Multilevel Image-Enhanced Sentence Representation Net for Natural Language Inference

Kun Zhang, Guangyi Lv, Le Wu, Enhong Chen, *Senior Member, IEEE*, Qi Liu, *Member, IEEE*, Han Wu, Xing Xie, *Senior Member, IEEE*, and Fangzhao Wu

***Abstract*—Natural language inference (NLI) task requires an agent to determine the semantic relation between a premise sentence (*p*) and a hypothesis sentence (*h*), which demands sufficient understanding about sentences semantic. Due to the issues, such as polysemy, ambiguity, as well as fuzziness of sentences, intense sentence understanding is very challenging. To this end, in this article, we introduce the corresponding image of sentences as reference information for enhancing sentence semantic understanding and representing. Specifically, we first propose an image-enhanced multilevel sentence representation net (IEMLRN), that utilizes the image features from pretrained models for enhancing the sentence semantic understanding at different scales, i.e., lexical-level, phrase-level, and sentence-level. The proposed model advances the performance on NLI tasks by leveraging the pretrained global features of images. However, as these pretrained image features are optimized on specific image classification datasets, they may not show the best performance on NLI tasks. Therefore, we further propose to design an adaptive image feature generator that extracts fine-grained image labels from the corresponding sentences. After that, we extend the IEMLRN to multilevel image-enhanced sentence representation net (MIESR) by utilizing not only the coarse-grained pretrained features of images, but also the fine-grained adaptive features of images. Therefore, sentence semantic can be evaluated and enhanced more comprehensively and precisely. Extensive experiments on two benchmark datasets (SNLI, SICK) clearly show our proposed IEMLRN significantly outperform the state-of-the-art baselines, and our proposed MIESR model achieves the best performance by considering not only the text but also images in an adaptive multigranularities way.**

***Index Terms*—Image-enhanced representation, multiple level, natural language inference (NLI), sentence semantic.**

Manuscript received November 27, 2018; revised May 7, 2019; accepted July 18, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1000904 and Grant 2017YFB0803301, in part by the National Natural Science Foundation of China under Grant U1605251, Grant 61727809, Grant 61672483, and Grant 61602147, and in part by CCF-Tencent Open Fund. This article was recommended by Associate Editor K. Panetta. *(Corresponding author: Enhong Chen.)*

K. Zhang, G. Lv, E. Chen, Q. Liu, and H. Wu are with the Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: zhkun@mail.ustc.edu.cn; gylv@mail.ustc.edu.cn; cheneh@ustc.edu.cn; qiliuql@ustc.edu.cn; wuhanhan@mail.ustc.edu.cn).

L. Wu is with the School of Computer and Information, Hefei University of Technology, Hefei 230029, China (e-mail: lewu@hfut.edu.cn).

X. Xie and F. Wu are with the Social Computing Group, Microsoft Research Asia, Beijing 100190, China (e-mail: xing.xie@microsoft.com; wufangzhao@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2019.2932410

I. Introduction

NATURAL language inference (NLI) or recognizing textual entailment (RTE) task requires an agent to determine the semantic relation between two sentences among *entailment* (if the semantic of hypothesis can be concluded from the premise), *contradiction* (if the semantic of hypothesis cannot be concluded from the premise) and *neutral* (neither entailment nor contradiction). As depicted in the following example from [1], where the semantic relation is *entailment*.

p: Several airlines polled saw costs grow more than expected, even after adjusting for inflation.

h: Some of the companies in the poll reported cost increases.

NLI is known as a fundamental and yet challenging task for natural language understanding (NLU) [2]. It requires NLI models to understand the sentence semantic as comprehensive as possible and model the semantic relations between two sentences, and it has broad applications, e.g., information retrieval [3]–[7], question answering [8], as well as dialog system [9]. With respect to the granularity, NLI task can be classified into two categories: 1) lexical-level inference [10]–[12] and 2) sentence-level inference [13]–[15]. Lexical-level inference focuses on representing word semantic with different methods and identifying whether one word can entail another [10]. Sentence-level inference concerns more about the contents of entire texts and representations of sentence semantic [16]. With the availability of large annotated datasets, such as SNLI [16], multi-NLI [2], and the advancement of semantic representation techniques [17]–[19], researchers have proposed various end-to-end neural models to understand sentence semantic and evaluate the inference relations between sentences [20]–[22].

However, most of these models focus on the text itself and do not take into consideration the reference information (or context, such as images), which is essential for sentence semantic understanding. Sentence semantic suffers from the issues, such as polysemy, ambiguity, as well as fuzziness [23]. Moreover, sentence semantic is highly related to the context. The information of the sentence itself may be insufficient for precisely semantic understanding. As shown in Fig. 1, both the premise and hypothesis describe that people are shopping at the market, in which the weather information is different. The weather in hypothesis sentence is *sunny day*. However, it is fuzzy in premise sentence. Since the market is outside, we may conclude that the weather is *sunny*, but we are not sure about it. Thus, we may conclude the inference relation is *neutral* when

Fig. 1. Example from SNLI dataset.

only texts are provided. On the contrary, when providing the reference information, i.e., the image in Fig. 1(a), we can make a confident decision. The image, which is corresponding to the sentence pair in SNLI, provides the reference information for us to verify the uncertain content. Moreover, when the reference information becomes the image in Fig. 1(b), there is no doubt that the inference relation is *contradiction*. Therefore, it is urgent to take into consideration the reference information for sentence semantic understanding and inference relation evaluation.

In fact, vision-to-language (V2L) work [24]–[28] has proven that images convey important information of associated sentences. However, the information that images contain may relate to the sentence semantic at different scales, e.g., lexical, phrase, or the entire sentence. Meanwhile, there is a big difference between image information expression and sentence semantic expression. Inappropriate use of images may have a negative impact on sentence semantic understanding [29], [30], which will deteriorate the performance of NLI models. Therefore, it is critical to find an effective method to integrate the image reference information with text information for better sentence semantic understanding and representations.

In order to utilize the image reference information for NLI task, the main challenge lies in how to properly enhancing the sentence semantic understanding and representations by leveraging the image information. Since the sentence representations show a multigranularities manner with lexical level, phase level or sentence level, it is natural to relate the given context (image) with the sentence semantics from these multiple levels. To this end, in our preliminary work [31], we propose an image-enhanced multilevel sentence representation net (IEMLRN), a novel architecture that utilize the pretrained image features of images to enhances the sentence semantic understanding with different granularities. To be specific, we first utilize the pretrained VGG19 [32] to extract the image features. To capture the different granularities of the sentences, we integrate the information among texts and images with three different granularities, i.e., lexical-level, phrase-level, and sentence-level. In each level, we utilize the attention mechanism to allow the corresponding semantic level to focus on the most relevant parts of the input image features. Thus, sentence semantic understanding can be enhanced with the help of the image reference information, which is in favor of tackling NLI task.

However, in the real world, human language is highly abstract and designed specifically so as to communicate information among humans. In contrast, even the most carefully composed image is the culmination of a complex set of physical processes over which humans have little control [33]. Directly integrating the pretrained image features and texts in our preliminary work may be too-coarse to leverage the image information for semantic understanding. Therefore, the problem of how to align the image semantics with sentence semantics to narrow down the gap between the images and texts for NLI remains pretty much open. To this end, we focus on the multigranularities image features utilization and design the multilevel image-enhanced sentence representation net (MIESR), a novel architecture that takes both coarse-grained pretrained features and fine-grained adaptive features of images into consideration. Specifically, we propose to extract the fine-grained adaptive features of images, which are represented by the words mined from the training captions. The fine-grained adaptive features can represent images from a local perspective and narrow down the gap between images and texts, which is very important for this article. Meanwhile, the coarse-grained pretrained image features, which are used in IEMLRN, can reveal the image information from a global perspective. They together enhance the sentence semantic representations more comprehensively and precisely. Moreover, we proposed a newly designed multigranularities image-enhanced unit (MIEU) to integrate the pretrained image features, adaptive image features, and word semantic representations effectively, which is very helpful for enhancing sentence semantic representations more comprehensively. Finally, we conduct systematic experiments on two benchmark NLI datasets. The experimental results clearly show that IEMLRN can effectively improve the performance over state-of-the-art baselines, and MIESR achieves the best performance by considering both the coarse-grained pretrained features and fine-grained adaptive features of images.

The remainder of this article is organized as follows. In Section II, we introduce the related work. Next, the structure and technical details of our proposed models are given in Sections IV and V. Then, we make the experiments and detailed analysis in Section VI. Finally, we discuss and conclude this article in Section VII.

II. Related Work

In general, the related work can be grouped into three categories: 1) *NLI*: focusing on the different methods for tackling NLI task; 2) *V2L*: focusing on the recent researches on understanding language through vision; and 3) *NLI Data*: focusing on the work that pays attention to NLI data generation and analysis.

A. Natural Language Inference Methods

Due to data limitation, early works on NLI have been performed on small datasets with conventional methods [1], [10].

Turney and Mohammad [10] proposed the similarity differences hypothesis: The tendency of a to entail b is correlated with some learnable function of the differences in their similarities to a set of reference words. Based on this hypothesis, they proposed the *SimDiffs* method, a second-order feature vector representations of p and h, in which the features were the differences in the similarities of p and h to a set of reference words. Among these differences, some were important for entailment while others might tend to indicate a lack of entailment. The reference words they utilized included 2086 basic English words [34]. Zhang *et al.* [35] introduced the neural network into lexical-level inference and proposed a method called CENN to represent words semantic with different contexts and integrated these representations with the consideration of inference relations.

With the development of large annotated data, e.g., SNLI [16], multi-NLI [2], and various neural network architectures, such as LSTM [36], GRU [37], as well as attention mechanism [18], [38], [39], a variety of methods have been developed to represent and evaluate sentence semantic for NLI.

Among all these methods, sentence encoding-based methods play an important role. They focus on the semantic representation of each sentence, which is essential for plenty of natural language tasks, e.g., information retrieval [3], question answering [8], as well as dialog system [9]. For example, Bowman *et al.* [16] encoded the sentences with different LSTMs. Many related works followed this framework, using different neural networks as encoders [14], [20]. Liu *et al.* [20] proposed inner-attention to imitate the human's behavior that paid more attention to the important words when reading. Then, they utilized mean pooling to generate the sentence representations for NLI. Shen *et al.* [40] developed a directional and multidimensional attention model without RNN/CNN structure. They calculated the attention on each dimension of word representations and utilized a multidimensional attention to compress the sequence to generate the sentence representations, followed by a classification model to tackle the NLI task. Im and Cho [21] adopted the masked multihead attention with distance to explore the sentence semantic. Then, they utilized densely connected operation to preserve all the information for better sentence semantic representation. However, most of these methods focus on the text itself and do not take into consideration the reference information (or context, such as images), which is capable of providing necessary information for avoiding the sentence semantic issues, such as polysemy, ambiguity, as well as fuzziness.

B. Vision-to-Language Methods

In recent years, integrating the language and vision modalities has become a hot topic. Much progress has been achieved in plenty of V2L problems, such as image captioning [26], [41], visual question answering [33], visual dialog [42], as well as visual reasoning [43]–[45].

Among all these methods, utilizing an architecture which connects a CNN and an RNN to mapping from images to sentences directly has become a dominant trend. Mao *et al.* [24], for instance, proposed a multimodal RNN (m-RNN) to estimate the probability distribution of the next word given previous words and the feature representations of an image at each time step. Moreover, Ma *et al.* [46] used CNNs to both extract image features, as well as sentence features. Then, they fused the features together with a multimodal CNN to answer the input questions. Furthermore, Wu *et al.* [33] suspected that these CNN-RNN methods did not represent high-level semantic concepts for better sentence semantic understanding. They proposed to incorporate high-level concepts into these CNN-RNN approaches and achieved a big improvement in both image captioning and visual question answering. The above work inspired us to take full advantage of image information with different image utilization, i.e., coarse-grained pretrained image features from global perspective and fine-grained adaptive image features from local perspective, to model sentence semantic more comprehensively and precisely.

C. Works on NLI Data

With the development of large annotated NLI datasets, e.g., SNLI [16], multi-NLI [2], more and more neural networks have been proposed to represent sentence semantic and tackle the NLI task. However, these datasets were created by crowd workers. Specific linguistic phenomena, such as negation and vagueness would be highly correlated with certain inference classes [47], making it possible to identify the label by looking only at the hypothesis. Thus, based on SNLI dataset, Gururangan *et al.* [47] proposed a challenging hard test set, in which the examples that premise-oblivious model classified accurately were removed. They intended to better evaluate the performance of NLI models with this test set.

Besides, recent NLI models concerned more about the structures and global semantic of whole sentences, but less about external lexical knowledge, which led them to fail to capture many simple inferences that require lexical and world knowledge [48]. In order to evaluate the generalization ability of NLI models, Glockner *et al.* [48] proposed a simple but challenging lexical test set. Since this test set was created based on SNLI too, all the models trained on SNLI data could be tested for better evaluation. Table I gives several examples from SNLI test, hard test, as well as lexical text.

III. Problem Statement

In this section, we formulate the NLI task as a supervised classification problem. Given a premise sentence $\boldsymbol{s}^p = \{\boldsymbol{w}_1^p, \boldsymbol{w}_2^p, \ldots, \boldsymbol{w}_{l_p}^p\}$, a hypothesis sentence $\boldsymbol{s}^h = \{\boldsymbol{w}_1^h, \boldsymbol{w}_2^h, \ldots, \boldsymbol{w}_{l_h}^h\}$ and the corresponding image $\boldsymbol{I}$, our goal is to learn a classifier ξ which is able to precisely predict the inference relation $y = \xi(\boldsymbol{s}^p, \boldsymbol{s}^h, \boldsymbol{I})$ between $\boldsymbol{s}^p$ and $\boldsymbol{s}^h$. Here, $\boldsymbol{w}_i^p$ and $\boldsymbol{w}_j^h$ are one-hot vectors which represent the ith and the jth word in the sentences, and l_p and l_h indicate the total number of words in $\boldsymbol{s}^p$ and $\boldsymbol{s}^h$.

To achieve this goal, we propose IEMLRN and MIESR, and show the details in the following sections.

IV. IEMLRN Model

The overall architecture is shown in Fig. 2. In order to understand sentences with multiple granularities, we

TABLE I
SOME EXAMPLES FROM DIFFERENT SNLI TEST SETS

Test set	Premise	Hypothesis	Label
SNLI Test	A man looks intent while sculpting a gargoyle.	The man is working on art.	Entailment
		The man is at the bank.	Contradiction
Hard Test	Here is a picture of a man waiting for the bus to pick him up and he is hiding his face	The man is driving himself somewhere	Contradiction
		The man is going somewhere.	Neutral
Lexical Test	The man is wearing a yellow shirt and playing a piano	The man is wearing a yellow shirt and playing an instrument.	Entailment
		The man is wearing a yellow shirt and playing a french horn.	Contradiction

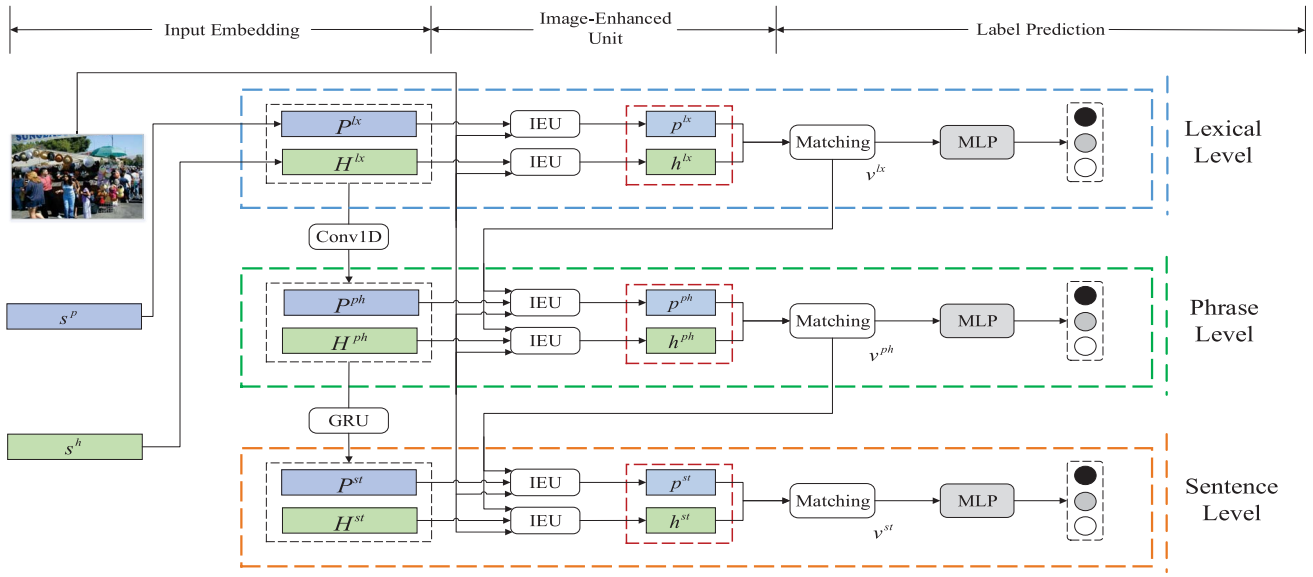


Fig. 2. Architecture of IEMLRN.

utilize three networks, i.e., lexical-level network, phrase-level network, and sentence-level network, as shown in the dashed boxes in Fig. 2.

Each network consists of three main components.

1) *Input Embedding:* Encoding the text inputs with different granularities, i.e., lexical-level, phrase-level, and sentence-level.
2) *Image-Enhanced Unit:* Generating the comprehensive and accurate sentence semantic representation with the image reference information.
3) *Label Prediction:* Utilizing the sentence representations from different granularities to predict the inference relation between sentences robustly.

Next, we will introduce the technical details of each component.

A. Input Embedding

The inputs of this article are two different types of data, i.e., images and texts. The image inputs of IEMLRN are represented by pretrained image features. As shown in Fig. 4(a), we select the pretrained VGG19 [32] to process the original image and employ the result of the last convolutional layer as the image feature representations. Then, we get the feature representation $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{l_c}\}, \boldsymbol{c}_i \in \mathbb{R}^{d_c}$, where d_c represents the dimension of each feature.

The text inputs of IEMLRN are one-hot representation sequences $\boldsymbol{s}^p = \{\boldsymbol{w}_1^p, \boldsymbol{w}_2^p, \ldots, \boldsymbol{w}_{l_p}^p\}$ for premise sentence and $\boldsymbol{s}^h = \{\boldsymbol{w}_1^h, \boldsymbol{w}_2^h, \ldots, \boldsymbol{w}_{l_h}^h\}$ for hypothesis sentence. In order to better represent each word, we utilize the concatenation of pretrained word embedding [49], character features [50], and syntactical features [47], [51] to represent each word in sentences. The character features are obtained by applying a convolutional neural network and a max-pooling to the learned character embeddings, which can represent words in a finer-granularity and help avoid the out-of-vocabulary (OOV) problem that pretrained word vectors suffer from. The syntactical features consist of the embedding of part-of-speech tagging feature, binary exact match feature, and binary antonym feature, which have been proved useful for sentence semantic understanding [47], [51], [52]. Then, we get the extravagant representations $\{\boldsymbol{p}_i^{lx} | i = 1, 2, \ldots, l_p\}$ and $\{\boldsymbol{h}_j^{lx} | j = 1, 2, \ldots, l_h\}$ for words $\{\boldsymbol{w}_i^p\}$ and $\{\boldsymbol{w}_j^h\}$ in premise and hypothesis sentences at lexical-level. Details about word embedding will be explained in Section VI-B.

However, these text representations only focus on lexical knowledge. Sentence semantic depends on not only lexical knowledge, but also other sentence features, such as word sequence, phrase structure, and the dependency among sentences. Thus, multilevel embedding methods are employed to encode the necessary information from different granularities.

To be specific, after getting the lexical-level representations $\{\boldsymbol{p}_i^{lx}\}$ and $\{\boldsymbol{h}_j^{lx}\}$ for premise and hypothesis sentences, we first concatenate these $\boldsymbol{p}_i^{lx}$ and $\boldsymbol{h}_j^{lx}$ by rows to form embedding matrices $\boldsymbol{P}^{lx} \in \mathbb{R}^{l_p * d}$ and $\boldsymbol{H}^{lx} \in \mathbb{R}^{l_h * d}$ for premise and hypothesis sentences. Then, one-dimensional (1-D) convolutions with different filter sizes (unigram, bigram,

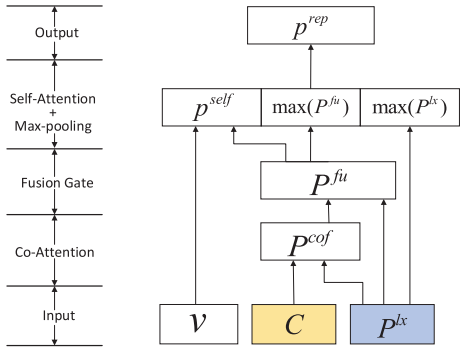


Fig. 3. Architecture of IEU.

and trigram) [8] are applied to them, followed by a max-pooling over different filters at each word. At last, we get the phrase-level representations $\boldsymbol{P}^{ph} \in \mathbb{R}^{l_p * d}$ and $\boldsymbol{H}^{ph} \in \mathbb{R}^{l_h * d}$, which extract the phrase structure information for sentence semantics as follows:

$$\boldsymbol{P}^{ph} = \text{Conv1D}\Big(\boldsymbol{P}^{lx}\Big), \ \ \boldsymbol{H}^{ph} = \text{Conv1D}\Big(\boldsymbol{H}^{lx}\Big). \tag{1}$$

Furthermore, in order to take the dependency, the words sequence, as well as the global semantic into consideration, we also send these phrase-level representations to a GRU [37] layer, resulting in the sentence-level representations $\boldsymbol{P}^{st} \in \mathbb{R}^{l_p * d}$ and $\boldsymbol{H}^{st} \in \mathbb{R}^{l_h * d}$, which can be formulated as follows:

$$\boldsymbol{p}_i^{st} = \text{GRU}\Big(\boldsymbol{p}^{ph}_{j=1,2,...,i}\Big), \ \ \boldsymbol{h}_i^{st} = \text{GRU}\Big(\boldsymbol{h}^{ph}_{j=1,2,...,i}\Big). \tag{2}$$

Therefore, we get different semantic representations at different granularities, i.e., lexical-level, phrase-level, and sentence-level. Then, we pass each level of representations to the following component to generate comprehensive and accurate sentence semantic representations with the corresponding image reference information.

B. Image-Enhanced Unit

As mentioned before, reference information is essential for sentence semantic understanding and helpful for comprehensive and accurate sentence semantic representations generating. However, how to make full use of reference information is still challenging. Among the core representation learning techniques, attention mechanism plays an important role. Attention mechanism is known for its alignment between representations, focusing one part of representation over another, and modeling the dependency regardless of sequence length [50]. Moreover, self-attention, which is a special case of attention mechanism, relates elements at different positions from a single sequence by computing the attention between each pair of tokens of the sequence [18], [40]. It is very flexible to model the long-range and local dependencies. Therefore, we intend to utilize attention mechanism to fully utilize the reference information for sentence semantics.

Fig. 3 shows the structure of image-enhanced unit (IEU). The inputs are one embedding sequence $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{l_s}\}$, one image feature sequence $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{l_c}\}$, as well as an inference relation vector $\boldsymbol{v}$ which we will introduce next. Please note that the embedding sequence $\boldsymbol{P}$ can be the premise vectors $\boldsymbol{P}^{lx}$, $\boldsymbol{P}^{ph}$, and $\boldsymbol{P}^{st}$ or the hypothesis vectors $\boldsymbol{H}^{lx}$, $\boldsymbol{H}^{ph}$, and $\boldsymbol{H}^{st}$ from different levels. Here, for simplicity, we take the lexical-level representations $\boldsymbol{P}^{lx}$ of the premise sentence as an example. We first employ co-attention [50] to model the relevance of each word in the premise sentence and the image features, which can be formulated as follows:

$$\begin{aligned} \alpha_{ij} &= \tanh\Big(\boldsymbol{p}_i^{lx} \boldsymbol{W}^{co} \boldsymbol{c}_j\Big) \\ \boldsymbol{p}_i^{\text{cof}} &= \sum_{j=1}^{l_c} \frac{\exp\big(\alpha_{ij}\big)}{\sum_{k=1}^{l_c} \exp\big(\alpha_{kj}\big)} \boldsymbol{c}_j, \quad i = 1, 2, \ldots, l_p \end{aligned} \tag{3}$$

where $\boldsymbol{p}_i^{\text{cof}}$ is actually a weight summation of the image context $\boldsymbol{c}_j$ for the *ith* word in the premise. $\boldsymbol{W}^{co} \in \mathbb{R}^{d * d^{'}}$ are the trainable parameters. According to *Similarity differences hypothesis* [10], the reference information can reveal some useful contents to indicate the inference relation between two sentences. Thus, we can utilize the most relevant information of image features to enhance the semantic understanding of each word in sentences.

After getting the representations $\{\boldsymbol{p}_i^{\text{cof}}\}$ from reference information, it is natural to consider integrating this representation and the original representation $\{\boldsymbol{p}_i^{lx}\}$. Inspired by GRU architecture, we introduce the fusion gate to integrate two types of representations [50], which can be formalized as follows:

$$\begin{aligned} \boldsymbol{z}_i &= \tanh\Big(\boldsymbol{W}_z\Big[\boldsymbol{p}_i^{lx}; \boldsymbol{p}_i^{\text{cof}}\Big] + \boldsymbol{b}_z\Big) \\ \boldsymbol{r}_i &= \sigma\Big(\boldsymbol{W}_r\Big[\boldsymbol{p}_i^{lx}; \boldsymbol{p}_i^{\text{cof}}\Big] + \boldsymbol{b}_r\Big) \\ \boldsymbol{f}_i &= \sigma\Big(\boldsymbol{W}_f\Big[\boldsymbol{p}_i^{lx}; \boldsymbol{p}_i^{\text{cof}}\Big] + \boldsymbol{b}_f\Big) \\ \boldsymbol{p}_i^{fu} &= \boldsymbol{r}_i \odot \boldsymbol{p}_i^{lx} + \boldsymbol{f}_i \odot \boldsymbol{z}_i \end{aligned} \tag{4}$$

where $\boldsymbol{W}_z, \boldsymbol{W}_r, \boldsymbol{W}_f \in \mathbb{R}^{d * 2d}$, and $\boldsymbol{b}_z, \boldsymbol{b}_r, \boldsymbol{b}_f \in \mathbb{R}^d$ are trainable parameters. tanh and σ are activation functions. $\odot$ is element-wise product. $\boldsymbol{z}_i$ can be regarded as the candidate activation of the inputs $[\boldsymbol{p}_i^{lx}; \boldsymbol{p}_i^{\text{cof}}]$. $\boldsymbol{r}_i$ is the update gate, which decides how much input $\boldsymbol{p}_i^{lx}$ will be retained. $\boldsymbol{f}_i$ is also the update gate, deciding how much candidate activation will be retained. By utilizing this fusion gate operation, we can integrate textual information as well as reference information. Thus, the semantics of each word can be represented in a more comprehensive way, which will be beneficial for sentence semantic understanding.

However, sentence semantic understanding requires not only lexical knowledge, but also the dependency and interaction of words among the sentence. In order to capture the dependency between words and significant properties in each sentence, we perform a variation of self-attention, a max-pooling on each fusion result, as well as max-pooling on the text input sequence. Then, we concatenate them together

$$\begin{aligned} \beta_i &= \boldsymbol{w}^T \sigma\Big(\boldsymbol{W}_\beta \boldsymbol{p}_i^{fu} + \boldsymbol{U}_\beta \boldsymbol{v} + \boldsymbol{b}_\beta\Big) \\ \boldsymbol{p}^{\text{self}} &= \sum_{i=1}^{p} \frac{\exp(\beta_i)}{\sum_{k=1}^{p} \exp(\beta_k)} \boldsymbol{p}^{fu}, \quad i = 1, 2, \ldots, l_p \end{aligned}$$

$$\boldsymbol{p}^{\text{rep}} = \left[\boldsymbol{p}^{\text{self}}; \max_i^p\left(\boldsymbol{p}^{fu}\right); \max_i^p\left(\boldsymbol{p}^{lx}\right)\right]. \quad (5)$$

Here, β_i denotes the attention weight of *ith* word in premise sentence before normalization. $\boldsymbol{p}^{\text{self}}$ represents the outputs of self-attention operation. $\boldsymbol{v}$ represents the inference relation vector of two sentences. Note that the input value of $\boldsymbol{v}$ depends on which level of the network is. For the lexical-level network, $\boldsymbol{v}$ will be zeros since it is the lowest level in our architecture. For a phrase-level network, $\boldsymbol{v}$ will be set as the output of the matching layer in lexical-level network, and so on. Details about computing that in matching layer will be discussed later.

As mentioned before, self-attention can solve the long-range dependency problem and choose the relevant information for sentence semantic. Since the sentence representations learning at each level aim at the same sentence, we try to inform the current level of the classification reason of the previous level. By utilizing self-attention operation, the model can grasp the most relevant parts for inference relation precisely and make the correct decision. Moreover, the max-pooling operation can select the most significant properties in each sentence and enhance the sentence representation. Therefore, self-attention and max-pooling together can generate a sufficient sentence representation, which is also the output of IEU.

As shown in the red box in Fig. 2, the sentence vector $\boldsymbol{p}^{\text{rep}}$ for the premise and $\boldsymbol{h}^{\text{rep}}$ for hypothesis represent the sentence semantics in a comprehensive way and guarantee the ability of models in sentence semantic understanding and inference relation classification.

C. Label Prediction

This component consists of two operations: 1) matching and 2) classification. In order to better evaluate the overall inference relation between two sentences, we employ the matching layer to integrate the information among the premise representation $\boldsymbol{p}^{\text{rep}}$ and hypothesis representation $\boldsymbol{h}^{\text{rep}}$. To be specific, we leverage heuristic matching methods to modify these representations, which can be formulated as follows:

$$\boldsymbol{v} = \text{relu}\left(\left[\left(\boldsymbol{p}^{\text{rep}}; \boldsymbol{h}^{\text{rep}}\right); \left(\boldsymbol{h}^{\text{rep}} - \boldsymbol{p}^{\text{rep}}\right); \left(\boldsymbol{h}^{\text{rep}} \odot \boldsymbol{p}^{\text{rep}}\right)\right]\right) \quad (6)$$

where [.; .] represents the concatenation operation, $\odot$ means element-wise product and relu is the nonlinear activation function. $\boldsymbol{v}$ is the inference relation vector of two sentences. In this operation, concatenation can retain all the information [35]. The element-wise product is a certain measure of "similarity" of premise and hypothesis [13]. Their differences can capture the degree of distributional inclusion on each dimension [53]. The output $\boldsymbol{v}$ will be used as the input of the classification layer. Besides, as mentioned in Section IV-B, it will also be sent to the IEU layer of the next level, e.g., $\boldsymbol{v}$ from lexical-level network is sent to phrase-level network.

After getting the inference relation vector $\boldsymbol{v}$, we utilize a multilayer perceptron (MLP) and one softmax output layer to classify the inference relation of two sentences. The output of this layer is the probability distribution of the inference relation between the sentence pair with the help of the reference information. The formulation is as follows:

$$P\left(y|\boldsymbol{s}^p, \boldsymbol{s}^h, \boldsymbol{I}\right) = \text{softmax}(\text{MLP}(\boldsymbol{v})). \quad (7)$$

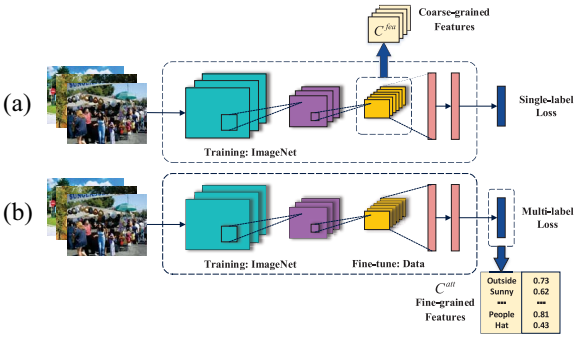


Fig. 4. Architecture of different image feature generators. (a) Pre-trained feature generator. (b) Adaptive feature generator.

We have to note that the input sentences are encoded at multiple levels. Thus, the probability $P(y|\boldsymbol{s}^p, \boldsymbol{s}^h, \boldsymbol{I})$ also can be calculated at multiple levels. As shown in Fig. 2, we utilize P^{lx}, P^{ph}, and P^{st} to represent the outputs of lexical-level network, phrase-level network, as well as sentence-level network. However, the lexical-level network or phrase-level network may misclassify the results due to the lack of global information of sentences. Utilizing these results may harm the performance of the models. Considering these results in the final classification result may harm the performance of the proposed model. Thus, the final classification result we use is the output of sentence-level network.

V. MIESR Model

In our IEMLRN model, we utilize pretrained VGG19 [32] to extract image features for enhancing sentence representations. These pretrained image features are designed on the external image classification datasets (e.g., ImageNet). Therefore, they mainly focus on features that contribute to the image classification and perform inferior to leverage the image information for sentence semantic understanding. Therefore, how to adaptively align the image semantics with sentence semantics to narrow down the gap between the images and texts for NLI remains pretty much open.

To this end, in this section, we focus on the utilization the image reference information and extend the current IEMLRN model to a novel MIESR model, which utilize both coarse-grained pretrained image features and fine-grained adaptive image features to enhance the sentence semantic representations more comprehensively and precisely. In the proposed model, our key contribution lies in designing an adaptive feature generator that extracts fine-grained image features by constructing a text vocabulary from the corresponding texts, and a multigranularities image enhanced unit that integrates multilevel image and text features for sentence semantic enhancement.

To be specific, our newly proposed MIESR also contains three networks, i.e., lexical-level network, phrase-level network, and sentence-level network. Each network also consists of three components: 1) input embedding; 2) MIEU; and 3) label prediction. In the input embedding, we propose the *Adaptive Feature Generator*, which is shown in Fig. 4(b), to predict the fine-grained adaptive image features. Thus, we

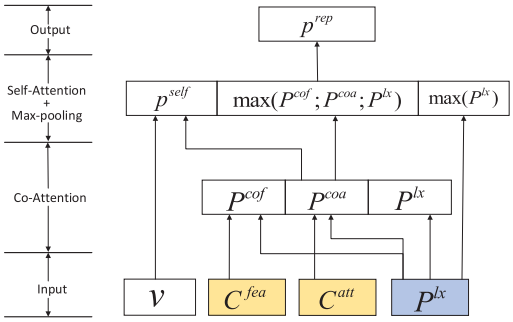


Fig. 5. Architecture of MIEU.

can represent image reference information not only with the coarse-grained features from pretrained VGG19 [32] but also with the fine-grained adaptive features. In MIEU component, which is shown in Fig. 5, we propose a more effective structure to integrate both the coarse-grained pretrained image features and fine-grained adaptive image features for better sentence representation enhancement. Since the text embedding and label prediction parts in our newly proposed MIESR are the same as the parts in our IEMLRN, we will only introduce the *Adaptive Feature generator* and MIEU in the following sections.

A. Adaptive Feature Generator

In order to explore the fine-grained image features and narrow down the gap between image and text, we leverage supervised learning to predict a set of adaptive features of images. Our newly adaptive feature generator is shown in Fig. 4(b). Since each image has several corresponding captioning sentences, we first build a text vocabulary with these captioning sentences. Thus, the most adaptive and salient features for each image can be extracted. To be specific, we first do POS tagging [54] on each word in the captioning sentences, including nouns, verbs, adjectives and so on. Then, we extract c most common words in each category and gather them together to determine the final text vocabulary.

Wu *et al.* [33] have proven that the most common words can reveal some semantics of images. However, the words with different parts of speech have different effects on semantic [35]. For example, nouns consider more about the objects in images, while prepositions pay more attention to the spatial relations among different objects. Moreover, the frequencies of words with different POS tags in sentences have a tremendous difference. Counting them with the same standard may conceal some important semantic concepts. In order to better extract adaptive features for images, we do POS tagging on words before counting their numbers. Along this line, we obtain a vocabulary with $c_{\text{pos}} = 465$ attributes. Table II shows some of the text features in parts of POS tagging categories.

With this text vocabulary, we can associate each image with a set of fine-grained adaptive features according to its captions. Then, a multilabel image dataset based on Flick30K [55] image captioning dataset and the attribute vocabulary is constructed. Next, we train an attribute extractor based on VGG19 model [32]. Fig. 4(b) shows the structure of the adaptive feature generator. We take advantage of the parameters in CNN layers and change the outputs of last fully connected layer to a c_{pos}-way softmax outputs. Moreover, we utilize the element-wise logistic loss function as the loss function of the attribute extractor, which can be formulated as follows:

$$J = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c_{\text{pos}}}\log\left(1+\exp\left(-y_i^j p_i^j\right)\right) \tag{8}$$

where y_i^j is a binary value whether the ith image contains the jth adaptive feature. p_i^j is the probability of the jth feature in the ith image that the adaptive feature generator predicts.

With the help of adaptive feature generator, we can predict the adaptive feature probability distribution $\boldsymbol{p}^{\text{att}}$ for the given image. However, this probability distribution can only describe the most likely semantic concepts of images. How to represent these semantic concepts properly is still challenging. Fortunately, each feature is represented by one word and extracted from the training captions. Thus, we can make use of the word embedding from the input embedding in IEMLRN to represent each attribute, in which the sentence semantic representation can be shared. Moreover, we multiply the word embeddings with the probability distribution. Along this line, the image is represented by adaptive features in sentence semantic space. We formulate this process as follows:

$$\begin{aligned}\boldsymbol{p}^{\text{att}} &= \text{Extractor}(\boldsymbol{I})\\ \boldsymbol{C}^{\text{att}} &= \boldsymbol{E}\odot\boldsymbol{p}^{\text{att}}\end{aligned} \tag{9}$$

where $\boldsymbol{C}^{\text{att}}$ represents the fine-grained adaptive feature representations of the image. $\boldsymbol{E}$ denotes the word embeddings from input embedding in IEMLRN.

TABLE II
ADAPTIVE FEATURE WORDS IN DIFFERENT POS TAGGING

POS Tagging	Adaptive Feature Words
Determiner	two, three, another, many, every, some, half, nine
Noun	people, man, woman, child, dog, jean, tree, hat clothes
Verb	stand, sit, walk, play, run, hold, smile, jump, take, shoot
Adjective	small, young, large, old, high, dark, high, white, blue
Preposition	in, on, from, over, behind, outside, across, above, indoor

B. Multigranularities Image-Enhanced Unit

In the input embedding, we have obtained the multilevel text representations and multigranularities image feature representations. How to integrate coarse-grained pretrained image features and fine-grained adaptive image features with text information at each level is still challenging. Thus, we propose the MIEU for better sentence representation enhancement.

In our IEMLRN [31], we utilize attention mechanism and fusion operation to process the image and text information, as shown in Fig. 3. However, this architecture is redundant. We have extracted the most relevant information with co-attention and self-attention operations. The importance of fusion gate is not so obvious. Therefore, we replace the fusion gate with concatenation operation in our newly proposed MIEU, which is shown in Fig. 5. To be specific, we concatenate the pretrained feature representations $\{\boldsymbol{p}_i^{\text{cof}}\}$, adaptive feature representations

$\{\boldsymbol{p}_i^{coa}\}$, as well as the original embeddings $\{\boldsymbol{p}_i^{lx}\}$. Then, we utilize the self-attention and max-pooling operations to capture the dependency between words and significant properties, which is the same as we did in *IEU*. Meanwhile, (4) and (5) will be modified as follows:

$$\begin{aligned}\boldsymbol{p}_i^{fu} &= \left[\boldsymbol{p}_i^{lx}; \boldsymbol{p}_i^{\text{cof}}; \boldsymbol{p}_i^{coa}\right] \\ \beta_i &= \boldsymbol{w}^T \sigma\left(\boldsymbol{W}_\beta \boldsymbol{p}_i^{fu} + \boldsymbol{U}_\beta \boldsymbol{v} + \boldsymbol{b}_\beta\right) \\ \boldsymbol{p}^{\text{self}} &= \sum_{i=1}^{p} \frac{\exp(\beta_i)}{\sum_{k=1}^{p} \exp(\beta_k)} \boldsymbol{p}^{fu}, \quad i = 1, 2, \ldots, l_p \\ \boldsymbol{p}^{\text{rep}} &= \left[\boldsymbol{p}^{\text{self}}; \max_i^p\left(\boldsymbol{p}^{fu}\right); \max_i^p\left(\boldsymbol{p}^{lx}\right)\right] \end{aligned} \tag{10}$$

where $\boldsymbol{p}_i^{coa}$ is the weight summation of the adaptive feature representations $\boldsymbol{C}^{\text{att}}$ for the *ith* word in the premise. By utilizing MIEU, MIESR is capable of measuring not only coarse-grained image features but also fine-grained image features. Thus, it can narrow down the gap between images and texts, and model the sentence semantic more comprehensively and precisely, which is essential for NLI.

VI. Experiment

In this section, we will first introduce the datasets that we evaluate the models on and the baselines that our proposed models compared with. Then, we will give a detailed analysis of the models and experimental results.

A. Data Description

In this section, we introduce two datasets, we evaluate the models on. Different from our preliminary work [31], we replace the DanMu dataset with SICK dataset, since the latter has more reliable images and sentences matching relations.

1) Standford Natural Language Inference (SNLI): SNLI [16] has 570k human-annotated sentence pairs. The premise sentences were drawn from the captions of Flickr30k corpus [55], and the hypothesis sentences were manually composed. The labels we use are *entailment*, *neutral*, and *contradiction*. Since each premise was drawn from the training captions, we extract the corresponding image from Flickr30k dataset as the reference information. In order to better evaluate the performances of models, we also select the challenging hard subset [47] and lexical subset [48] as our test sets. Table I gives several examples from different SNLI test sets.

2) Sentences Involving Compositional Knowledge (SICK): SICK dataset [56] consists of about 10 000 English sentence pairs, generated from two existing sets: 1) the 8K ImageFlickr dataset [57] and 2) the STS MSR-video description dataset.[1] It is a randomly selected subset of sentence pairs from each of these sources. Each sentence pair is annotated for the semantic inference relations (*entailment*, *contradiction*, and *neutral*) by means of crowdsourcing techniques. In order to make the results more reliable, we only extract the corresponding reference information from the 8K ImageFlickr dataset.

[1]https://www.cs.york.ac.uk/semeval-2012/

TABLE III
Statistical Information of Each Test Set

Test set	Data Size			Average Token Count	
	E	C	N	premise	hypothesis
SNLI Full Test	3,368	3,237	3,219	13.91	7.48
SNLI Hard Test	1,058	1,135	1,068	13.81	7.71
SNLI Lexical Test	782	5,164	43	11.42	11.60
SICK Test	2,243	239	2,513	11.69	11.47

Since we introduce the image information into the original dataset, a small part of the data that do not contain the corresponding images will be removed. Therefore, we recount the basic statistics and show them in Table III.

B. Model Learning

In this section, we will introduce the details about the model learning, which consists of two parts: 1) loss function and 2) model initialization.

1) Loss Function: Since it is a classification problem, we utilize *cross-entropy* as the loss function. The following is the loss function of the lexical-level network, where n is the number of training examples:

$$L = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \log P\left(y_i | \boldsymbol{s}_i^p, \boldsymbol{s}_i^h, \boldsymbol{I}_i\right) \tag{11}$$

$\boldsymbol{y}_i$ is the one-hot representation for the true class of the *ith* example, and $P(y_i|\boldsymbol{s}_i^p, \boldsymbol{s}_i^h, \boldsymbol{I}_i)$ is the probability distribution over the classes that our proposed models output. As mentioned in Section IV-C, each network in our model has an output. We intend that each network in our models should make the correct classification. Therefore, we apply cross-entropy function to each-level output. Considering the model complexity, we also add the L2-norm of all parameters in *IEU* or MIEU to the entire loss function. Then, we get the loss function for the whole model as follows:

$$L = L^{lx} + L^{ph} + L^{st} + \epsilon \|\boldsymbol{\theta}\|_2. \tag{12}$$

2) Model Initialization: In order to get the best performance, we have tuned the hyper-parameters on the validation set. Specially, we utilized the validation set to monitor the training process. If the loss on validation set did not decrease in 1000 batches, we would stop the training process and select the trained model that has the best performance on the validation set as the final model, in which all the parameters are determined.

For pretrained image feature generator, we utilize the VGG19 [32] in Keras[2] to process the images and employ the result of the last convolutional layer as the image feature representations. For adaptive image feature generator, we modify the output layer in VGG19 to change it from single-label classification to multilabel classification. Then, we fine-tune the generator on the Flickr30k corpus [55].

For both IEMLRN and MIESR, we set the word embedding dimension as 300, character-level embedding level as 100, phrase-level embedding are also set as 100, the dropout as 0.6, and ϵ as 0.01. The word embedding we use are obtained

[2]https://keras.io/

TABLE IV
PERFORMANCE (ACCURACY) OF MODELS ON SNLI AND SICK DATASETS

Model	#Paras	Full test	Hard test	Lexical test	SICK test
(1) LSTM encoders [59]	3.0m	80.6%	58.5%	52.3%	81.8%
(2) Inner-Attention BiLSTM [20]	2.8m	84.5%	62.7%	58.6%	85.2%
(3) CENN [36]	≈700k	82.1%	60.4%	51.9%	82.5%
(4) Gated-Att BiLSTM [14]	12m	85.5%	65.5%	65.6%	85.7%
(5) CAFE [60]	3.7m	85.9%	66.1%	65.5%	86.1%
(6) Distance-based Self-Attention [21]	4.7m	86.3%	67.4%	68.5%	86.7%
(7) DRCN [22]	5.6m	86.5%	68.3%	69.4%	87.4%
(8) CENN with image [36]	≈700k	83.1%	61.7%	66.8%	84.2%
(9) NIC [42]	-	84.7%	63.6%	67.1%	85.5%
(10) m-RNN [24]	-	85.1%	64.9%	69.4%	85.9%
(11) VQA-model [61]	-	79.7%	56.2%	62.4%	84.9%
(12) *IEMLRN*	3.9m	**87.5**%	**75.4**%	**78.1**%	**87.7**%
(13) *MIESR*	3.7m	**87.8**%(+0.3)	**76.8**%(+1.4)	**78.7**%(+0.6)	**88.3**%(+0.6)

from a pretrained word vectors (840B GloVe) [49]. The hidden state size of GRU is 512. The hidden size in co-attention and self-attention calculation is set as 200. The sizes of two-layer MLP in label prediction layer are set as 512 and 256. To initialize the model, we randomly initialize all weights such as $\boldsymbol{W}_\beta$ following the uniform distribution in the range between $-\sqrt{6/(n\text{in}+n\text{out})}$ and $\sqrt{6/(n\text{in}+n\text{out})}$ as suggested by [61]. All biases such as $\boldsymbol{b}_\beta$ are initialized as zeros. We use Adam optimizer with learning rate 10^{-4}. During implementation, we utilize tensorflow[3] and *Photinia*[4] to build our entire model.

C. Baselines

In this part, we compare our model against the following start-of-the-art sentence-encoding baselines.

1) *LSTM Encoders [58]:* Leveraging different LSTMs to encode the premise and hypothesis sentences.
2) *CENN [35]:* Integrating different context for better sentence semantic representations.
3) *Inner-Attention BiLSTM [20]:* Utilizing inner-attention to extract the important parts for sentence representations.
4) *Gated-Att BiLSTM [14]:* Using intrasentence gated-attention method to generate sentence representations.
5) *CAFE [59]:* Utilizing a compare, compress and propagate architecture to generate sentence representations.
6) *Distance-Based Self-Attention [21]:* Utilizing self-attention and distance mask to model the local and global dependencies among sentences for NLI.
7) *DRCN [22]:* Using a densely connected co-attentive recurrent neural network (RNN) to preserve all the information for better sentence representations.

We also select three V2L models to better verify the performances of IEMLRN and MIESR. Since these models aim to generate image descriptions or predict the scores of candidate answers, we just leveraging their fusion representations of images and sentences. Then, we employ the same label prediction component as our proposed models did.

1) *NIC [41]:* A neural network consisting of a vision CNN followed by a language RNN.
2) *m-RNN [24]:* Utilizing a deep RNN for sentences and a deep CNN for images to generating the captioning.
3) *VQA Model [60]:* Adopting a combined bottom-up and top-down attention mechanism to better model the interaction between images and sentences.

D. Experimental Results

In this section, we utilize the accuracy on different test sets to evaluate the performance of all models.

1) Overall Performance: The overall results are summarized in Table IV. We can observe that IEMLRN and MIESR achieve state-of-the-art performance on all test sets. To be specific, the corresponding images are introduced as reference information. Thus, the sentence semantic of the premise and hypothesis can be evaluated more comprehensively and precisely. For example, the image can be helpful for distinguishing the exact weather in premise sentence, as shown in Fig. 1. Moreover, IEMLRN integrates the reference information and evaluates the inference relation between two sentences with different granularities, which means IEMLRN can understand the sentence semantic from lexical knowledge to global semantic. Furthermore, MIESR utilizes coarse-grained pretrained features and fine-grained adaptive features to represent the image from different granularities. Thus, MIESR has the capability to narrow down the gap between texts and images, and take full advantage of image information for better sentence semantic representation enhancement.

LSTM encoder [58] utilizes different LSTMs to encode sentences and leads many related works, such as inner-attention BiLSTM [20] and CENN [35]. However, they encode each sentence separately. The interactions between two sentences, which are essential for NLI, have not been utilized effectively. Distance-based self-attention [21] and DRCN [22] are current state-of-the-art sentence encoding-based models. The former utilizes the masked multihead attention with distance to model the sentence semantic. The latter adopts densely connected co-attentive network to generate sentence representations. They are capable of modeling sentences from multiple aspects comprehensively. However, they take only the text information into consideration, which is insufficient for tackling the issues that sentence semantic suffers from. Moreover, if the words in

[3]https://www.tensorflow.org/
[4]https://github.com/XoriieInpottn/photinia

TABLE V
ACCURACY(%) OF IEMLRN AND MIESR WITH DIFFERENT IMAGE SETTINGS OVER EACH INFERENCE RELATION

Model		Original images			W/O Images			Foil images		
		Full	Hard	Lexical	Full	Hard	Lexical	Full	Hard	Lexical
IEMLRN	Entailment	88.3%	81.4%	72.3%	88.1%(-0.2)	74.5%(-6.9)	76.2%(+3.9)	87.0%(-1.3)	74.2%(-7.2)	58.3%(-14.0)
	Contradiction	89.2%	79.7%	79.4%	85.0%(-4.2)	70.0%(-9.7)	68.4%(-11.0)	84.9%(-4.3)	65.6%(-14.1)	60.8%(-18.6)
	Neutral	84.9%	64.9%	27.2%	82.0%(-2.9)	58.1%(-6.8)	23.3%(-3.9)	77.1%(-7.8)	55.3%(-9.6)	23.3%(-3.9)
	Overall	87.5%	75.4%	78.1%	85.1%(-2.4)	67.4%(-8.0)	69.1%(-9.0)	83.1%(-4.4)	65.0%(-10.4)	60.2%(-17.9)
MIESR	Entailment	88.3%	82.3%	73.3%	88.1%(-0.2)	74.5%(-7.8)	76.2%(+2.9)	86.2%(-2.1)	69.1%(-13.2)	60.7%(-12.6)
	Contradiction	88.6%	81.5%	79.9%	85.0%(-3.6)	69.8%(-11.7)	68.4%(-11.5)	86.3%(-2.3)	69.1%(-12.4)	69.5%(-10.4)
	Neutral	86.4%	66.3%	34.9%	82.0%(-4.4)	58.2%(-8.1)	23.3%(-11.6)	80.0%(-6.4)	58.0%(-8.3)	30.2%(-4.7)
	Overall	87.8%	76.8%	78.7%	85.1%(-2.7)	67.4%(-9.4)	69.1%(-9.6)	84.2%(-3.6)	65.3%(-11.5)	68.1%(-10.6)

sentences have high overlap, they might be treated as the same sentences, which has a bad influence on the final decision.

2) Performance on Hard and Lexical Test: As mentioned in Section II-C, hard test and lexical test are capable of better evaluating the performances of NLI models. We observe from Table IV that IEMLRN outperforms all baselines by a large margin, e.g., distance-based self-attention model (+8.0%) and gated-Att BiLSTM model (+9.9%). Meanwhile, with most of the models perform worse on lexical test than their own performances on hard test, IEMLRN and MIESR achieve the accuracy 78.1% and 78.7% separately, which are 2.7% and 1.9% higher than their own performances on Hard test. These phenomena suggest that our proposed models have better generalization ability and grasp the lexical knowledge indeed. Moreover, MIESR achieves better performance over IEMLRN, which indicates that utilizing both the coarse-grained pre-trained features and fine-grained adaptive features of images will be quite helpful for enhancing the sentence representation.

3) Comparison Between IEMLRN and MIESR: As illustrated in Tables IV and V, MIESR achieves better performance than IEMLRN. We can obtain that MIESR has a more stable improvement on hard test (1.4%) and lexical test (0.6%), which indicates its superiority. Moreover, we can observe that MIESR not only achieve better performance on challenging test sets (i.e., hard test and lexical test), but also on challenging category (i.e., *Neutral*). As shown in Table V, MIESR achieves the accuracy 86.4%, 66.3%, and 34.9% on the *Neutral* category of different test sets, which are 1.5%, 1.4%, and 7.7% higher than the performances of IEMLRN. These phenomena indicate that by utilizing both the coarse-grained pre-trained features and fine-grained adaptive features of images, MIESR is capable of understanding sentence semantic more precisely and comprehensively. Thus, it can better deal with more complex situations.

E. Analysis of the Importance of Images

In this article and our previous work [31], we introduce the images as reference information into NLI. There are three important questions should be answered to validate the importance of image reference information.

1) Since the premise sentences are drawn from the training captions of images, whether the premise sentences can be replaced by the corresponding images.
2) Our proposed models achieve the best performance. How many of the improvements are achieved by the image reference information. Whether the corresponding images are beneficial for all inference relations.
3) Since the image is invisible when annotating the data, image introduction may change the inference relation between sentences. How is the impact degree of the image reference information. Does this change invalidates the model evaluation when still using the original labels.

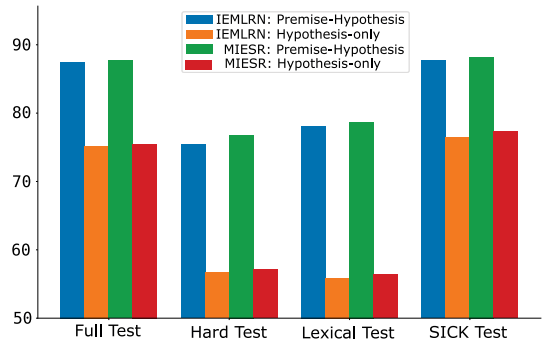


Fig. 6. Comparison between premise-hypothesis and hypothesis-only.

In order to answer these questions, we did extensive experiments and detailed analysis in the following part.

1) Problem A: The focus of this problem is whether the premise sentences can be replaced by the corresponding images. If so, we can just evaluate the inference relation with the corresponding images and hypothesis sentences, and achieve comparable performance. Thus, we remove all the premise data and evaluate the models on all test sets.

As shown in Fig. 6, there is a big gap between the performances of complete data and the hypothesis-only data. Moreover, the performances on the challenging hard test and lexical test become a lot worse. This phenomenon indicates that the images cannot replace the premise sentences. Even though the premise sentences describe the content in the images, there is still a big difference between them. The images can be treated as reference information to enhance the sentence semantic understanding and assist inference relation classification, but they cannot be treated as the replacement of the premise sentences. In other words, this article still focuses on the inference relation between sentences and has a big difference with image and sentence retrieval alike work.

2) Problem B: The concentration of this problem is the importance of images. Thus, we further explore the impact

TABLE VI
ACCURACY ON IEMLRN AND MIESR WITH DIFFERENT IMAGE SETTINGS

Model	Test set	Images	W/O Images	Foil Images
IEMLRN	Full	87.5%	85.1%	83.1%
	Hard	75.4%	67.4%	65.0%
	Lexical	78.1%	69.1%	60.2%
MIESR	Full	87.8%	85.1%	84.2%
	Hard	76.8%	67.4%	65.3%
	Lexical	78.7%	69.1%	68.1%

TABLE VII
LABEL CHANGING COLLECTION WITH IMAGES ON FULL AND HARD TESTS. (E: ENTAILMENT, C: CONTRADICTION, N: NEUTRAL, O: OVERALL)

Re-Labeled	Full Test				Hard Test			
	E	C	N	O	E	C	N	O
Entailment	313	4	71	**388**	297	6	94	**397**
Contradiction	11	314	53	**378**	17	311	54	**382**
Neutral	16	12	206	**234**	26	13	182	**221**
Overall	*340*	*330*	*330*	1000	*340*	*330*	*330*	1000

of different image settings in IEMLRN and MIESR by comparing their performance on full test and hard test of SNLI. Tables V and VI show the corresponding overall and detailed results. *W/O Images* represents removing the inputs of image features. *Foil Images* represents replacing the original image with an unrelated random one from the whole image set.

Compared with the original images, the performances without images or with foil images both have a big drop, in which foil image setting leads to worse performance. It indicates that images play an important role in sentence semantic understanding, and foil images will introduce more irrelevant information that deteriorates the model performances. Moreover, we can observe that MIESR has better performance on *neutral* relation recognition, which is the hardest to classify among three relations. This phenomenon indicates that multigranularities representations of images are really helpful for sentence representation enhancement. Furthermore, we observe that IEMLRN has better performance over lexical test on *entailment* relation without images than with original images. Since this test set needs explicit lexical knowledge and precise attention on the different words, only coarse-grained pretrained image features may be insufficient for inference relation classification. On the contrary, MIESR adopts both coarse-grained pretrained features and fine-grained adaptive features for images. Therefore, we can observe from Table V that MIESR has very stable performance on lexical test too.

3) Problem C: The core of this problem is to figure out whether the image introduction will invalidate the model evaluation when we still use the original labels. First of all, we sampled 1000 instances (340 entailment, 330 contradiction, and 330 neutral) from full test and hard test separately, and invited ten NLP researchers to reannotate these instances. Inspired by confusion matrix, we utilize Table VII to illustrate the statistic results in each category. Each column (row) represents the count of gold labels (relabeled labels) in each category. For example, the first column indicates that 313 *entailment* instances are relabeled *Entailment*, 11 *entailment* instances are relabeled *Contradiction*, and 16 *entailment* instances are relabeled *Neutral*. We utilize Cohen's κ to validate the consistency of relabeled results and original results. The inner-annotator agreement κ is 0.749 for full test, 0.685 for hard test, and 0.717 for all of them, which indicate the credibility of relabeled data. This phenomenon illustrates that the image introduction will not invalidate the model evaluation when the gold labels are still used.

Moreover, the biggest impact of image reference information is that some of *neutral* instances are relabeled *entailment* or *contradiction*. Since the image reference information is invisible when annotating the data, it is natural that the images provide necessary information to avoid the ambiguity or fuzziness that sentence semantic suffers from. As mentioned in [47], *neutral* hypothesis is often generated by introducing the plausible information, e.g., *talking on the phone* to *talking to his wife on the phone*, or replacing approximates with exact descriptions, e.g., *some people* to *four adults*. For the latter situation, image reference information provides sufficient information, or grounded information [62] to distinguish the inference relation. Thus, image reference information will be helpful for avoiding this kind of annotation artifacts. As for the former situation, though image reference information cannot provide corresponding information, it can help the model distinguish the plausible information and classify these instances into *neutral* category. In other words, though the image reference information may change the gold labels, it does not invalidate the model evaluation, but helps to avoid some annotation artifacts.

F. Ablation Performance

In the previous section, we have proven that image information has a big impact on the sentence semantic understanding. However, the importance of multilevel sentence semantic representations in our proposed models is still unclear. Thus, we conduct an ablation study to examine the effectiveness of each component. The results are shown in Table VIII.

From the results, our proposed models perform better when consider multigranularities text representations. When considering multigranularities image representations, MIESR performs better than IEMLRN. These phenomena indicate that considering the semantics of sentences and images from multiple granularities is important and necessary for semantic understanding. In other words, fine-grained representations consider more about the local information, while coarse-grained representations concern more about global information. They all should be considered for better sentence representation enhancement and inference relation classification.

Moreover, compared with *IEU*, we remove the fusion gate from MIEU and declare that this operation is redundant in Section V-A. However, whether this operation is redundant is still unclear. Therefore, we also make experiments to validate the performance of IEMLRN and MIESR by retaining or removing the fusion gate. The results are shown in Table VIII (7) and (8). As illustrated from the results, we can obtain that the performance of fusion gate is not consistent, with

Fig. 7. Comparison of gold labels, changing labels, and predicted labels in some examples.

TABLE VIII
ABLATION PERFORMANCE (ACCURACY) OF MODELS ON SNLI AND SICK DATASETS

Model	Full Test		Hard Test		Lexical Test		SICK Test	
	IEMLRN	*MIESR*	*IEMLRN*	*MIESR*	*IEMLRN*	*MIESR*	*IEMLRN*	*MIESR*
(1) Only Lexical Feature	34.3%	42.5%	42.1%	45.6%	67.5%	70.1%	51.7%	58.5%
(2) Only Phrase Feature	52.7%	58.4%	45.2%	48.5%	66.2%	69.4%	57.5%	63.8%
(3) Only Sentence Feature	76.2%	79.3%	57.9%	60.3%	65.0%	68.7%	62.3%	69.2%
(4) Lexical-Phrase Features	65.5%	69.7%	64.5%	68.5%	69.4%	72.2%	63.3%	68.7%
(5) Lexical-Sentence features	82.9%	83.1%	66.8%	67.2%	74.6%	75.6%	70.2%	71.8%
(6) Phrase-Sentence Features	83.2%	83.8%	65.7%	66.4%	73.2%	74.7%	71.5%	72.4%
(7) With Fusion Gate	87.5%	87.6%	75.4%	77.8%	78.1%	78.5%	87.7%	88.7%
(8) Without Fusion Gate	87.6%	87.8%	75.6%	76.8%	74.8%	78.7%	87.4%	88.3%
(9) Whole architecture	**87.5%**	**87.8%**	**75.4%**	**76.8%**	**78.1%**	**78.7%**	**87.7%**	**88.3%**

little improvement on some test sets. However, this part also performs worse on other test sets. With the consideration of the complexity of models, we remove the fusion gate from MIESR.

G. Error Analysis

As discussed in Section VI-E, the image plays an important role in precise sentence semantic understanding. However, it would modify the gold labels and import some unrelated information to some degree. In order to better validate the importance of images and the performance of proposed models, we make error analysis on several misclassification examples, which is shown in Fig. 7. Next, we will group these examples into three categories and analyze them separately.

1) Wrong Classification: For the top left example, the main differences between two sentences are *clothing store* and *restaurant*. However, the image reference information makes the model focus on the working worker and the outside place. Thus, both the models are confused about the sentence semantic and make the wrong classification *neutral*. For the top right example, the models meet the same circumstance. The image makes IEMLRN confused about the place that the bird stands. Fortunately, attribute representations give necessary information for MIESR to make the right decision.

2) Gold Label Changing: The image reference information provides more information for the sentences. Thus, the gold labels of some instances might be changed. The middle left and middle right examples indicate that the images provide sufficient information to distinguish the inference relations between two sentences. Thus, the inference relation will become clearer. Both of our proposed models utilize this information and do the correct classification. Moreover, Section VI-E shows that most of the affected examples are *neutral* examples. The image reference information can provide necessary information for this situation, which proves that image does play an important role in sentence semantic understanding.

3) External Knowledge: Though our proposed models make full use of images to help to understand sentence, there are still some weaknesses. From the bottom left example, we find that both the sentences describe the same person. However, the hypothesis presents some human prior knowledge, which is difficult for the models. The image cannot tackle this problem. Thus, both models make the wrong classification. The bottom right example demonstrates the same phenomenon. In this example, both the sentences are highly consistent with the image. Thus, the models classify this example into *entailment* category. However, there are still some differences between the words *woman runs* and *runner*, which need more human prior knowledge to distinguish. Therefore, the inference relation should be *neutral*.

H. Case Study

In order to better validate the proposed multilevel architecture, we visualize the self-attention at different levels in IEMLRN since it focuses on understanding sentence semantic with different granularities. Fig. 8 shows the attention distribution and classification probability distribution of each level over the example shown in Fig. 1.

Compared with the attention results at the three-level structure, we can observe that IEMLRN pay more attention to the important words, i.e., *shopping, outside market* in

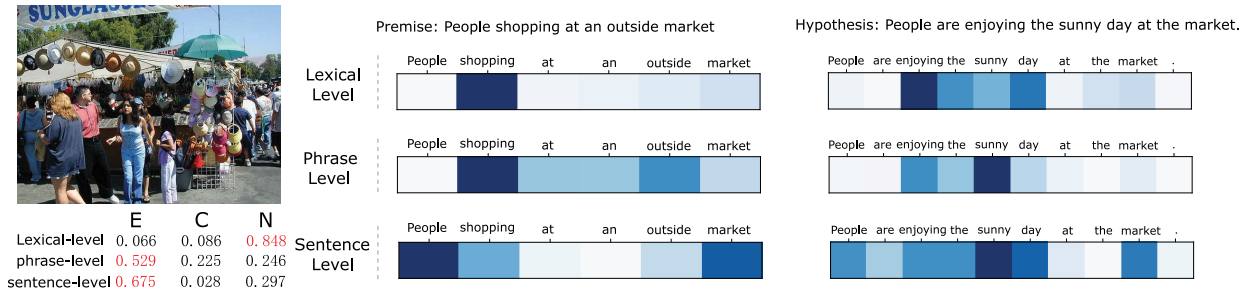


Fig. 8. Visualization of self-attention with different text granularities.

premise and *enjoy, sunny day, market* in hypothesis. This indicates that IEMLRN pays attention to not only the weather information but also people's activity. Therefore, sentence semantic can be evaluated and represented more comprehensively and precisely. Moreover, the classification probability distributions at each level indicate that IEMLRN makes a wrong classification at lexical level since the absence of global semantic is missing. When taking more global semantic (i.e., phrase-level or sentence-level information) into consideration, IEMLRN turns to the right classification result and becomes more and more confident about the decision. In other words, by evaluating sentence semantic and relations from lexical knowledge to global semantic, our proposed models are capable of achieving good performance in NLI task.

VII. Conclusion

In this article, we presented a study on NLI. Specifically, we introduced the corresponding images of sentences as reference information into NLI for sentence representation enhancement. Moreover, we proposed an IEMLRN, a novel architecture that allowed the model to utilize the image reference information to understand sentence semantic from lexical knowledge to global semantics. By integrating vision and language information from multilevel granularities, i.e., lexical-level, phrase-level, and sentence-level, IEMLRN can model the sentence semantic comprehensively and accurately. Furthermore, we explored the image utilization and extended the IEMLRN to MIESR, which adopted coarse-grained pretrained features and fine-grained adaptive features of images. This architecture was capable of narrowing down the gap between images and texts, as well as enhancing the sentence semantic representations more comprehensively. Finally, experimental results on two benchmark NLI datasets demonstrated that IEMLRN and MIESR were able to understand sentence semantic, generate sentence representation, and evaluate the inference relation between sentences in a comprehensive and precise way.

In the future, we will consider more different reference information and more efficient processing methods (e.g., objection detection) for more precise sentence semantic understanding and representations. Since there has been relatively little work on utilizing reference information directly for inference relation classification, we hope this article could inspire the relative researches and lead to many future works.

References

[1] B. MacCartney, *Natural Language Inference*, Stanford Univ., Stanford, CA, USA, 2009.
[2] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proc. NAACL*, 2018, pp. 1112–1122.
[3] P. Clark *et al.*, "Combining retrieval, statistics, and inference to answer elementary science questions," in *Proc. AAAI*, 2016, pp. 2580–2586.
[4] J. Su, "Representation and inference of user intention for Internet robot," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 8, pp. 995–1002, Aug. 2014.
[5] G. Liu, Y. Wang, M. A. Orgun, and E.-P. Lim, "Finding the optimal social trust path for the selection of trustworthy service providers in complex social networks," *IEEE Trans. Services Comput.*, vol. 6, no. 2, pp. 152–167, Apr./Jun. 2013.
[6] 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.*, to be published.
[7] G. Liu *et al.*, "MCS-GPM: Multi-constrained simulation based graph pattern matching in contextual social graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1050–1064, Jun. 2018.
[8] 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.*, vol. 4, 2017, pp. 3909–3918.
[9] I. V. Serban, A. Sordoni, 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. 3776–3784.
[10] P. D. Turney and S. M. Mohammad, "Experiments with three approaches to recognizing lexical entailment," *Nat. Lang. Eng.*, vol. 21, no. 3, pp. 437–476, 2015.
[11] M. Baroni, R. Bernardi, N.-Q. Do, and C.-C. Shan, "Entailment above the word level in distributional semantics," in *Proc. EACL*, 2012, pp. 23–32.
[12] L. Kotlerman, I. Dagan, I. Szpektor, and M. Zhitomirsky-Geffet, "Directional distributional similarity for lexical inference," *Nat. Lang. Eng.*, vol. 16, no. 4, pp. 359–389, 2010.
[13] L. Mou *et al.*, "Natural language inference by tree-based convolution and heuristic matching," in *Proc. ACL*, vol. 2, 2016, pp. 130–136.
[14] Q. Chen, X.-D. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Recurrent neural network-based sentence encoder with gated attention for natural language inference," in *Proc. RepEval@EMNLP*, 2017, pp. 36–40.
[15] T. Munkhdalai and H. Yu, "Neural tree indexers for text understanding," in *Proc. Conf. Assoc. Comput. Linguist. Meeting*, vol. 1, 2017, p. 11.
[16] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proc. EMNLP*, 2015, pp. 632–642.
[17] 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.
[18] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[19] Q. Liu *et al.*, "Finding similar exercises in online education systems," in *Proc. SIGKDD*, 2018, pp. 1821–1830.
[20] Y. Liu, C. Sun, L. Lin, and X. Wang, "Learning natural language inference using bidirectional LSTM model and inner-attention," *CoRR*, vol. abs/1605.09090, 2016.
[21] J. Im and S. Cho, "Distance-based self-attention network for natural language inference," *CoRR*, vol. abs/1712.02047, 2017.

[22] S. Kim, J.-H. Hong, I. Kang, and N. Kwak, "Semantic sentence matching with densely-connected recurrent and co-attentive information," *CoRR*, vol. abs/1805.11360, 2018.

[23] X. Zheng, J. Feng, Y. Chen, H. Peng, and W. Zhang, "Learning context-specific word/character embeddings," in *Proc. AAAI*, 2017, pp. 3393–3399.

[24] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Deep captioning with multimodal recurrent neural networks (M-RNN)," *CoRR*, vol. abs/1412.6632, 2014.

[25] J. Atif, C. Hudelot, and I. Bloch, "Explanatory reasoning for image understanding using formal concept analysis and description logics," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 5, pp. 552–570, May 2014.

[26] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. IEEE CVPR*, 2015, pp. 1473–1482.

[27] K. Ramasamy and D. Ganesan, "A systematic analysis of transform coefficients and block decomposition for texture enhancement with orthogonal polynomials model," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 12, pp. 3245–3255, Dec. 2017.

[28] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2018.

[29] J. Zhang, L. Cui, Y. Fu, and F. B. Gouza, "Fake news detection with deep diffusive network model," *CoRR*, vol. abs/1805.08751, 2018.

[30] L. Cui, Z. Chen, J. Zhang, L. He, Y. Shi, and P. S. Yu, "Multi-view collective tensor decomposition for cross-modal hashing," in *Proc. ACM ICMR*, 2018, pp. 73–81.

[31] K. Zhang *et al.*, "Image-enhanced multi-level sentence representation net for natural language inference," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, 2018, pp. 747–756.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[33] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE CVPR*, 2016, pp. 203–212.

[34] C. K. Ogden, *Basic English: A General Introduction With Rules and Grammar*, vol. 29, K. Paul and T. Trench, Eds. London, U.K.: Kegan, 1944.

[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] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proc. EMNLP*, 2016, pp. 551–562.

[37] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014.

[38] L. Wu *et al.*, "Personalized multimedia item and key frame recommendation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 1431–1437.

[39] K. Zhang *et al.*, "DRr-Net: Dynamic re-read network for sentence semantic matching," in *Proc. AAAI*, 2019, pp. 7442–7449.

[40] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "DiSAN: Directional self-attention network for RNN/CNN-free language understanding," *CoRR*, vol. abs/1709.04696, 2017.

[41] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE CVPR*, 2015, pp. 3156–3164.

[42] A. Das *et al.*, "Visual dialog," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2017, pp. 326–335.

[43] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. CVPR*, 2017, pp. 1988–1997.

[44] A. Suhr, M. Lewis, J. Yeh, and Y. Artzi, "A corpus of natural language for visual reasoning," in *Proc. ACL*, vol. 2, 2017, pp. 217–223.

[45] K. Zhang, G. Lv, E. Chen, L. Wu, Q. Liu, and C. P. Chen, "Context-aware dual-attention network for natural language inference," in *Proc. Pac.–Asia Conf. Knowl. Disc. Data Min.*, 2019, pp. 185–198.

[46] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *Proc. AAAI*, vol. 3, 2016, pp. 3567–3573.

[47] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," in *Proc. NAACL*, 2018, pp. 107–112.

[48] M. Glockner, V. Shwartz, and Y. Goldberg, "Breaking NLI systems with sentences that require simple lexical inferences," in *Proc. ACL*, 2018, pp. 650–655.

[49] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[50] Y. Gong, H. Luo, and J. Zhang, "Natural language inference over interaction space," *CoRR*, vol. abs/1709.04348, 2017.

[51] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *Proc. ACL*, 2017, pp. 1870–1879.

[52] M. T. Mills and N. G. Bourbakis, "Graph-based methods for natural language processing and understanding—A survey and analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 1, pp. 59–71, Jan. 2014.

[53] J. Weeds, D. Clarke, J. Reffin, D. Weir, and B. Keller, "Learning to distinguish hypernyms and co-hyponyms," in *Proc. COLING*, 2014, pp. 2249–2259.

[54] A. Voutilainen, "Part-of-speech tagging," in *The Oxford Handbook of Computational Linguistics*. Oxford, U.K.: Oxford Univ. Press, 2003, pp. 219–232.

[55] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 67–78, Feb. 2014.

[56] 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.

[57] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, no. 1, pp. 853–899, 2013.

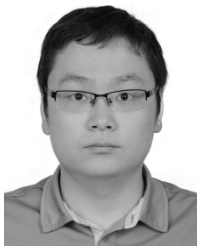
[58] S. R. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C. D. Manning, and C. Potts, "A fast unified model for parsing and sentence understanding," in *Proc. ACL*, 2016, pp. 1466–1477.

[59] Y. Tay, L. A. Tuan, and S. C. Hui, "A compare-propagate architecture with alignment factorization for natural language inference," *CoRR*, vol. abs/1801.00102, 2017.

[60] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. CVPR*, vol. 3, 2018, p. 6.

[61] G. B. Orr and K.-R. Müller, *Neural Networks: Tricks of the Trade*. Heidelberg, Germany: Springer, 2003.

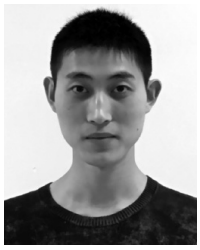
[62] H. T. Vu *et al.*, "Grounded textual entailment," in *Proc. COLING*, 2018, pp. 2354–2368.



Kun Zhang received the B.E. degree in computer science and technology from the University of Science and Technology of China, Hefei, China, in 2014, where he is currently pursuing the Ph.D. degree in natural language processing with the School of Computer Science and Technology.

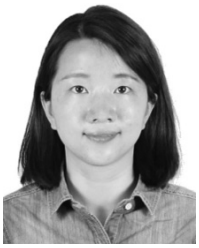
He has published several papers in refereed conference proceedings, such as AAAI, KDD, and ICDM. His current research interests include natural language processing and text mining.

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



Guangyi Lv received the B.E. degree in computer science and technology from Sichuan University, Chengdu, China, in 2013. He is currently pursuing the Ph.D. degree in natural language processing with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China.

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



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

She is currently a Faculty Member with the Hefei University of Technology, Hefei. She has published several papers in referred journals and conferences, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *ACM Transactions on Intelligent Systems and Technology*, AAAI, IJCAI, KDD, SDM, and ICDM. Her current research interests include data mining, recommender system, and social network analysis.

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

Enhong Chen (SM'07) received the Ph.D. degree in data mining and machine learning from the University of Science and Technology of China (USTC), Hefei, China, in 1996.

He is a Professor and the Vice Dean of the School of Computer Science, USTC. He has published over 100 papers in refereed conferences and journals, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MOBILE COMPUTING, KDD, ICDM, NIPS, and CIKM. His research was supported by the National Science Foundation for Distinguished Young Scholars of China. His current research interests include data mining and machine learning, social network analysis, and recommender systems.

Prof. Chen was on the program committees of numerous conferences, including KDD, ICDM, and SDM.

Qi Liu (M'15) received the Ph.D. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2013.

He is an Associate Professor with USTC. He has published prolifically in refereed journals and conference proceedings, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *ACM Transactions on Information Systems*, *ACM Transactions on Knowledge Discovery From Data*, *ACM Transactions on Intelligent Systems and Technology*, KDD, IJCAI, AAAI, ICDM, SDM, and CIKM. His current research interests include datamining and knowledge discovery.

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

Han Wu received the B.E. degree in software engineering from the Dalian University of Technology, Dalian, China, in 2015. She is currently pursuing the Ph.D. degree in data mining with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China.

She has published several papers in refereed conference proceedings, such as International Joint Conference on Artificial Intelligence, IEEE International Conference on Data Mining, and the International Conference on Database Systems for Advanced Applications. Her current research interest includes data mining, with a focus on patent analysis such as patent litigation prediction and patent technology representation.

Xing Xie (SM'09) received the B.S. and Ph.D. degrees in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 1996 and 2001, respectively.

He is currently a Senior Researcher with Microsoft Research Asia, Beijing, China, and a Guest Ph.D. Advisor with USTC. His current research interests include spatial data mining, location-based services, social networks, and ubiquitous computing.

Dr. Xie was involved in the program or organizing committees of over 70 conferences and works. Especially, he initiated the LBSN workshop series and served as the Program Co-Chair of ACM Ubicomp 2011. He is a Senior Member of ACM, and a Distinguished Member of China Computer Federation.

Fangzhao Wu received the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2017.

He is currently an Associate Researcher with Microsoft Research Asia, Beijing. He has published several papers in TKDE, ACL, SIGIR, AAAI, CIKM, and ICDM. His current research interests include natural language processing and data mining.