

LadRa-Net: Locally Aware Dynamic Reread Attention Net for Sentence Semantic Matching

Kun Zhang¹, Member, IEEE, Guangyi Lv², Le Wu³, Member, IEEE, Enhong Chen⁴, Senior Member, IEEE, Qi Liu⁵, Member, IEEE, and Meng Wang⁶, Fellow, IEEE

Abstract—Sentence semantic matching requires an agent to determine the semantic relation between two sentences, which is widely used in various natural language tasks, such as natural language inference (NLI) and paraphrase identification (PI). Much recent progress has been made in this area, especially attention-based methods and pretrained language model-based methods. However, most of these methods focus on all the important parts in sentences in a static way and only emphasize how important the words are to the query, inhibiting the ability of the attention mechanism. In order to overcome this problem and boost the performance of the attention mechanism, we propose a novel dynamic reread (DRr) attention, which can pay close attention to one small region of sentences at each step and reread the important parts for better sentence representations. Based on this attention variation, we develop a novel DRr network (DRr-Net) for sentence semantic matching. Moreover, selecting one small region in DRr attention seems insufficient for sentence semantics, and employing pretrained language models as input encoders will introduce incomplete and fragile representation problems. To this end, we extend DRr-Net to locally aware dynamic reread attention net (LadRa-Net), in which local structure of sentences is employed to alleviate the shortcoming of byte-pair encoding (BPE) in pretrained language models and boost the performance of DRr attention. Extensive experiments on two popular sentence semantic matching tasks demonstrate that *DRr-Net* can significantly improve the performance of sentence semantic matching. Meanwhile, LadRa-Net is able to achieve better performance by considering the local structures of sentences. In addition, it is exceedingly interesting that some discoveries in our experiments are consistent with some findings of psychological research.

Index Terms—Dynamic reread (DRr) attention, local structure, representation learning, sentence semantic matching.

I. INTRODUCTION

SENTENCE semantic matching is a fundamental technology in natural language processing (NLP), which requires an agent to predict the semantic relation between two sentences. For example, in natural language inference (NLI), sentence semantic matching is leveraged to determine whether a hypothesis sentence can be inferred reasonably from a given premise sentence [1]. In paraphrase identification (PI), it is utilized to identify whether two sentences have identical meaning or not [2]. Fig. 1 illustrates two representative examples about NLI and PI, respectively.

As a fundamental yet challenging task, sentence semantic matching has been applied in many fields, e.g., information retrieval [3], question answering [4], [5], and dialogue system [6]. With the large annotated datasets, such as Stanford NLI (SNLI) [7], SciTail [8], Quora [9], and Microsoft Research Paraphrase Corpus (MSRP) [2], and advancement of representation learning techniques, such as convolutional neural network (CNN) [10], long short-term memory (LSTM) [11], gated recurrent unit (GRU) [12], and attention mechanism [13], [14], rapid development on sentence semantic matching has been enabled. Early work often focuses on designing different structures for matching modeling [1], [14], [15]. Recently, pretrained language models (bidirectional encoder representations from transformer (BERT) [16] and generative pre-trained transformer (GPT) [17]) have become the new scheme for language understanding. Researchers employ these pretrained methods to process the input sentences and design task-aware network structures for the final tasks, which have achieved much progress in many NLP tasks.

However, there are still some limitations in most existing methods. First, most of them adopt an attention mechanism to select all the important parts of sentences in a static way, restricting the ability of the attention mechanism. Inspired by the idea that pretrained methods leverage dynamic embedding methods to replace the static embedding methods (e.g., Word2Vec [18] and Glove [19]), we intend to make full use of the attention mechanism dynamically, which is one of our main contributions. In this direction, human reading behaviors can bring us plenty of inspiration. “*When reading a piece of text, people will constantly change their focal point for in-depth understanding based on what they already know.*” Taking case one in Fig. 1 as an example, when classifying the inference relation between *a* and *b1*, we first try to understand the global meaning of the sentence in a sequential manner (i.e., reading it word by word) and grasp information as much

Manuscript received 21 June 2020; revised 4 December 2020 and 20 April 2021; accepted 31 July 2021. Date of publication 18 August 2021; date of current version 6 February 2023. This work was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62006066; in part by the National Natural Science Foundation of China under Grant 61727809, Grant 61922073, and Grant 61972125; and in part by the Fundamental Research Funds for the Central Universities, Hefei University of Technology (HFUT), under Grant JZ2021HGTB0075 and Grant JZ2020HGPA0114; and in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 202100029. (Corresponding author: Kun Zhang.)

Kun Zhang, Le Wu, and Meng Wang are with the Key Laboratory of Knowledge Engineering With Big Data, Hefei University of Technology, Hefei, Anhui 230029, China, and also with the School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230029, China (e-mail: zhang1028kun@gmail.com; lewu.ustc@gmail.com; eric.mengwang@gmail.com).

Guangyi Lv is with AI Lab at Lenovo Research, Beijing 100094, China, and also with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: gylv@mail.ustc.edu.cn).

Enhong Chen and Qi Liu are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: chenh@ustc.edu.cn; qiliuq@ustc.edu.cn).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TNNLS.2021.3103185>.

Digital Object Identifier 10.1109/TNNLS.2021.3103185

2162-237X © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See <https://www.ieee.org/publications/rights/index.html> for more information.

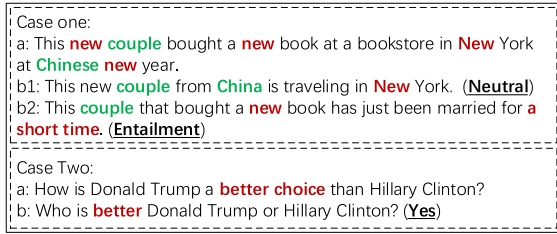


Fig. 1. Examples for SNLI and Quora datasets.

as possible, such as the object: *new couple* and *new book*, the place: *bookstore* and *New York*, and the time: *Chinese new year*. After reading *b1*, we will pay more attention to the object and the place. When confirming that the place is the same, we will turn to identify whether the couple comes from China. Finally, we can conclude that the relation is *Neutral*. Meanwhile, psychological research has also shown that there are plenty of reading behaviors of humans for efficient language understanding and communication. For example, humans only pay attention to a small region of information at one time, i.e., people only focus on 1.5 words each time when intensively reading a piece of text [20]. Koch and Tsuchiya [21] have demonstrated that people focused on less than seven different objects at the same time. Therefore, how to capture these important parts in a dynamic way and make full use of them for in-depth semantic understanding is one of the challenges that we should focus on.

Another easily overlooked limitation is the usage of byte-pair encoding (BPE) or its variations in BERT-based methods. Though this encoding method is capable of avoiding out-of-vocab (OOV) problems, it splits a word into subword units, which suffers from two problems: *incomplete modeling* and *fragile representation* [22]. Therefore, we try to leverage local structures of input sentences to alleviate these problems and further improve the ability of our proposed dynamic attention mechanism. For this direction, we can also learn from human reading behaviors. As shown in Fig. 1, when predicting the inference relation between *a* and *b2*, we will focus on the *couple* and try to figure out how long they have been married. Traditional methods may be confused by the four “*new*”s in *a*. When taking the local structure into consideration, we obtain that the first “*new*” is the attribute of the *couple*, indicating that they are newly married. Thus, the inference relation between *a* and *b2* is “*Entailment*.” Biological researchers have also observed that people will focus on what they care about most and leverage peripheral vision [23], [24] to get the local structure for understanding enhancement. Similar observations have been found by Zheng *et al.* [25] through a lab study. To this end, how to select and use the local structure effectively while paying close attention to important words is another problem that we need to tackle.

In order to solve the challenge of how to employ attention mechanism in a dynamic way for enhancing its ability of selecting important parts, in our preliminary work [26], we proposed the dynamic reread (DRr) attention, a novel architecture that selects one important word at each reading step and reads the important words repeatedly for precise sentence semantic representation. Based on this novel attention, we developed a dynamic reread network (DRr-Net) for sentence semantic matching, in which global sentence encoding

is used to model sentences comprehensively and the DRr unit is used to capture the important parts precisely for better sentence semantic representation and matching. In this way, DRr-Net was able to select the most important word to process based on the learned information, which was in favor of tackling the sentence semantic matching task.

However, there still exists some space in *DRr-Net* for further improvement. First, the encoding ability of global sentence encoding is insufficient for input embedding. Replacing with pretrained methods (e.g., BERT) will cause the problem that BPE suffers from on the other hand. Second, selecting only one important word at each reading step in *DRr-Net* may be insufficient since words can express tremendously different meanings with different local structures. Therefore, in this article, we focus on using local structures to alleviate the above problems. We propose the locally aware dynamic reread attention net (LadRa-Net), an advanced architecture that leverages pretrained language models to encode input sentences and uses the local structure to alleviate the weaknesses of BPE in pretrained methods and further enhance the ability of our proposed dynamic reread (DRr) attention. In concrete details, we first take advantage of pretrained BERT [16] to encode input sentences from a global perspective. Meanwhile, we develop a *Phrase-CNN* unit to capture local structures of the input sentence with different composite kernels (e.g., bigram and trigram) over the output of BERT so that the input sentences can be encoded sufficiently. After getting the words and local phrase representations, a newly designed dynamic sequential attention (DSA) unit is employed to perform dynamic attention in a sequential manner. Along this line, not only the important parts can be measured dynamically but also the corresponding local structure can be fully explored. Meanwhile, the ability of BPE in pretrained models can also be strengthened. Therefore, LadRa-Net is capable of making better modeling of sentence semantics and doing better prediction of sentence semantic relations.

The remainder of this article is organized as follows. We will first introduce the related work in Section II. Then, in Section III, we give a formal definition of a sentence semantic matching task and propose two important issues that will be tackled in this article. Next, the structure and technical details of our proposed models are given in Sections IV and V. The experiments and detailed analysis are given in Section VI. Finally, we discuss and conclude our work in Section VII.

II. RELATED WORK

Our work is related to two lines of literature: 1) *Sentence Semantic Matching*: focusing on the semantic representations and relation verification of sentence pairs and 2) *Human Attention Behavior*: focusing on the study of attention mechanism in human reading behavior.

A. Sentence Semantic Matching

Based on the learning schema, this part can be grouped into two categories: *traditional neural network (NN)-based methods* and *pretrained language model-based methods*. We will introduce each of them in the following parts.

1) *Traditional Neural Network-Based Methods*: With the developments of various NN technologies, such as CNN [10], recurrent NN (RNN) [11], attention mechanism [13], [27],

and graph NN (GNN) [28], [29], plenty of methods have been exploited to model semantic matching on large datasets, such as SNLI [7], SciTail [8], and Quora [9]. Traditionally, researchers try to make full use of NN technologies to model sentence semantic meanings in an end-to-end fashion. Among them, RNNs mainly focus on the sequential information and the semantic dependence of sequences [30]–[32]. CNNs tend to capture the local context with different convolutional filters [10], [33]. The attention mechanism is usually utilized to extract the most important parts in sentences, capture the semantic relations, and align the elements of two sentences properly [34], [35]. For example, Tay *et al.* [36] employed a stacked multilayer Bi-LSTM with alignment factorization to measure all feature hierarchies among two sentences. Yang *et al.* [37] utilized multilayer encoding and fusion block based on CNN structure to build a fast and well-performed sentence matching model. Kim *et al.* [1] proposed a coattention network to model the interaction between two sentences and developed a densely connected structure to retain as much information as possible. Dong *et al.* [38] adopted GNN to access the structure information of input sentences for comprehensive sentence relation modeling. Kun *et al.* [14] combined these three architectures into a hybrid architecture, in which CNN is utilized to generate phrase-level semantic meanings, GRU is used for the word sequence and dependence among sentences, and attention is adopted to integrate different features for the final classification.

2) *Pretrained Language Model-Based Methods*: In order to make full use of existing large language corpora, various pretrained language models have been proposed. For example, Vaswani *et al.* [13] proposed the transformer architecture, in which multihead self-attention and residual connection are used to analyze the inputs. By stacking multiple layers of transformers, input sentences could be fully explored. Based on the transformer and large language corpora, GPT [17] and BERT [16] were proposed to learn language representations in a pretrained manner. Among them, BERT achieved better performance. It made full use of self-supervised mask language model (MLM) and next sentence prediction (NSP) tasks to get a comprehensive understanding of input sentences. Moreover, by modifying the downstream tasks to the required form, BERT was capable of providing promising representations for inputs and improving the model performance effectively [16]. To further improve the performance of pretrained models, one direction is modifying the input encoding and self-supervised pretrained task usage. For example, XLNet [39] leveraged a newly designed permutation language model (PLM) task to narrow down the gap between pretrained tasks and downstream tasks. In addition, there still exists much promising work, such as RoBERTa [40], CharBERT [22], and bidirectional and auto-regressive transformer (BART) [41]. Moreover, pretrained methods still have some weaknesses in accessing external knowledge. To this end, many external knowledge enriched methods have been proposed, such as incorporating knowledge base [42], integrating entity knowledge into pretrained stage [43], and considering syntax and semantics knowledge [44].

The above work has made great progress in sentence semantic matching and inspired us to fully utilize advanced

NNs and pretrained methods for sentence semantic modeling. On the other hand, current methods still have some weaknesses in dealing with sentence semantics and need to be further improved. First, Ma *et al.* [22] has proved that BPE structure in BERT is insufficient for word representations and easy to be attacked. Meanwhile, attention operation in BERT is calculated among the entire input sequence, in which the irrelevant parts would interfere with the weights of important parts. Thus, better modeling for input embedding should be considered. Second, most existing methods only treat the structured information (knowledge base) as prior knowledge, ignoring the potential of reading behaviors of humans. Humans invent many reading behaviors (e.g., skipping and repeating) for language understanding and communication efficiently [25], which can also be treated as prior knowledge. Therefore, in this article, we take human reading behaviors into consideration and propose the novel DRr and DSA units to boost the ability of attention mechanisms for better sentence semantic modeling. We design two-sentence semantic matching models (i.e., *DRr-Net* and *LadRa-Net*) based on DRr and DSA units separately to tackle the above problems.

B. Human Attention Behavior

Human learning has inspired various algorithm designs throughout the development of machine learning [45]. For example, curriculum learning [46] tries to train a model from easier data to harder data, which imitates the meaningful learning order in human curricula. Attention mechanism helps an agent to focus on the most relevant parts of input for output to achieve better performance, which imitates the human ability of quickly perceiving necessary information by selectively attending to parts of what they saw [47]. Despite the success of imitating human actions, we have to admit that these algorithm designs still have large performance gaps with human beings in more practical settings [25]. Learning better from human behaviors may help models to achieve better performances.

Taking attention mechanism as an example, Zheng *et al.* [25] observed some user behavior patterns for better attention usage with a lab study. For example, in a specific scenario (e.g., answer selection), users tended to pay more attention to possible segments that are relevant to what they want. They would reread more snippets of candidate answers with more *skip* and *up* behaviors while ignoring the irrelevant parts [25], [48]. Moreover, Sen *et al.* [49] obtained significant similarities between human attention and machine attention on a large text classification dataset by experiments. Peng *et al.* [50] demonstrated that imitating reverse thinking and inertial thinking of humans can improve the model performance on reading comprehension tasks. These observations indicated that human attention was helpful for guiding NN design, and it is a very promising direction.

Besides, psychologists also have similar observations. For example, Yarbus [51] described that human attention “*is dependent on not only what is shown in the picture but also the problem facing the observer and the information that he hopes to gain.*” By building an eye tracker, Wang *et al.* [20] found that people may center on 1.5 words each time when

reading a piece of text intensively. In addition, researchers have already demonstrated that human eyes have a central vision and peripheral vision. Central vision concentrated on what a person needs at the current time. Peripheral vision used the coarse-grained observation of the surroundings to support the central vision [23], [24], [52].

These studies encourage us to learn from human learning behaviors and improve model performance by incorporating this behavior knowledge. Therefore, in this article, we focus on utilizing behavior knowledge to further improve the model performance on sentence semantic modeling.

III. PROBLEM STATEMENT

In this section, we formulate the sentence semantic matching task as a supervised classification problem. Given the triple (s^a, s^b, y) , where $s^a = \{w_1^a, w_2^a, \dots, w_{l_a}^a\}$ and $s^b = \{w_1^b, w_2^b, \dots, w_{l_b}^b\}$ are the given sentences, w_i^a and w_j^b are one-hot vectors that represent the i th and j th words in the sentences, and l_a and l_b indicate the total number of words in s^a and s^b . The true label $y \in \mathcal{Y}$ indicates the semantic relationship between the given sentence pair, where $\mathcal{Y} = \{\text{entailment, contradiction, neutral}\}$ for the NLI task and $\mathcal{Y} = \{\text{Yes, No}\}$ for the PI task. Our goal is to learn the classifier ζ , which is able to compute the conditional probability $P(y|s^a, s^b)$ and predict the label for the given sentence pair in the test set by $y^* = \operatorname{argmax}_{y \in \mathcal{Y}} P(y|s^a, s^b)$.

In order to do better sentence semantic matching, the following important challenges should be considered.

- 1) Static usage of attention may inhibit the ability of attention mechanism. How to adopt an attention mechanism dynamically for important parts selection and sentence semantic understanding?
- 2) The local structures are helpful for overcoming the weaknesses of BPE in BERT and enhancing the ability of DRr attention. How to select and use this information efficiently while selecting the important words?

To this end, we propose *DRr-Net* and *LadRa-Net* to tackle the above issues for sentence semantic matching.

IV. DRr NETWORK

The overall architecture is shown in Fig. 2(a), which contains four main components: 1) *input embedding*; 2) *global sentence encoding*; 3) *DRr*; and 4) *label prediction*. Next, we will introduce each of them in detail.

A. Input Embedding

The input embedding converts each word of sentences into a vector representation and constructs the representation matrix for the sentences. We combine multiple features as the semantics representations of words. To be specific, the inputs of *DRr-Net* are one-hot representations $s^a = \{w_1^a, w_2^a, \dots, w_{l_a}^a\}$ and $s^b = \{w_1^b, w_2^b, \dots, w_{l_b}^b\}$. For more comprehensive access to the semantics of each word in sentences, we adopt the concatenation of pretrained word embedding [19], character features [53], and syntactical features [54], [55] to represent each word in sentences. The word embedding is obtained

by mapping each token to a high-dimensional vector space by pretrained word vector (840B Glove [19]). The character features are obtained by applying a CNN with a max-pooling layer to the learned character embeddings, which can represent words in a finer granularity and help to avoid the out-of-vocabulary problem that pretrained word vectors often suffer from. The syntactical features consist of a one-hot part-of-speech (POS) tagging feature, binary exact match feature, and binary antonym feature, which have been proved useful for sentence semantic understanding [54]. Next, we pass these representations through a two-layer highway network [56] and get the extravagant representations $\{a_i | i = 1, 2, \dots, l_a\}$ and $\{b_j | j = 1, 2, \dots, l_b\}$ for the words in sentences s^a and s^b .

B. Global Sentence Encoding

As is known to us all, humans can leverage enormous prior knowledge to extract important parts for sentence semantics directly. However, it is quite formidable for models. They have to learn as much information as possible from the input data. Since RNN is powerful to process the sequence and helps models to capture the intradependence and interaction of the input sequence, we select Stack-RNN [57] to process sentences. It is composed of multiple RNN layers on top of each other. Note that we utilize GRU as the base unit. Specifically, let H_l be the l th GRU layer. At the time step t , Stack-RNN can be expressed as follows:

$$h_t^l = H_l(x_t^l, h_{t-1}^l), \quad x_t^l = h_{t-1}^{l-1} \quad (1)$$

where x_t^l is the input of the t th step in the l th GRU layer. While this architecture enables us to build up deeper representations, it cannot model sentence semantics comprehensively and preserve all the learned information, and even worse, this architecture might trigger explosion or make gradient problems vanish. Thus, in order to optimize the capability of Stack-RNN and motivated by Kim *et al.* [1], we concatenate the inputs $x^{(l-1)}$ and the states $h^{(l-1)}$ of the $(l-1)$ th GRU layer as the inputs of the l th GRU layer and modified (1) as follows:

$$h_t^l = H_l(x_t^l, h_{t-1}^l), \quad x_t^l = [h_{t-1}^{l-1}; x_{t-1}^{l-1}] \quad (2)$$

where $[\cdot; \cdot]$ denotes the concatenation operation. The final outputs are denoted as $\{h_i^a | i = 1, 2, \dots, l_a\}$ and $\{h_j^b | j = 1, 2, \dots, l_b\}$, which can preserve all the information, as well as the previous feature work in word embedding part.

However, this architecture only models the sentence and stores all the information into vectors in a comprehensive way. How to compress these vectors into one sentence representation is still unclear. Since natural language has the redundancy mechanism [58], different words have different contributions to sentence semantics. Moreover, self-attention can select the important parts at different positions from a single sequence [13]. Therefore, it is natural to leverage self-attention to generate sentence representations

$$\begin{aligned} A^a &= [h_1^a, h_2^a, \dots, h_{l_a}^a] \\ \alpha^a &= \omega^T \tanh(WA^a + b) \\ h^a &= \sum_{i=1}^{l_a} \frac{\exp(\alpha_i^a)}{\sum_{k=1}^{l_a} \exp(\alpha_k^a)} h_i^a, \quad i = 1, 2, \dots, l_a \end{aligned} \quad (3)$$

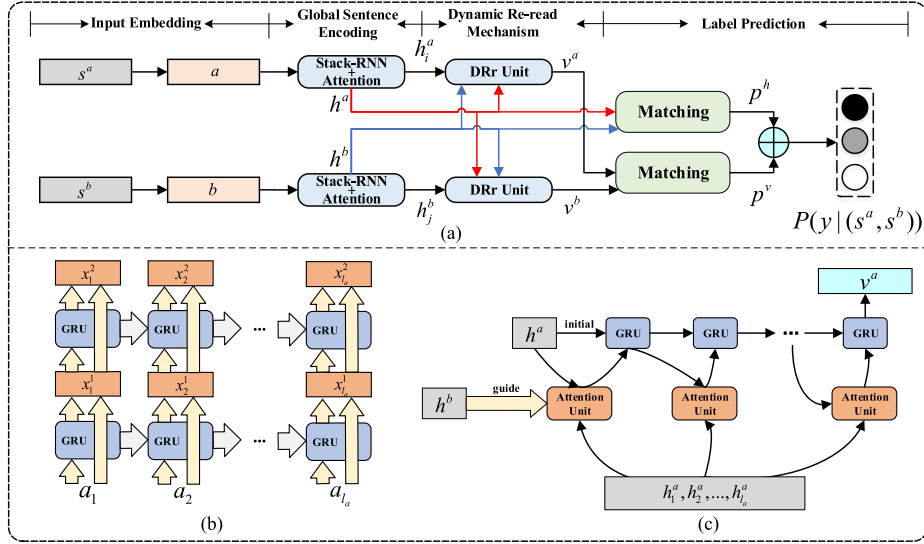


Fig. 2. (a) Architecture of *Dynamic Re-read Network (DRr-Net)*. (b) Stack-RNN processes the original sentence and preserves all the information from bottom-most word embedding input to up-most recurrent output. (c) DRr Unit pays close attention to only one important word at each step with the learned information and reread these important words for sentence semantic matching more precisely. (a) Architecture of DRr-Net. (b) Two-layer Stack-RNN unit without attention. (c) DRr Unit.

where $\{W \in \mathbb{R}^{d_g * (d_w + l_s * d_g)}, \omega \in \mathbb{R}^{d_g * 1}\}$ are the trainable parameters. d_g, d_w and l_s denotes the hidden size of GRU, the embedding size of input words, and the stack layer number, respectively. h_i^a is the i th output of Stack-RNN for sentence s^a . h^a denotes the global contextual representation of sentence s^a , which is actually a weighted summation of the outputs of Stack-GRU. The same operation will be done on sentence s^b to get the global contextual representation h^b .

C. DRr Mechanism

As mentioned in Section I, static usage of the attention mechanism will inhibit its ability. Meanwhile, psychological research has shown that people usually pay close attention to the small part that they care about most and dynamically change their focus point for an in-depth understanding of the sentence [20]. Therefore, we develop the DRr mechanism to solve the first challenge mentioned in Section III. As shown in Fig. 2(c), the DRr unit selects the most important word at each step with the consideration of global contextual representations and the selections in previous steps.

In detail, the inputs of DRr unit are the final outputs $\{h_i^a | i = 1, 2, \dots, l_a\}$ and $\{h_i^b | i = 1, 2, \dots, l_b\}$ of previous component. In each reading step, we adopt attention mechanism to choose one important word \bar{a}_t from the whole input sequence based on the learned information (i.e., \bar{h}_{t-1}^a, h^b). Then, we utilize GRU to encode the chosen word and generate the DRr representation v^a for sentence s^a as follows:

$$\begin{aligned} \bar{a}_t &= F([\mathbf{h}_1^a, \mathbf{h}_2^a, \dots, \mathbf{h}_{l_a}^a], \bar{h}_{t-1}^a, \mathbf{h}^b) \\ \bar{h}_t^a &= \text{GRU}(\bar{a}_t, \bar{h}_{t-1}^a), \quad t = 1, 2, \dots, T \\ v^a &= \bar{h}_T^a \end{aligned} \quad (4)$$

where h^b is the global contextual representation for sentence s^b . T is the DRr length. We also employ the global contextual representation h^a as the initial state of GRU for sentence a . $F(\cdot)$ is the choosing function, and we utilize attention

mechanism to achieve this function as follows:

$$\begin{aligned} \bar{A}^a &= [\mathbf{h}_1^a, \mathbf{h}_2^a, \dots, \mathbf{h}_{l_a}^a] \\ \bar{m}^a &= \omega_d^T \tanh(W_d \bar{A}^a + (U_d \bar{h}_{t-1}^a + M_d \mathbf{h}^b) \otimes \mathbf{e}_{l_a}) \\ \bar{\alpha}^a &= \sum_{i=1}^{l_a} \frac{\exp(\bar{m}_i^a)}{\sum_{k=1}^{l_a} \exp(\bar{m}_k^a)} \\ \bar{a}_t &= h_{j_t}^a, \quad (j_t = \text{Index}(\max(\bar{\alpha}^a))) \end{aligned} \quad (5)$$

where $\{W_d \in \mathbb{R}^{d_a * (d_w + l_s * d_g)}, U_d \in \mathbb{R}^{d_a * d_g}, M_d \in \mathbb{R}^{d_a * (d_w + l_s * d_g)}, \omega_d \in \mathbb{R}^{d_a * 1}\}$ are trainable parameters. $\mathbf{e}_{l_a} \in \mathbb{R}^{l_a}$ is a row vector of 1. $\text{Index}(\max(\bar{\alpha}^a))$ denotes getting the corresponding index of the maximum value in the attention vector $\bar{\alpha}^a$. The outer product $(U_d \bar{h}_{t-1}^a + M_d \mathbf{h}^b) \otimes \mathbf{e}_{l_a}$ means repeating l_a times of the results $(U_d \bar{h}_{t-1}^a + M_d \mathbf{h}^b)$. Specifically, the global contextual representation h^b of sentence s^b is beneficial for choosing the most relevant information in sentence s^a . The previous hidden state \bar{h}_{t-1}^a is capable of memorizing what has been selected in the previous steps. Treating them as the input of the attention unit, *DRr-Net* can select the most important part from sentence s^a at the t th time step with the consideration of the information from sentence s^b and the previous steps.

However, $\text{Index}(\max(\cdot))$ operation has no derivative, which means its gradient could not be calculated. Fortunately, our goal is to select the most important part, which requires one word at one step. Inspired by softmax function, we modify (5) as follows to solve the nonderivative problem:

$$\begin{aligned} \bar{A}^a &= [\mathbf{h}_1^a, \mathbf{h}_2^a, \dots, \mathbf{h}_{l_a}^a] \\ \bar{m}^a &= \omega_d^T \tanh(W_d \bar{A}^a + (U_d \bar{h}_{t-1}^a + M_d \mathbf{h}^b) \otimes \mathbf{e}_{l_a}) \\ \bar{a}_t &= \sum_{i=1}^{l_a} \frac{\exp(\beta \bar{m}_i^a)}{\sum_{k=1}^{l_a} \exp(\beta \bar{m}_k^a)} \mathbf{h}_i^a \end{aligned} \quad (6)$$

where β is an arbitrarily big value. With this operation, the weight \bar{a}_t of the most important word will be very close to 1, and other weights will be very close to 0.

D. Label Prediction

This component consists of three operations: matching, fusion, and classification. In order to determine the overall relation between two sentences, we leverage heuristic matching [59] between global contextual representations \mathbf{h}^a , \mathbf{h}^b , and DRr representations \mathbf{v}^a and \mathbf{v}^b . Specifically, we use the elementwise product, their difference, and concatenation. Then, we send them to multilayer perceptron (MLP) to calculate the relation probability between two sentences. The MLP has two hidden layers with rectified linear unit (ReLU) activation and a softmax output layer

$$\begin{aligned} \mathbf{h} &= [(\mathbf{h}^a; \mathbf{h}^b); (\mathbf{h}^b \odot \mathbf{h}^a); (\mathbf{h}^b - \mathbf{h}^a)] \\ \mathbf{v} &= [(\mathbf{v}^a; \mathbf{v}^b); (\mathbf{v}^b \odot \mathbf{v}^a); (\mathbf{v}^b - \mathbf{v}^a)] \\ \mathbf{p}^h &= \text{MLP}_1(\mathbf{h}) \\ \mathbf{p}^v &= \text{MLP}_2(\mathbf{v}) \end{aligned} \quad (7)$$

where \mathbf{p}^h and \mathbf{p}^v denote the probability distributions of different classes with global contextual sentence representations and dynamic important part representations, respectively.

After getting different probability distributions of the semantic relations using different sentence semantic representations, we intend to integrate this information to achieve a more robust performance. Thus, we utilize a fusion gate and an MLP to integrate \mathbf{p}^h , \mathbf{p}^v and make the final classification, which can be formulated as follows:

$$\begin{aligned} \alpha_h &= \sigma(\mathbf{w}_h^T \mathbf{p}^h + b_h) \\ \alpha_v &= \sigma(\mathbf{w}_v^T \mathbf{p}^v + b_v) \\ P(y|(\mathbf{s}^a, \mathbf{s}^b)) &= \text{MLP}_3(\alpha_h \mathbf{p}^h + \alpha_v \mathbf{p}^v). \end{aligned} \quad (8)$$

V. LOCALLY AWARE DRr ATTENTION NET

As mentioned in Section I, our proposed *DRr-Net* still has some space for further improvement. First, the encoding capability of GRU has been proven weaker than transformer [16]. Replacing it with pretrained models (e.g., BERT) will import the incomplete modeling and fragile representation that BPE suffers from. Second, selecting only one important word at each reading time is insufficient since words with different contexts can express tremendously different meanings.

To solve the previous shortcomings, in this section, we focus on the local information utilization and extend *DRr-Net* to a novel *LadRa-Net*, which employs pretrained BERT to encode sentences and a newly designed DSA for the usage of local structures and the selection of important parts. Our key contributions lie in two parts. The first is using a CNN-based structure to alleviate the problem that BPE suffers from and enhance the encoding ability of BERT. The second is designing a novel DSA unit to integrate the chosen important part and corresponding local structure in a sequential way at each reading step for performance improvement of our proposed dynamic attention mechanism. Moreover, the DSA unit utilizes a coverage mechanism to ensure the diversity of dynamic selection, which can help to evaluate the semantic relation between the sentence pair more comprehensively. Next, we will introduce the technical details of *LadRa-Net*.

A. Global Encoding

In order to encode sentences more comprehensively and make full use of the pretrained language model, *LadRa-Net* chooses BERT-base [16] to encode each word in the sentences from a global perspective. Specifically, the input sentence pair is first concatenated with “[SEP]” token and added “[CLS]” at the beginning and the end as BERT requires. Then, BPE tokenizer is employed to tokenize each word into BPE tokens. Suppose that the final number of tokens in the sentence pair is l_{ab} , and BERT generates L hidden states for each BPE token \mathbf{BERT}_t^l , $1 \leq l \leq L$, $1 \leq t \leq l_{ab}$. As shown in Fig. 3(a), the contextual representation for the t th token in input sentence pair at token level is then a per-layer weighted sum of transformer block output, with weights $\alpha_1, \alpha_2, \dots, \alpha_L$

$$\mathbf{h}_t^0 = \sum_{l=1}^L \alpha_l \mathbf{BERT}_t^l, \quad 1 \leq t \leq l_{ab} \quad (9)$$

where α_l is the weight for the l th layer in BERT and is trained during model training. \mathbf{h}_t^0 is the contextual representation for the t th token. Here, 0 denotes the index of weighted output from BERT and is consistent with the description in the following parts. Moreover, the output \mathbf{h}^0 of the first special token “[CLS]” in the last block is treated as the contextual representation for input sentence pair at sentence level.

B. Local Structure Encoding

The semantic relationships within the sentence pair are not only connected with the important words but also affected by the corresponding local structure. Selecting only one important word at each reading step in *DRr-Net* seems insufficient for semantic relation modeling. Meanwhile, though BERT leverages multilayer operation to perceive coarser granularity and learns more about words or phrases, the used BPE will import incomplete modeling and fragile representation problems for input sentences. To overcome the above shortcomings, we adapt CNN with different composite kernels (e.g., bigram and trigram) and propose the *Phrase-CNN (PCNN)* unit to extract local contexts of input sentence pair. As illustrated in Fig. 3(b), let \mathbf{h}_t^r be the local structure representation for the t th token with the r th kernel. This process can be formulated as follows:

$$\mathbf{h}_t^r = \text{CNN}_r([\mathbf{h}_{t-k}^0, \mathbf{h}_{t-k+1}^0, \dots, \mathbf{h}_{t+k-1}^0, \mathbf{h}_{t+k}^0]), \quad t = 1, 2, \dots, l_{r-1}, l_r \quad (10)$$

where k denotes the kernel size of the r th PCNN unit. l_r denotes the number of PCNN units. In order to maintain consistency of the output of different encoding units in this layer, the output length (i.e., number of output channels) of each PCNN unit is the same as that of the input sequence. Moreover, we intend to generate the contextual representations \mathbf{h}^r based on the local structure information from PCNN more comprehensively. Thus, we apply self-attention to generate

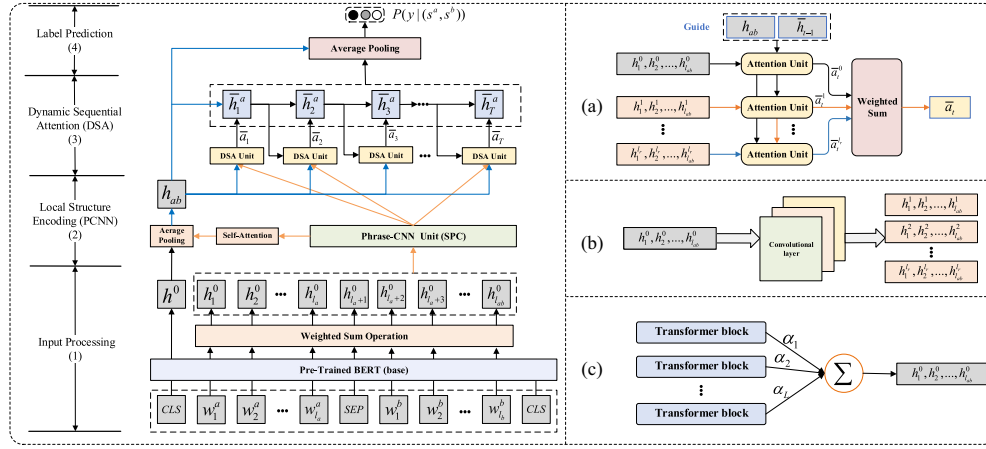


Fig. 3. Overall architecture of *Locally Aware Dynamic Re-read Attention Net (LadRa-Net)* and DSA. (a) Weighted sum operation. (b) PCNN unit. (c) DSA unit.

contextual representations over the output of PCNN

$$\begin{aligned}
 A_{ab}^r &= [h_1^r, h_2^r, \dots, h_{l_{ab}}^r] \\
 \alpha^r &= \omega_r^T \tanh(W_r A_{ab}^r + b) \\
 h^r &= \sum_{i=1}^L \frac{\exp(\alpha_i^r)}{\sum_{j=1}^L \exp(\alpha_j^r)} h_i^r \\
 h_{ab} &= \text{avg}([h^0, h^1, h^2, \dots, h^{l_{ab}}]) \quad (11)
 \end{aligned}$$

where $\{W_r \in \mathbb{R}^{d_a \times 768}, \omega_r \in \mathbb{R}^{d_a \times 1}\}$ are trainable parameters. h_i^r is the i th token representation from the r th PCNN encoding kernel. avg denotes the average-pooling. h_{ab} denotes the sentence-level representation for input sentence pair. We have to note that we adopt BERT-base as the basic encoder. Thus, its output size 768 for each token affects the parameter size of the following network structures of *LadRa-Net*.

C. Dynamic Sequential Attention

In our preliminary work [26], we leverage the DRr unit to select one important word at each reading step. However, one word can express tremendously different semantic meanings due to different local structures. Only selecting one word may be insufficient for precise sentence semantic modeling. In order to leverage the corresponding local structures to enrich the analysis of the selected important parts and represent these parts more precisely, we develop a novel DSA unit, which is shown in Fig. 3(c). DSA employs the learned information to select the most important part in a sentence and uses the local structure to boost its representation in a sequential manner at each reading step. Then, a weighted sum operation is employed to integrate the selected important part and the corresponding local structure. After that, the enhanced result will be treated as the input of GRU.

Concretely, DSA first employs the choosing function $F(\cdot)$ and learned information (i.e., $\bar{h}_{t-1}, h_{ab}, \bar{a}_t^{[0:(k-1)]}$) to select the most important part \bar{a}_t^r from BERT output $[h_1^0, h_2^0, \dots, h_{l_{ab}}^0]$ at the current step. Then, the same operation is applied on PCNN output $([h_1^r, h_2^r, \dots, h_{l_{ab}}^r], r = 1, 2, \dots, l_r)$ to select the proper local structures \bar{a}_t^r in a sequential manner, which can support the understanding to the important part. Next, it adopts a weight sum to fuse all these selected information. Since

humans generally select these important parts in a sequential manner, DSA also selects GRU to process the fusion results. The fusion result \bar{a}_t is sent to a GRU and treated as the t th input of this GRU as follows:

$$\begin{aligned}
 \bar{a}_t^r &= F([h_1^r, h_2^r, \dots, h_{l_{ab}}^r], \bar{h}_{t-1}, h_{ab}, \bar{a}_t^{[0:(r-1)]}) \\
 \alpha_t^r &= \omega_f^T \tanh(W_f \bar{a}_t^r) \\
 \bar{a}_t &= \sum_{r=1}^{l_r} \frac{\exp(\alpha_t^r)}{\sum_{j=1}^{l_r} \exp(\alpha_t^j)} \bar{a}_t^r \\
 \bar{h}_t &= \text{GRU}(\bar{a}_t, \bar{h}_{t-1}), \quad t = 1, 2, \dots, T \quad (12)
 \end{aligned}$$

where $\bar{a}_t^{[0:(r-1)]}$ is previous r selections at time step t . \bar{h}_{t-1} denotes the $(t-1)$ th hidden state of GRU. \bar{h}_0 is initialized with sentence-level representation h_{ab} . T is the length of dynamic selection. $F(\cdot)$ is the choosing function. We select the additive attention [60] to implement $F(\cdot)$ and modify it so that only the most important feature is selected

$$\begin{aligned}
 \bar{A}^{r,a} &= [h_1^r, h_2^r, \dots, h_{l_{ab}}^r] \\
 \bar{m}_t^r &= \omega_{rd}^T \tanh \left(W_{rd} \bar{A}^r + \left(U_{rd} \bar{h}_{t-1} + M_{rd} h_{ab} \right. \right. \\
 &\quad \left. \left. + \sum_{j=0}^{r-1} V_{rd} \bar{a}_t^j \right) \otimes e_{l_{ab}} \right) \\
 \bar{a}_t^r &= \sum_{i=1}^{l_{ab}} \frac{\exp(\beta \bar{m}_{t,i}^r)}{\sum_{j=1}^{l_{ab}} \exp(\beta \bar{m}_{t,j}^r)} h_i^r \quad (13)
 \end{aligned}$$

where $\{W_{rd} \in \mathbb{R}^{d_a \times 768}, U_{rd} \in \mathbb{R}^{d_a \times d_g}, M_{rd} \in \mathbb{R}^{d_a \times 768}, V_{rd} \in \mathbb{R}^{d_a \times 768}, \omega_{rd} \in \mathbb{R}^{d_a \times 1}\}$ are trainable parameters. Similar to *DRr-Net*, we also modify the softmax with this arbitrary big value β to imitate the process of the most important part selection. One step further, DSA unit selects the important parts of input sentences in an unsupervised manner, and this unit is learned along with the end-to-end training of the entire model. Therefore, the DSA unit might result in two problems: 1) *Overselection*: some important parts are selected repetitively too many times and 2) *Underselection*: some important parts are mistakenly ignored. Inspired by the coverage mechanism

used in neural machine translation [61], we also propose to leverage the coverage mechanism to remember the dynamic selection at previous steps, and it will be updated after each selection. Specifically, a coverage vector c_t^r is used to indicate the degree of which word is not concerned at the t th reading step for the r th encoding unit, enabling the diversity of dynamic selections in the model. This vector is initialized as an l_{ab} dimension vector of 1, indicating that all words in sentences have not been selected. After selecting the important part at each reading step, we update this vector based on the calculated attention weights \bar{m}_t^r and the hidden states $\bar{A}^{r,a}$ of input sentences. Thus, 13 will be modified as follows:

$$\begin{aligned}
 c_0^r &= [1, 1, \dots, 1] \in \mathbb{R}^{l_{ab}} \\
 \bar{m}_t^r &= c_{t-1}^r \omega_{rd}^T \tanh \left(\mathbf{W}_{rd} \bar{A}^r + \left(\mathbf{U}_{rd} \bar{h}_{t-1} + \mathbf{M}_{rd} \mathbf{h}_{ab} \right. \right. \\
 &\quad \left. \left. + \sum_{j=0}^{r-1} \mathbf{V}_{rd} \bar{a}_t^j \right) \otimes e_{l_{ab}} \right) \\
 c_t^r &= c_{t-1}^r - \frac{1}{\phi} \bar{m}_t^r, \quad \phi = T \cdot \sigma(\mathbf{W}_\phi \bar{A}^r). \quad (14)
 \end{aligned}$$

D. Label Prediction

After getting the dynamic reading results at each important position, it is natural to integrate the information for the final prediction. Since the sentence-level outputs from BERT, PCNN, and hidden states from GRU indicate the semantic relations from different perspectives, we utilize average pooling to process the results $[\mathbf{h}_{ab}, \bar{h}_1, \bar{h}_2, \dots, \bar{h}_T]$ for decision-making. Then, we send the result \mathbf{v} to a MLP for final classification. The MLP consists of one hidden layer with ReLU an activation and a softmax output layer

$$\begin{aligned}
 \mathbf{v} &= \text{avg}([\mathbf{h}_{ab}, \bar{h}_1, \bar{h}_2, \dots, \bar{h}_T]) \\
 P(y|(s^a, s^b)) &= \text{MLP}(\mathbf{v}). \quad (15)
 \end{aligned}$$

VI. EXPERIMENT

In this section, we first give a brief introduction to the datasets and evaluation methods. Then, we introduce implementation details about the proposed models. Next, we present empirical results on all datasets and give a detailed analysis of the model and experimental results. For all reported results, we employ boldface and underline for the best and the second best results, respectively.

A. Datasets and Evaluation Methods

To evaluate the proposed *DRr-Net* and *LadRa-Net* models, we conduct an empirical evaluation based on two well-known tasks (i.e., SNLI and PI). For each task, we select three benchmark datasets for evaluation. Specifically, for NLI task, we select SNLI [7], the Sentences Involving Compositional Knowledge (SICK) [62], and SciTail [8] datasets to evaluate the model performance. For PI task, we choose Quora [9], MSRP [2], and Twitter-URL [63]. These two tasks cover the asymmetric and symmetric sentence relations separately and can exhibit different characteristics.

TABLE I

HYPERPARAMETERS CONFIGURATION IN *DRr-Net* AND *LadRa-Net*

Model	Hyper-parameters	Value
<i>DRr-Net</i>	Word embedding size	$d = 300$
	Character embedding size	$d_c = 100$
	embedding size of input words	$d_w = 412$
	GRU hidden size	$d_g = 256$
	attention size	$d_a = 200$
	Stack layer number	$l_s = 3$
	Dynamical Reading Length	$T_1 = 6$
	Initial learning rate	$\alpha = 0.001$
<i>LadRa-Net</i>	Adam β_1	$\beta_1 = 0.9$
	Adam β_2	$\beta_2 = 0.999$
	Kernel size of PCNN	$d_k = 2, 3$
	Output size of PCNN	$d_c = 500, 768$
	Dynamical Reading Length	$T_2 = 5$
	Initial learning rate	$\alpha = 0.001$

Since both NLI and PI tasks can be treated as classification problems, and most of the baselines employ *accuracy* as the evaluation metric, we also use *accuracy* to compare the model performance. Moreover, we leverage *F1* to testify the models on the Twitter-URL dataset as other baselines did. We have to note that, for each experiment, we repeat the evaluation process five times with different seeds and report the best results.

B. Implementation Details and Model Training

In order to get the best performance, we leverage the validation set to tune the hyperparameters. If the loss on the validation set does not decrease in 1000 batches, we will stop the training process and accept the best model on validation as the final model. For MSRP [2] and Twitter-URL [63] that do not have the validation set, we randomly split 5% data from the training set as the validation set. Since the hyperparameters will be various for different datasets, we list some common hyperparameters of *DRr-Net* and *LadRa-Net* in Table I. They will be described in the following part.

For *DRr-Net*, the word embedding size is set to 300, and the character embedding size is set to 100. The stack layer number in global sentence encoding is $d_{s1} = 3$. The reread length in the DRr attention mechanism is $T_1 = 6$. For initialization, we obtained the word embedding from pretrained word vectors (840B GloVe) [19], and the out-of-vocabulary (OOV) words are randomly initialized. They are kept to be fixed when training the entire model. All the model parameters are initialized with the uniform distribution in the range between $-(6/(\text{nin} + \text{nout}))^{1/2}$ and $(6/(\text{nin} + \text{nout}))^{1/2}$ as suggested by Orr and Müller [64]. All biases are set as zeros.

For *LadRa-Net*, the kernel sizes of PCNN are set as 2 and 3, and the output dimensions of two-layer CNN are set as 500 and 768, the same as the output size of BERT. The length of dynamic selection in DSA is set as 5. The optimizer that we use is the Adam optimizer, and the initial learning rate is 0.001.

For both methods, we leverage Adam as the optimizer, and the initial learning rate is 0.001. During training, we design a learning rate decay methods called *Cosine Warm Up Decay*, in which cosine operation and training epochs are utilized to update the learning rate in a nonlinear manner. Corresponding details can be found in the released code.¹

¹<https://github.com/little1tow/Cosine-Warm-Up-Decay-method>

TABLE II
EXPERIMENTAL RESULTS (ACCURACY) ON SNLI AND SICK DATASETS

Model	SNLI Full test	SNLI Hard test	SICK test
(1) CAFE [65]	85.9%	66.1%	86.1%
(2) Distance-based SAN [66]	86.3%	67.4%	86.7%
(3) DRCN [1]	86.5%	68.3%	87.4%
(4) Dynamic Self-Attention [67]	87.4%	71.5%	87.8%
(5) CSRAN [36]	88.5%	76.8%	89.7%
(6) RE2 [37]	88.9%	77.3%	89.8%
(7) BERT-(base) [16]	90.3%	80.6%	88.7%
(8) BERT-(large) [16]	90.7%	81.3%	88.3%
(9) RoBERTa-(base) [40]	90.9%	81.5%	90.3%
(10) ALBERT-(base) [68]	86.2%	77.5%	87.3%
(11) <i>DRr-Net</i>	87.7%	71.4%	88.3%
(12) <i>LadRa-Net</i> -BERT-(base)	91.0%	81.1%	89.5%
(13) <i>LadRa-Net</i> -RoBERTa-(base)	91.2%	81.6%	89.8%

C. NLI Results

1) *Performance on SNLI and SICK*: As illustrated in Table II, for BERT-free methods, Co-Stack Residual Affinity Networks (CSRAN) [36] and RE2 [37] achieve the best performance. The following is our proposed *DRr-Net*. Though *DRr-Net* did not perform the best, it leverages a relatively simple structure (DRr unit) to achieve competitive results compared with the complex multiple RNN and attention calculation mixing layers used in CSRAN and RE2. This phenomenon supports the effectiveness of our proposed DRr attention. For powerful pretrained models, we first note that “[CLS]” output from the last layer of pretrained models is leveraged for final classification, and we employ the same training strategy as our proposed methods did for a fair comparison. This is also the reason why the results of pretrained models are a little different from the General Language Understanding Evaluation (GLUE) leaderboard. From the table, we can first observe that BERT-based methods outperform all the BERT-free methods by a large margin, demonstrating the powerful encoding capability of pretrained methods. Moreover, *LadRa-Net* performs better than these pretrained models. Since they all leverage the pretrained methods as the basic encoders, the performance improvement of *LadRa-Net* can demonstrate the usefulness of local structure. By employing the local structure, *LadRa-Net* can access the complete words or phrases in the input sentences, alleviating the incomplete modeling or fragile representation of Binance Chain Tokenization Standard (BEP) [22] and achieving better performance. On the other hand, local structure and DRr attention utilization require more parameters to be trained. Thus, it may be a little underfitting when the dataset is small, which is the possible reason why *LadRa-Net* did not achieve the best performance on the SICK dataset. A similar performance can also be observed on the MSRP dataset from Table IV.

Among all baselines, CSRAN and RE2 are the current best BERT-free methods. RoBERTa-base [40] achieves the best performance among BERT-based methods. After a detailed analysis, we conclude the following observations. First, both CSRAN and RE2 stack multilayer blocks to achieve impressive performance, which is similar to the transformers in BERT-based methods. This observation proves that stacking multilayer blocks is, indeed, helpful for semantic representation. Second, RoBERTa-base achieves better performance than BERT-large, which is counterfactual. We speculate the possible reason is that we did not fine-tune BERT-large well on these

TABLE III

EXPERIMENTAL RESULTS (ACCURACY) ON THE SciTail DATASET

Model	SciTail test
(1) CAFE [65]	83.3%
(2) HBMP [15]	86.0%
(3) Finetuned Transformer LM [17]	88.3%
(4) CSRAN [65]	86.5%
(5) RE2 [37]	86.2%
(6) BERT-(base) [16]	93.1%
(7) BERT-(large) [16]	93.6%
(8) RoBERTa-(base) [40]	93.8%
(9) ALBERT-(base) [68]	91.4%
(10) <i>DRr-Net</i>	87.4%
(11) <i>LadRa-Net</i> -BERT-(base)	93.9%
(12) <i>LadRa-Net</i> -RoBERTa-(base)	93.9%

datasets since there are huge amounts of parameters. Based on this reason, we only select BERT-base and RoBERTa-base as the basic encoders in *LadRa-Net*. Third, the improvement between results (13) and (9) is smaller than the improvement between results (12) and (7) in Table II. This phenomenon supports the fact that these BERT-base models are not well trained [40]. Selecting RoBERTa-base as the basic encoder will be a better choice in sentence semantic modeling.

2) *Performance on SciTail Test*: Table III reports the experimental results on SciTail [8]. Compared with the other two NLI datasets, the SciTail dataset requires models to focus on entailment relation recognition. Moreover, both premise and hypothesis sentences are derived from the existing data source, which makes sentence pairs more linguistically challenging [8]. From the table, we can observe that, even using multilayer blocks, the performance of CSRAN is still 0.9% lower than *DRr-Net*. This observation indicates that our proposed DRr attention can deal with more complex and realistic situations. However, GRU used in *DRr-Net* still has some weaknesses in encoding input sentences, which leads *DRr-Net* to be 0.9% lower than fine-tuned transformer [17]. Compared with these BERT-based methods, we can observe that *LadRa-Net* with BERT-base can outperform RoBERTa on this dataset, which not only indicates the usefulness of DRr attention but also demonstrates that local structure can alleviate the weaknesses of BPE and boost the performance of DRr attention.

D. PI Results

In addition to the NLI task, we also select the PI task to verify the performance of our proposed models. Different from the NLI task, the PI task focuses on the symmetric semantic relation and requires an agent to classify whether

TABLE IV

EXPERIMENTAL RESULTS (ACCURACY) ON QUORA AND MSRP DATASETS

Model	Quora test	MSRP test
(1) CENN [35]	80.7%	76.4%
(2) L.D.C [69]	85.6%	78.4%
(3) BiMPP [70]	88.2%	-
(4) DIIN [53]	89.1%	-
(5) DRCN [1]	90.2%	82.5%
(6) RE2 [37]	89.3%	78.5%
(7) BERT-(base) [16]	91.0%	84.2%
(8) BERT-(large) [16]	91.4%	85.4%
(9) RoBERTa-(base) [40]	90.6%	87.1%
(10) ALBERT-(base) [68]	90.3%	88.6%
(11) <i>DRr-Net</i>	89.8%	82.9%
(12) <i>LadRa-Net</i> -BERT-(base)	92.1%	84.8%
(13) <i>LadRa-Net</i> -RoBERTa-(base)	91.3%	86.8%

TABLE V

EXPERIMENTAL RESULTS (F1) ON THE TWITTER-URL DATASET

Model	Twitter-URL Test
(1) Shortcut-Stacked Encoder [57]	0.650
(2) Decomposable Attention [72]	0.652
(3) InferSent [73]	0.746
(4) <i>ESIM_{seq}</i> [59]	0.752
(5) PWIM [74]	0.761
(6) RE2 [37]	0.724
(7) BERT-(base) [16]	0.765
(8) BERT-(large) [16]	0.768
(9) RoBERTa-(base) [40]	0.763
(10) ALBERT-(base) [68]	0.762
(11) <i>DRr-Net</i>	0.758
(12) <i>LadRa-Net</i> -BERT-(base)	0.772
(13) <i>LadRa-Net</i> -RoBERTa-(base)	0.767

two sentences express the same meaning, which is critical for question answering community, such as Quora and Zhihu.²

1) *Performance on Quora and MSRP*: Table IV reports the results on Quora and MSRP datasets. Still, our proposed methods achieve highly competitive performance by making full use of DRr attention and local structure. Besides, we observe that all methods have a better performance on the Quora dataset than the MSRP dataset. One possible reason is that the Quora dataset has more data (over 400k sentence pairs) than the MSRP dataset (only 5801 sentence pairs). In addition to the data size, we speculate that intersentence interaction is probably another possible reason. The Quora dataset contains many sentence pairs with less complicated interactions (e.g., many identical words in two sentences) [71]. Thus, all the methods can achieve better performance on the Quora dataset. When analyzing the performance on MSRP, we can observe that A Lite BERT (ALBERT)-base [68] achieves the best performance on the MSRP dataset. Meanwhile, our proposed *DRr-Net* outperforms RE2 by a large margin (+4.4%). On the other hand, *LadRa-Net* with RoBERTa-base is not even as good as RoBERTa-base (-0.3%). Since the MSRP dataset only contains 5801 sentence pairs, relatively complex structures are harder to optimize well, let alone a better performance. This phenomenon is also consistent with the performance of our proposed models on the SICK dataset, which is shown in Table II.

2) *Performance on Twitter-URL*: Table V shows the results (F1 value) on the Twitter-URL dataset [63]. In addition to the impressive performance, we also find some interesting phenomena. First, pairwise word interaction modeling (PWIM) [74] is the best model among all BERT-free baselines

and even has a highly competitive performance compared with BERT-based methods. By utilizing Bi-LSTM to encode the input sentences and a 19-layer deep CNN to integrate the features, PWIM makes a fully exploration on the local structure on sentences and achieves very competitive performance. However, the encoding and representing abilities of Bi-LSTM are still weaker than BERT. Thus, we observe that BERT-based methods still outperform PWIM on this dataset. On the contrary, *LadRa-Net* not only selects BERT as the encoder to obtain comprehensive representations of input sentences but also leverages local structure to alleviate the problems of BPE in BERT and strengthen the ability of our proposed DRr attention. Thus, we can observe that *LadRa-Net* achieves the best and the second best performances with different BERT-like encoders, surpassing these baselines.

E. Ablation Performance

The overall performance has proven the superiority of our proposed models. However, which part is more important for performance improvement is still unclear. Consequently, we conduct an ablation study on different test sets to examine the effectiveness of each component in *DRr-Net* and *LadRa-Net*. Since it is very flexible to change different basic encoders in *LadRa-Net*, we just examine the performance of *LadRa-Net* with BERT-base as the encoder and employ *LadRa-Net* to represent “*LadRa-Net* with BERT-base” for simplicity.

1) *Performance of DRr-Net*: First, we select the global contextual sentence vector \mathbf{h}^a as the initial vector of the GRU in DRr processing and \mathbf{h}^b as the guide vector of the dynamic selection operation. Therefore, *DRr-Net* can have a comprehensive understanding of sentence s^a and search the most relevant part with the consideration of sentence s^b and learned information at each step. When we remove each of them individually, we can obtain the results from Table VI(1) and (2). These results demonstrate that the performance of both models decreases by nearly 2.0%, which is indicative that the initial vector and guide vector are critical for deciding which part should be concerned more at each step.

Second, we are curious whether only one of the representations (i.e., global result or dynamic result) is enough for semantic matching. To this end, we remove the global results and reread results from *DRr-Net*. As illustrated in Table VI(3) and (4), both results are extremely important for classification. Among them, global results play a more important role, which is consistent with human behavior. People generally read the text from the beginning to the end and then continue reading till reaching the end so that they can understand the sentence semantics from a global perspective [25]. Moreover, humans have the capability to leverage enormous prior knowledge to extract the important parts, which is quite arduous for models to access the knowledge. Therefore, the performance of the input encoder is essential for our proposed models, and a better encoder (e.g., BERT) can boost the model performance.

2) *Performance of LadRa-Net*: There are three main components in *LadRa-Net*: 1) BERT encoder; 2) PCNN unit; and 3) DSA unit. Therefore, we first examine their effectiveness by removing each of them. Table VII(1)–(4) reports the corresponding results. We can observe that the BERT encoder

²<https://www.quora.com/>, <https://www.zhihu.com/>

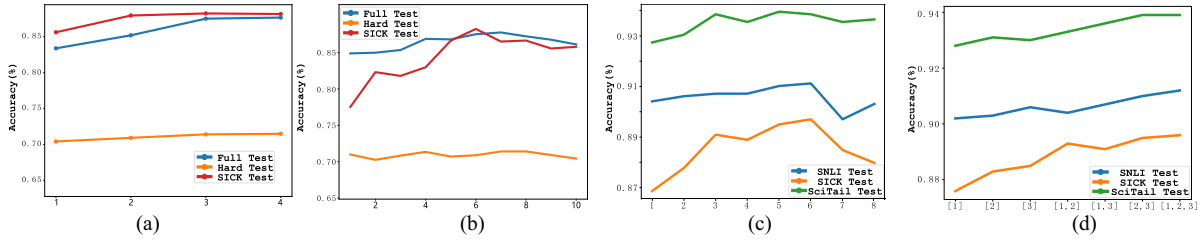


Fig. 4. Performance (accuracy) of *DRr-Net* and *LadRa-Net* with different hyperparameters, including the number stack layers (1–4) and reading lengths (1–10) for *DRr-Net*, and dynamical reading lengths (1–8) and different kernel size (1–3) for *LadRa-Net*. (a) Number of GRU layers in Stack-RNN in *DRr-Net*. (b) DRr length of the DRr unit in *DRr-Net*. (c) Dynamic reading length of DSA in *LadRa-Net*. (d) Different kernel size of PCNN in *LadRa-Net*.

TABLE VI

ABLATION PERFORMANCE (ACCURACY) OF *DRr-Net*

Model	SNLI test	SICK test
(1) <i>DRr-Net</i> (w/o initial vector)	85.3%	85.7%
(2) <i>DRr-Net</i> (w/o guide vector)	85.1%	86.1%
(3) <i>DRr-Net</i> (w/o global result)	81.2%	80.5%
(4) <i>DRr-Net</i> (w/o DRr result)	85.6%	86.4%
(5) <i>DRr-Net</i>	87.7%	88.3%

is still the most important component, which is consistent with the observation in the ablation performance in *DRr-Net*. As a new scheme of NLP, BERT-based methods can help models access knowledge from large language corpora and surpass other NN methods by a large percentage.

One of the main contributions of *LadRa-Net* is to model the local structure of input sentences. To verify the usefulness of the local structure, we remove the PCNN unit from *LadRa-Net* and add the PCNN unit into BERT-base for the evaluation. Compared with the results between Table VII(1) and (2), we can observe that the local structure can improve the performance of the BERT-base. Moreover, by considering the local structure, BERT-base can achieve the second best performance on the SNLI dataset. Meanwhile, the comparison between Table VII(3) and (7) also reports that the local structure can further boost the model performance. This phenomenon is also the same as the observation from Twitter-URL experiments in Section VI-D. All of these indicate the usefulness of local structure. Furthermore, when removing h_{ab} from the Label Prediction component and only taking the results of the DSA unit into account, model performance has a big drop. Since DSA only focuses on the important parts and local contexts, it lacks a global understanding of input sentences. Meanwhile, removing h_{ab} also means ignoring the superiority of BERT. Therefore, its performance would be greatly reduced.

F. Sensitivity of Parameters

As mentioned in Sections IV and V, two hyperparameters affect the performance of each model separately:

- 1) the stack layer number l_s of Stack-RNN and the dynamic reading length T_1 of DRr for *DRr-Net*;
- 2) the kernel size d_k of PCNN and the dynamic reading length T_2 of DSA for *LadRa-Net*.

Therefore, we evaluate their impact on NLI tasks with different hyperparameter settings. The results are reported in Fig. 4.

First, the most important hyperparameter that both models have is the dynamic reading length (i.e., T_1 and T_2). We observe from Fig. 4(b) that the performance of *DRr-Net* first becomes better with the increasing of reread length. When the reread length T_1 is between 5 and 7,

TABLE VII

ABLATION PERFORMANCE (ACCURACY) OF *LadRa-Net*

Model	SNLI test	SciTail test
(1) BERT-base	90.3%	93.1%
(2) BERT + PCNN	90.6%	93.3%
(3) BERT + DSA	90.4%	92.7%
(4) Without h_{ab}	85.6%	82.7%
(5) Stack-GRU + PCNN + DSA	88.2%	87.9%
(6) Max-pooling fusion	91.0%	93.7%
(7) <i>LadRa-Net</i> (BERT+ PCNN + DSA)	91.0%	93.9%

DRr-Net achieves the best performance. This phenomenon is consistent with the psychological findings that human attention focuses on nearly seven words [75]. When the length is bigger than 7, the accuracy of *DRr-Net* decreases to varying degrees. In other words, too short a reading length may be possible to cause the model to ignore some important parts, and too long a reading length may weaken the capability of capturing the important part, as well as understanding the semantic meaning intensively.

When it comes to *LadRa-Net*, we can obtain a similar phenomenon from Fig. 4(c). Different from *DRr-Net*, the best shot is around $T_2 = 5$, and *LadRa-Net* is more sensitive to the dynamic reading length. With the help of a BERT-based encoder and local structure utilization, *LadRa-Net* can achieve better performance with short reread length and save some time. However, *LadRa-Net* is more complex than *DRr-Net*, especially the BERT-base encoder. We need to tune the hyperparameters carefully for the model performance.

Second, Fig. 4(a) reports the model performance with different l_s 's. We can observe that the performance becomes better with the increase in l_s . With the continuous increase in GRU layers, the increasing rate of accuracy will slow down and even become worse on some test sets. With the consideration of model architecture, the scale of parameters will grow rapidly with the increasing of GRU layers, which may trigger the model hard to optimize, and even worse, the gradient might explode or vanish. Moreover, the encoding capability of GRU is not comparable with BERT. To this end, we replace GRU with BERT-base in *LadRa-Net*.

Third, Fig. 4(D) illustrates the performance of *LadRa-Net* with different combinations of kernel size d_k . We observe that the model performance will increase with the increase in d_k . As mentioned before, PCNN is capable of capturing the local structure to enhance semantic modeling. This result demonstrates that a large kernel size can help *LadRa-Net* to consider more local information and also proves the usefulness of the local structure. However, the effect of $d_k = 1$ is not obvious. With the consideration of model complexity, we select 2 and 3 as our final kernel sizes.

TABLE VIII
SOME EXAMPLES OF REREAD SEQUENCE AND THE CLASSIFICATION

Sentence	Re-read sequence	Gold Label	Predicted Label
s^a : a couple walk hand in hand down a street.	walk walk couple couple street street	Contradiction	Contradiction
s^b : a couple is sitting on a bench.	couple sitting sitting bench bench bench		
s^a : a person in a red shirt and black pants hunched over.	red shirt red shirt red shirt	Entailment	Entailment
s^b : a person wears a red shirt.	wears red shirt red shirt red		
s^a : a person with a purple shirt is painting an image of a woman on a white wall.	person painting painting woman woman woman	Neutral	Neutral
s^b : a woman paints a portrait of her best friend.	woman paints paints best best best		
s^a : the man in the black t-shirt is trying to throw something.	black t-shirt is trying throw throw	Entailment	Neutral
s^b : the man is in a black shirt.	black black shirt shirt shirt		
s^a : a person with a purple shirt is painting an image of a woman on a white wall.	person painting painting woman woman woman	Neutral	Entailment
s^b : a woman paints a portrait of a person.	paints portrait portrait portrait portrait person		

G. Case Study and Error Analysis

To visually demonstrate the validity of *DRr-Net* and *LadRa-Net*, we give some random examples from the SNLI dataset since the core idea of *DRr-Net* and *LadRa-Net* is the DRr attention. For simplicity, we list the dynamic reading sequences of *DRr-Net* on these examples. The results are shown in Table VIII.

For the first example in Table VIII, *DRr-Net* pours attention to words “walk, couple, street” in sentence s^a and “couple, sitting, bench” in sentence s^b . Then, *DRr-Net* repeatedly processes these important words for the final decision. From these words, we can conclude that the relation of this sentence pair was contradiction easily. Moreover, when checking the entailment relation in the second example, *DRr-Net* processes the same important words repeatedly, i.e., reading “red shirt” multiple times. In other words, *DRr-Net* does choose the important region and reread these important parts multiple times for the final decision. In order to better verify the ability of *DRr-Net*, we make an error analysis on the misclassification examples. In the fourth example, when sentence s^a contains more information than sentence s^b , *DRr-Net* may consider the unseen information more and make a misclassified decision. Moreover, when the sentence pair has very complex semantic relations, e.g., the last example in Table VIII, the model may be confused about its semantics and suffer from one of the important words, which leads to a wrong classification result.

VII. CONCLUSION AND FUTURE WORK

In this article, we presented a study on sentence semantic representation and matching. Specifically, we investigated that most attention-based methods attempted to select all the important parts in a static way, restricting the ability of the attention mechanism. Inspired by the dynamic embedding methods used in pretrained language models and the observation that people would constantly change their focal point for an in-depth understanding of sentences, we proposed a newly designed DRr attention, which is able to pay close attention to a small region of sentences at each time and reread the important parts for better sentence semantic matching. Based on this attention mechanism, we developed a novel *DRr-Net* for sentence semantic matching. Moreover, the GRU encoder used in *DRr-Net* still has some weaknesses in representing sentence semantics. Therefore, we employ a pretrained language model (e.g., BERT and RoBERTa) as the input encoder. One step further, the BPE used in pretrained language

models still suffered from incomplete modeling and fragile representation problems. In order to alleviate these problems and further boost the ability of DRr attention, we proposed to take local structures of sentences into consideration and extended *DRr-Net* to *LadRa-Net*. Finally, extensive experiments on two sentence semantic matching tasks (i.e., NLI and PI) demonstrated that *DRr-Net* could significantly improve the performance of sentence semantic matching. Furthermore, *LadRa-Net* was able to model sentence semantics more precisely and achieve better performance by considering the local structures of sentences.

In the future, we will focus on providing more information for dynamic rereading attention to better sentence semantic representing and matching. We also hope that our work could inspire relative research and lead to many future works.

REFERENCES

- [1] S. Kim, I. Kang, and N. Kwak, “Semantic sentence matching with densely-connected recurrent and co-attentive information,” in *Proc. AAAI*, 2019, pp. 6586–6593.
- [2] W. B. Dolan and C. Brockett, “Automatically constructing a corpus of sentential paraphrases,” in *Proc. IJW*, 2005, pp. 9–16.
- [3] P. Clark *et al.*, “Combining retrieval, statistics, and inference to answer elementary science questions,” in *Proc. AAAI*, 2016, pp. 2580–2586.
- [4] P. Wang, Q. Wu, C. Shen, and A. van den Hengel, “The VQA-machine: Learning how to use existing vision algorithms to answer new questions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1173–1182.
- [5] Q. Liu *et al.*, “Finding similar exercises in online education systems,” in *Proc. SIGKDD*, 2018, pp. 1821–1830.
- [6] I. V. Serban, A. Sordani, Y. Bengio, A. C. Courville, and J. Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” in *Proc. AAAI*, vol. 16, 2016, pp. 3776–3783.
- [7] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 632–642.
- [8] T. Khot, A. Sabharwal, and P. Clark, “SCITAIL: A textual entailment dataset from science question answering,” in *Proc. AAAI*, 2018, pp. 5189–5197.
- [9] S. Iyer, N. Dandekar, and K. Csernai, “First quora dataset release: Question pairs,” Tech. Rep., 2017. [Online]. Available: <https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs>
- [10] Y. Kim, “Convolutional neural networks for sentence classification,” 2014, *arXiv:1408.5882*. [Online]. Available: <http://arxiv.org/abs/1408.5882>
- [11] J. Cheng, L. Dong, and M. Lapata, “Long short-term memory-networks for machine reading,” in *Proc. EMNLP*, 2016, pp. 551–561.
- [12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, *arXiv:1412.3555*. <http://arxiv.org/abs/1412.3555>
- [13] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. NIPS*, 2017, pp. 5998–6008.
- [14] Z. Kun, L. Guangyi, W. Le, C. Enhong, L. Qi, and W. Han, “Image-enhanced multi-level sentence representation net for natural language inference,” in *Proc. IEEE ICDM*, Nov. 2018, pp. 747–756.

- [15] A. Talman, A. Yli-Jyrä, and J. Tiedemann, “Natural language inference with hierarchical BiLSTM max pooling architecture,” 2018, *arXiv:1808.08762*. [Online]. Available: <http://arxiv.org/abs/1808.08762>
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, *arXiv:1810.04805*. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” openAI, Tech. Rep., 2018.
- [18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” 2013, *arXiv:1310.4546*. [Online]. Available: <http://arxiv.org/abs/1310.4546>
- [19] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proc. EMNLP*, 2014, pp. 1532–1543.
- [20] J. Wang, H.-C. Chen, R. Radach, and A. Inhoff, *Reading Chinese Script: A Cognitive Analysis*. London, U.K.: Psychology Press, 1999.
- [21] C. Koch and N. Tsuchiya, “Attention and consciousness: Two distinct brain processes,” *Trends Cogn. Sci.*, vol. 11, no. 1, pp. 16–22, Jan. 2007.
- [22] W. Ma, Y. Cui, C. Si, T. Liu, S. Wang, and G. Hu, “CharBERT: Character-aware pre-trained language model,” 2020, *arXiv:2011.01513*. [Online]. Available: <http://arxiv.org/abs/2011.01513>
- [23] T. Brandt, J. Dichgans, and E. Koenig, “Differential effects of central versus peripheral vision on egocentric and exocentric motion perception,” *Exp. Brain Res.*, vol. 16, no. 5, pp. 476–491, Mar. 1973.
- [24] W. H. Warren and K. J. Kurtz, “The role of central and peripheral vision in perceiving the direction of self-motion,” *Perception Psychophysics*, vol. 51, no. 5, pp. 443–454, Sep. 1992.
- [25] Y. Zheng, J. Mao, Y. Liu, Z. Ye, M. Zhang, and S. Ma, “Human behavior inspired machine reading comprehension,” in *Proc. SIGIR*, 2019, pp. 425–434.
- [26] K. Zhang *et al.*, “DRR-Net: Dynamic re-read network for sentence semantic matching,” in *Proc. AAAI*, 2019, pp. 7442–7449.
- [27] L. Wu, L. Chen, R. Hong, Y. Fu, X. Xie, and M. Wang, “A hierarchical attention model for social contextual image recommendation,” *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 10, pp. 1854–1867, Oct. 2020.
- [28] M. Song, W. Zhao, and E. HaiHong, “KGAnet: A knowledge graph attention network for enhancing natural language inference,” *Neural Comput. Appl.*, vol. 32, no. 18, pp. 14963–14973, 2020.
- [29] L. Wu, Y. Yang, K. Zhang, R. Hong, Y. Fu, and M. Wang, “Joint item recommendation and attribute inference: An adaptive graph convolutional network approach,” in *Proc. SIGIR*, 2020, pp. 679–688.
- [30] Y. Liu, C. Sun, L. Lin, and X. Wang, “Learning natural language inference using bidirectional LSTM model and inner-attention,” 2016, *arXiv:1605.09090*. [Online]. Available: <http://arxiv.org/abs/1605.09090>
- [31] S. Peng, H. Cui, N. Xie, S. Li, J. Zhang, and X. Li, “Enhanced-RCNN: An efficient method for learning sentence similarity,” in *Proc. WWW*, 2020, pp. 2500–2506.
- [32] Z. Hu, Z. Fu, C. Peng, and W. Wang, “Enhanced sentence alignment network for efficient short text matching,” in *Proc. 6th Workshop NUT*, 2020, pp. 34–40.
- [33] S. Xu, S. E. and Y. Xiang, “Enhanced attentive convolutional neural networks for sentence pair modeling,” *Expert Syst. Appl.*, vol. 151, Aug. 2020, Art. no. 113384.
- [34] K. Cho, A. Courville, and Y. Bengio, “Describing multimedia content using attention-based encoder-decoder networks,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.
- [35] K. Zhang, E. Chen, Q. Liu, C. Liu, and G. Lv, “A context-enriched neural network method for recognizing lexical entailment,” in *Proc. AAAI*, 2017, pp. 3127–3134.
- [36] Y. Tay, L. A. Tuan, and S. C. Hui, “Co-stack residual affinity networks with multi-level attention refinement for matching text sequences,” 2018, *arXiv:1810.02938*. [Online]. Available: <http://arxiv.org/abs/1810.02938>
- [37] R. Yang, J. Zhang, X. Gao, F. Ji, and H. Chen, “Simple and effective text matching with richer alignment features,” in *Proc. ACL*, 2019, pp. 4699–4709.
- [38] J. Dong, M.-A. Rondeau, and W. L. Hamilton, “Distilling structured knowledge for text-based relational reasoning,” in *Proc. EMNLP*, 2020, pp. 6782–6791.
- [39] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized autoregressive pretraining for language understanding,” in *Proc. NeurIPS*, vol. 32, 2019, pp. 1–11.
- [40] Y. Liu *et al.*, “RoBERTa: A robustly optimized BERT pre-training approach,” 2019, *arXiv:1907.11692*. [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [41] M. Lewis *et al.*, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proc. ACL*, 2020, pp. 7871–7880.
- [42] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, “ERNIE: Enhanced language representation with informative entities,” 2019, *arXiv:1905.07129*. [Online]. Available: <http://arxiv.org/abs/1905.07129>
- [43] Y. Sun *et al.*, “ERNIE: Enhanced representation through knowledge integration,” 2019, *arXiv:1904.09223*. [Online]. Available: <http://arxiv.org/abs/1904.09223>
- [44] T. Liu, X. Wang, C. Lv, R. Zhen, and G. Fu, “Sentence matching with syntax-and semantics-aware bert,” in *Proc. COLING*, 2020, pp. 3302–3312.
- [45] X. Wang, Y. Chen, and W. Zhu, “A survey on curriculum learning,” 2020, *arXiv:2010.13166*. [Online]. Available: <http://arxiv.org/abs/2010.13166>
- [46] B. Xu, L. Zhang, Z. Mao, Q. Wang, H. Xie, and Y. Zhang, “Curriculum learning for natural language understanding,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6095–6104.
- [47] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, “Human attention in visual question answering: Do humans and deep networks look at the same regions?” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 932–937.
- [48] X. Li, J. Mao, C. Wang, Y. Liu, M. Zhang, and S. Ma, “Teach machine how to read: Reading behavior inspired relevance estimation,” in *Proc. SIGIR*, 2019, pp. 795–804.
- [49] C. Sen, T. Hartvigsen, B. Yin, X. Kong, and E. Rundensteiner, “Human attention maps for text classification: Do humans and neural networks focus on the same words?” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4596–4608.
- [50] W. Peng *et al.*, “Bi-directional cognitive thinking network for machine reading comprehension,” in *Proc. COLING*, 2020, pp. 2613–2623.
- [51] A. L. Yarbus, “Eye movements during perception of complex objects,” in *Eye Movements and Vision*. Boston, MA, USA: Springer, 1967, pp. 171–211.
- [52] H. Strasburger, I. Rentschler, and M. Jüttner, “Peripheral vision and pattern recognition: A review,” *J. Vis.*, vol. 11, no. 5, p. 13, 2011, doi: 10.1167/11.5.13.
- [53] Y. Gong, H. Luo, and J. Zhang, “Natural language inference over interaction space,” 2017, *arXiv:1709.04348*. [Online]. Available: <http://arxiv.org/abs/1709.04348>
- [54] D. Chen, A. Fisch, J. Weston, and A. Bordes, “Reading wikipedia to answer open-domain questions,” in *Proc. ACL*, 2017, pp. 1870–1879.
- [55] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith, “Annotation artifacts in natural language inference data,” in *Proc. NAACL-HLT*, 2018, pp. 107–112.
- [56] R. K. Srivastava, K. Greff, and J. Schmidhuber, “Highway networks,” 2015, *arXiv:1505.00387*. <http://arxiv.org/abs/1505.00387>
- [57] Y. Nie and M. Bansal, “Shortcut-stacked sentence encoders for multi-domain inference,” 2017, *arXiv:1708.02312*. [Online]. Available: <http://arxiv.org/abs/1708.02312>
- [58] E. Luuk and H. Luuk, “The redundancy of recursion and infinity for natural language,” *Cognit. Process.*, vol. 12, no. 1, pp. 1–11, Feb. 2011.
- [59] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen, “Enhanced LSTM for natural language inference,” in *Proc. ACL*, 2017, pp. 1657–1668.
- [60] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, *arXiv:1409.0473*. [Online]. Available: <http://arxiv.org/abs/1409.0473>
- [61] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, “Modeling coverage for neural machine translation,” 2016, *arXiv:1601.04811*. [Online]. Available: <http://arxiv.org/abs/1601.04811>
- [62] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, “SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment,” in *Proc. SemEval*, 2014, pp. 1–8.
- [63] W. Lan, S. Qiu, H. He, and W. Xu, “A continuously growing dataset of sentential paraphrases,” in *Proc. EMNLP*, 2017, pp. 1235–1245.
- [64] G. B. Orr and K.-R. Müller, *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2003.
- [65] Y. Tay, L. A. Tuan, and S. C. Hui, “A compare-propagate architecture with alignment factorization for natural language inference,” 2017, *arXiv:1801.00102*. [Online]. Available: <http://arxiv.org/abs/1801.00102>
- [66] J. Im and S. Cho, “Distance-based self-attention network for natural language inference,” 2017, *arXiv:1712.02047*. [Online]. Available: <http://arxiv.org/abs/1712.02047>

- [67] D. Yoon, D. Lee, and S. Lee, "Dynamic self-attention: Computing attention over words dynamically for sentence embedding," 2018, *arXiv:1808.07383*. [Online]. Available: <http://arxiv.org/abs/1808.07383>
- [68] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*. [Online]. Available: <http://arxiv.org/abs/1909.11942>
- [69] Z. Wang, H. Mi, and A. Ittycheriah, "Sentence similarity learning by lexical decomposition and composition," in *Proc. COLING*, 2016, pp. 1340–1349.
- [70] Z. Wang, W. Hamza, and R. Florian, "Bilateral multi-perspective matching for natural language sentences," 2017, *arXiv:1702.03814*. [Online]. Available: <http://arxiv.org/abs/1702.03814>
- [71] W. Lan and W. Xu, "Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering," in *Proc. COLING*, 2018, pp. 3890–3902.
- [72] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proc. EMNLP*, 2016, pp. 2249–2255.
- [73] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proc. EMNLP*, 2017, pp. 670–680.
- [74] H. He and J. Lin, "Pairwise word interaction modeling with deep neural networks for semantic similarity measurement," in *Proc. NAACL*, 2016, pp. 937–948.
- [75] G. Tononi, "Consciousness as integrated information: A provisional manifesto," *Biol. Bull.*, vol. 215, no. 3, pp. 216–242, Dec. 2008.



Kun Zhang (Member, IEEE) received the Ph.D. degree in computer science and technology from the University of Science and Technology of China, Hefei, China, in 2019.

He is currently a Faculty Member with Hefei University of Technology (HFUT), Hefei. He has published several articles in refereed journals and conferences, such as the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, the ACM Transactions on Knowledge Discovery from Data, AAAI, KDD, ACL, and ICDM. His

research interests include natural language understanding and recommendation systems.

Dr. Zhang was a recipient of the KDD 2018 Best Student Paper Award.



Guangyi Lv received the Ph.D. degree in computer science and technology from the University of Science and Technology of China, Hefei, China, in 2019.

He has published several articles in refereed conference proceedings, such as PAKDD'15, AAAI'16, and AAAI'17. His major research interests include deep learning, natural language processing, and recommendation systems.



Le Wu (Member, IEEE) received the Ph.D. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2015.

She is an Associate Professor with Hefei University of Technology (HFUT), Hefei. She has published several articles in refereed journals and conferences, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the ACM Transactions on Intelligent Systems and Technology, AAAI, IJCAI, KDD, SDM, and ICDM. Her general

areas of research are data mining, recommender systems, and social network analysis.

Dr. Wu was a recipient of the Best of SDM 2015 Award.



Enhong Chen (Senior Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 1996.

He is a Professor and a Vice Dean with the School of Computer Science, USTC. He has authored or coauthored more than 100 articles in refereed conferences and journals, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON MOBILE COMPUTING, KDD, ICDM, NIPS, and CIKM. He was on program committees of numerous conferences, including KDD, ICDM, and SDM. His research is supported by the National Science Foundation for Distinguished Young Scholars of China. His general area of research includes data mining and machine learning, social network analysis, and recommender systems.



Qi Liu (Member, IEEE) received the Ph.D. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2013.

He is a Professor with USTC. He has published prolifically in refereed journals and conference proceedings, e.g., the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the ACM Transactions on Information Systems, the ACM Transactions on Knowledge Discovery from Data, the ACM Transactions on Intelligent Systems and Technology, KDD, IJCAI, AAAI, ICDM, SDM, and CIKM. His general area of research is data mining and knowledge discovery.

Dr. Liu is a member of the ACM. He was a recipient of the ICDM 2011 Best Research Paper Award and the Best of SDM 2015 Award.



Meng Wang (Fellow, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively.

He is a Professor with Hefei University of Technology (HFUT), Hefei. He has authored or coauthored more than 200 book chapters, journals, and conference papers in these areas. His current research interests include multimedia content analysis, computer vision, and pattern recognition.

Dr. Wang was a recipient of the ACM SIGMM Rising Star Award 2014. He is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.