effect of components on SciTail dataset.

tures. DEISTE (Yin et al., 2018) propose deep explorations of inter-sentence interaction. AF-DMN (re-imp) is our re-implementation of the multi-layered attention model in (Duan et al., 2018) that have not reported the results on this dataset.

On this dataset, our single OSOA-DFN significantly outperforms these strong baselines, achieving the state-of-the-art performance with 86.8% accuracy on the test set. It demonstrates that our model has the ability to improve semantic matching on the challenging SciTail dataset.

5.5 Comparison on Paraphrase Identification

Quora We compare our model with the following previous models on Quora dataset, and show the results in Table 4. The Siamese-CNN model and Siamese-LSTM model encode sentences with CNN and LSTM respectively, and then predict the relationship between them based on the cosine similarity (Wang et al., 2017). Multi-Perspective-CNN and Multi-Perspective-LSTM adopt multiple perspective cosine matching function (Wang et al., 2017). L.D.C (Wang et al., 2016) and BiMPM (Wang et al., 2017) adopt attention-based framework that performs word-level matching.

As we can see, our single OSOA-DFN outperforms the baselines and achieves 89.03% accuracy on the test set. The results prove that our model is very effective for paraphrase identification task.

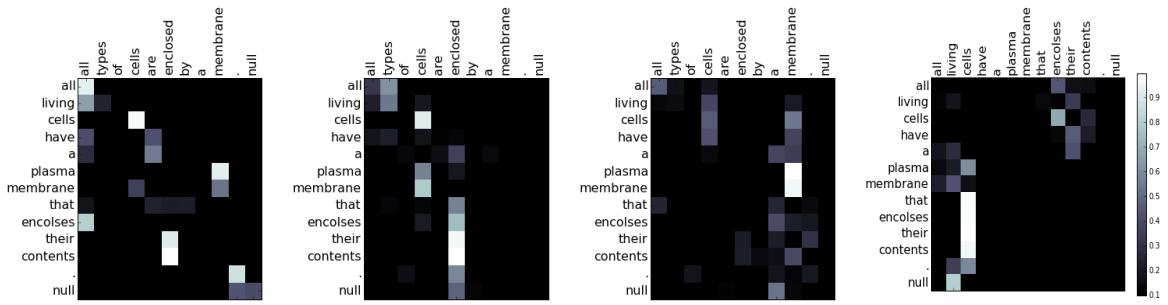
5.6 Effect of Original Semantics-Oriented Cross Sentence Attention

To verify the effect of original semantics-oriented cross sentence attention, we first implement a variant of our model, namely OSOA-DFN (inter-attention), as shown in Table 5. As in the model of (Duan et al., 2018), we make the cross sentence attention oriented to the intermediate representations from the preceding layer of another sentence, where cross attention is performed on parallel layers between two sentences. The results show that our method achieves higher accuracy on the test set of SciTail dataset, which proves the effect of original semantics-oriented cross attention on extracting expressive features from another sentence for semantic matching.

We further verify the effect of the depth of cross sentence attention on performance, as shown in Table 6. As the number of stacked attention layers increases from 1 to 5, we can see that the performance increases both on the development set and the test set of SciTail dataset. But from 3 to 5, the increased accuracy is slower. We can conclude that the multiple original semantics-oriented attentions are effective in improving matching performance. However, the parameters will grow rapidly with the increasing of the number of stacked attention layer, and a large of number of parameters will increase model complexity. Because of computational cost, we just set the number of cross attention layers to 3 in our experiment.

5.7 Effect of Deep Fusion and Self-Attention Mechanism

We verify the effect of deep fusion and self-attention for better understanding of the performance improvement of OSOA-DFN, as shown in Table 7. Only using BiLSTM fusion without deep fusion, the accuracy drops by 2.1% on the test set of SciTail dataset. This indicates the augmented deep fusion at different layers is important in prop-



(a) 1st cross attention. (b) 2nd cross attention. (c) 3rd cross attention. (d) self-attention.

Figure 2: The visualization of alignment matrices in the three cross sentence attention layers and the self-attention layer.

agating attention information through the network for deep interaction. Without the self-attention, the accuracy is degraded to 84.8%. This indicates that the self-attention mechanism is effective in capturing global context information for augmenting attention-aware semantic representation.

We also verify the effect of BiLSTM fusion, without the BiLSTM fusion, the accuracy is degraded to 83.2%. This shows contextual information gathered by BiLSTM fusion is important for interaction between two sentences.

5.8 What is Learned by Attention ?

We further investigate the results of the multi-layered cross sentence attention and the self-attention and then visualize the results in Figure 2. This is an instance from the test set of the SciTail dataset: {P: *all living cells have a plasma membrane that encloses their contents. Q: all types of cells are enclosed by a membrane.* The label y: *entailment.*}. The results are produced by OSOA-DFN with 3 original semantics-oriented cross sentence attentions $P(t) \rightarrow Q$ and 1 self-attention. We visualize the attention matrices for each layer to show the dynamic attention changes.

From the results, we observe that the first attention layer may have erroneous alignments. We can find that the premise word "*encloses*" is incorrectly aligned with hypothesis word "*all*". In the second attention layer, the alignment quality is improved dramatically, where the "*encloses*" is correctly aligned to "*enclosed*". It shows that the second attention layer effectively revises the errors from the first attention layer. In the second and third attention layers, the attention gradually tends to capture phrase-level alignments, such as "*that encloses their contents*" and "*enclosed*", and "*cells have a plasma membrane*" and "*membrane*". Meanwhile, with the increment of interaction, the high attention layers also tend to capture new alignment information from another sentence that is not captured in low attention layers.

In the self-attention layer, we observe that the phrase "*plasma membrane that encloses their contents*" is strongly aligned to the phrase "*living cells*". This layer captures long-distance semantic dependency within the sentence. The visualization of attention further shows that our proposed model is capable of capturing alignment information from two sentences for better semantic matching.

6 Related Works

Recently, deep neural network models have achieved promising results in modeling sentence matching. A standard practice is to encode each sentence as a vector with a neural network (Bowman et al., 2015; Mou et al., 2015; Tan et al., 2015), and then the relation is decided based on the two sentence vectors. This kind of framework ignores the interaction between two sentences.

Most recent works (Wang and Jiang, 2016a; Chen et al., 2017; Duan et al., 2018) employ attention mechanism to model interaction between two sentences. The attention-based framework matches two sentences at the word level. (Wang and Jiang, 2016b) design a specific LSTM called matching-LSTM that performs word-by-word matching of the hypothesis with the premise. Furthermore, (Wang et al., 2017) and (Chen et al., 2017) propose a new framework to model the relationship between two sentences, which performs the matching from two directions. To improve the attention-based framework, (Duan et al., 2018) propose an attention-fused deep matching

network (AD-DMN), which is based on multi-layered attention mechanism and shows that multiple stacked attention layers can improve matching performance. Besides cross sentence attention, the self-attention mechanism is proposed to solve the limitations of RNN model on the long-term dependency problem, which aims to align the sentence with itself and has been used in a variety of tasks (Lin et al., 2017; Duan et al., 2018).

Our proposed OSOA-DFN conducts original semantics-oriented cross sentence attention to model the matching. We design deep fusion to augment the propagation of attention information. At last, we introduce a self-attention mechanism to capture global context to enhance semantic representation. Compared to AF-DMN (Duan et al., 2018), we just use one self-attention layer instead of multiple layers, which reduces model complexity but achieves outperformed accuracy.

7 Conclusions and Future Work

In this paper, we propose an original semantics-oriented attention and deep fusion network (OSOA-DFN) for sentence matching. It leverages original semantics-oriented cross sentence attention, deep fusion and self-attention mechanism jointly. We compare our model with the previous models on two sentence matching tasks: natural language inference and paraphrase identification. Experiment results show that OSOA-DFN has the ability to model sentence matching more precisely and significantly improves the performance.

In the future, we will further investigate the effect of the network depth on sentence matching and explore introducing external knowledge, such as pre-trained language model BERT (Devlin et al., 2018) and paraphrase database (Ganitkevitch et al., 2013), to help learning more accurate and robust sentence representation.

Acknowledgement

The research work descried in this paper has been supported by the National Nature Science Foundation of China (Nos. 61876198, 61976015, 61370130 and 61473294), and also supported by the Fundamental Research Funds for the Central Universities (No. 2018YJS025), the Beijing Municipal Natural Science Foundation (No. 4172047), and the International Science and Technology Cooperation Program of China under Grant No. K11F100010. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proc. ACL*.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chaoqun Duan, Lei Cui, Xinchi Chen, Furu Wei, Conghui Zhu, and Tiejun Zhao. 2018. Attention-fused deep matching network for natural language inference. In *IJCAI*, pages 4033–4040.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323.

Yichen Gong, Heng Luo, and Jian Zhang. 2017. Natural language inference over interaction space. *arXiv preprint arXiv:1709.04348*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jinbae Im and Sungzoon Cho. 2017. Distance-based self-attention network for natural language inference. *arXiv preprint arXiv:1712.02047*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of AAAI*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2015. Natural language inference by tree-based convolution and heuristic matching. *arXiv preprint arXiv:1512.08422*.

Biswajit Paria, KM Annervaz, Ambedkar Dukkipati, Ankush Chatterjee, and Sanjay Podder. 2016. A neural architecture mimicking humans end-to-end for natural language inference. *arXiv preprint arXiv:1611.04741*.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

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*.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Lei Sha, Baobao Chang, Zhifang Sui, and Sujian Li. 2016. Reading and thinking: Re-read lstm unit for textual entailment recognition. In *COLING*, pages 2870–2879.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2017. A compare-propagate architecture with alignment factorization for natural language inference. *arXiv preprint arXiv:1801.00102*.

Shuohang Wang and Jing Jiang. 2016a. A compare-aggregate model for matching text sequences. *arXiv preprint arXiv:1611.01747*.

Shuohang Wang and Jing Jiang. 2016b. Learning natural language inference with lstm. In *Proceedings of NAACL-HLT*, pages 1442–1451.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. corr abs/1702.03814.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence similarity learning by lexical decomposition and composition. *arXiv preprint arXiv:1602.07019*.

Wenpeng Yin, Hinrich Schütze, and Dan Roth. 2018. End-task oriented textual entailment via deep explorations of inter-sentence interactions. *arXiv preprint arXiv:1804.08813*.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv:1512.05193*.