# Image-Enhanced Multi-Level Sentence Representation Net for Natural Language Inference

Kun Zhang[1], Guangyi Lv[1], Le Wu[2], Enhong Chen[1,*], Qi Liu[1], Han Wu[1], Fangzhao Wu[3]
[1]Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China
{zhkun, gylv, wuhanhan}@mail.ustc.edu.cn, {cheneh, qiliuql}@ustc.edu.cn
[2]Hefei University of Technology, China, lewu@hfut.edu.cn
[3] Microsoft Research Asia, China, wufangzhao@gmail.com

***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 from lexical knowledge to global semantic. Due to the issues such as polysemy, ambiguity, as well as fuzziness of sentences, fully understanding sentences is still challenging. To this end, we propose an Image-Enhanced Multi-Level Sentence Representation Net (*IEMLRN*), a novel architecture that is able to utilize the image to enhance the sentence semantic understanding at different scales. To be specific, we introduce the corresponding image of sentences as reference information, which can be helpful for sentence semantic understanding and inference relation evaluation. Since image information might be related to the sentence semantics at different scales, we design a multi-level architecture to understand sentences from different granularity and generate the sentence representation more precisely. Experimental results on the large-scale NLI corpus and real-world NLI alike corpus demonstrate that *IEMLRN* can simultaneously improve the performance. It is noteworthy that *IEMLRN* significantly outperforms the state-of-the-art sentence-encoding based models on the challenging hard subset and challenging lexical subset of SNLI corpus.**

***Index Terms*—Natural Language Inference, Sentence Semantic, Image-Enhanced Representation, Multiple Level**

## 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, which has broad applications, e.g. information retrieval [3], question answering [4], as well as dialog system [5]. With respect to the granularity, NLI task can be classified into two categories: Lexical-level inference [6]–[8] and Sentence-level inference [9]–[11]. Lexical-level inference focuses on representing word semantic with different methods and identifying whether one word can entail another [6], [12]. Sentence-level inference concerns more about the content of entire texts and representation of sentence semantics [13]. With the availability of large annotated datasets, such as SNLI [13], MultiNLI [2], and the advancement of semantic representation technique [14]–[17], researchers have proposed various end-to-end neural models to understand sentence semantic and evaluate the inference relations between sentences.


Fig. 1. Example from SNLI dataset.

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. Moreover, it is highly related to the context. The information of sentence itself may be insufficient for precise semantic understanding. As shown in Figure 1, both the premise and hypothesis describe that people are shopping at the market, while the weather information is different. The weather in hypothesis sentence is "sunny day", while it is fuzzy in premise sentence. Since the market is outside, we may conclude that the weather is "sunny", which we are not sure about. Thus, we may conclude the inference relation is neutral when texts are

provided singly. On the contrary, when providing the reference information, i.e. the image in Figure 1(A), we can make a confident decision. The image, which is the 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 Figure 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, image captioning work [18]–[21] have 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-level, phrase-level, or the whole sentence. Inappropriate use of the image reference information may have a negative impact on sentence semantic understanding [22], [23], which is crucial for natural language inference. Therefore, it is critical to find an effective method to integrate the image reference information into sentence semantic understanding and representation.

To this end, we propose an Image-Enhanced Multi-Level Sentence Representation Net (*IEMLRN*), a novel architecture that is able to utilize the image to enhance the sentence semantic understanding at different scales to tackle the above issue. To be specific, we introduce the corresponding image as reference information and utilize the attention mechanism, which allows the model to focus on the most relevant parts of inputs for outputs [24], to integrate the information among text and images at different scales, i.e. lexical-level, phrase-level, and sentence-level. Thus, sentence semantic can be enhanced with the help of reference information and evaluated with the same standard, which is in favor of tackling NLI task. Extensive evaluations on a large-scale NLI corpus and a real-world NLI alike corpus demonstrate the effectiveness of *IEMLRN* over state-of-the-art sentence encoding-based baselines.

The remainder of this paper is organized as follows. In Section II, we introduce the related work. Then the structure and technical details of our proposed approach are given in Section III. In Section IV, experiments on different test sets are presented. Finally, we conclude our work in Section V.

## II. Related Work

In this section, we will introduce the related work, which can be classified into three parts: methods about NLI, methods about image captioning, as well as works about NLI data.

### A. Natural Language Inference Methods

Due to data limitation, early works on NLI have been performed on small datasets with conventional methods [1]. Turney et al. [6] 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 representation 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 [25]. Zhang et al. [26] introduced the neural network into lexical-level inference and proposed a method called *CENN* to represent words semantic with different context and integrated these representations with the consideration of inference relations.

With large annotated corpora for NLI, i.e. SNLI [13], MultiNLI [2], a variety of methods have been developed to represent and evaluate sentence semantic for NLI. These models could be classified into two frameworks: sentence representation framework and words matching framework.

The sentence representation framework focused on semantic representation of each sentence and modeled their semantic similarity. For example, Bowman et al. [13] encoded the premise and hypothesis with different LSTMs. Many related works followed this framework, using different neural networks as encoders. They also proposed a Stack-augmented Parser-Interpreter Neural Network (SPINN), which combined parsing and interpretation within a single tree-sequence hybrid model [27]. Munkhdalai et al. [28] proposed Neural Tree Indexers (NTI) to provide a middle ground between sequential RNNs and syntactic tree-based recursive models. In addition to network structures and sentence structures, inner information also attracted researchers' attention. Mou et al. [9] proposed a tree-based convolutional neural network (TBCNN) for NLI, which captured the sentence level semantics. Liu et al. [29] proposed a bi-directional LSTM model with inner-attention of a sentence to generate sentence representation, which could help re-weight words according to their importance. Im et al. [30] employed multi-head attention and distance mask, which could grasp as many aspects of sentences as possible, to generate a better sentence representation.

The second framework took into consideration more about words matching. Rocktäschel et al. [31] proposed a word-by-word attention model to capture the attention information among words and sentences. Cheng et al. [32] proposed an LSTM with deep attention fusion model to process text incrementally from left to right. There were still other models developed for NLI, such as decomposable attention model with intra-sentence attention [33], full tree matching NTI-SLSTM-LSTM with global attention [28], Bilateral Multi-Perspective Matching [34], etc.

### B. Image Captioning Methods

It has been observed that the use of the intermediate representation from Convolutional Neural Network (CNN) as an image descriptor significantly boosts subsequent tasks such as object localization, object detection, and fine-grained recognition [22], [35], [36]. What's more, image captioning [18]–[21] has been found benefiting from using the image descriptors from a pre-trained CNN.

Junhua et al. [19] proposed a multimodal Recurrent Neural Network (m-RNN) to model the probability distribution of generating a word given previous words and the image.

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 |

They took the image information into account in each step of generating a new word. Andrej et al. [37] proposed a Multimodal Recurrent Neural Network that used the inferred alignments to learn to generate novel descriptions of image regions. Their model utilized the images and sentences to learn about the inter-modal correspondences between them. Oriol et al. [21] proposed a neural network consisting of a vision CNN followed by a language generating RNN. The initial state of language generating RNN was the image representation from the vision CNN.

### C. Works on NLI Data

With the development of large annotated NLI corpora, i.e. SNLI [13], MultiNLI [2], more and more neural networks have beed proposed to represent and evaluate sentence semantic, as well as tackle the NLI tasks. However, these datasets were created by crowd workers. Specific linguistic phenomena such as negation and vagueness would be highly correlated with certain inference classes [38], making it possible to identify the label by looking only at the hypothesis. Thus, based on the SNLI test set, Gururangan et al. [38] proposed a challenging hard test set, in which the examples that premise-oblivious model classified accurately were removed. They intended to better evaluate NLI models' performance with this test set.

Besides, recent models that intended to tackle the NLI task concerned more about the structures and global semantic of sentences, but less about external lexical knowledge, which led them to failing to capture many simple inferences that require lexical and world knowledge [39]. In order to evaluate NLI models' generalization ability, Glockner et al. [39] proposed a simple but challenging lexical test set. Since this test set is created based on SNLI too, all the models trained on SNLI data could be tested for better evaluation. Table I gives some example from SNLI test, hard test, as well as lexical text.

## III. PROBLEM STATEMENT AND MODEL STRUCTURE

In this section, we first formulate the natural language inference (NLI) problem as a supervised classification and then introduce the structure and technical details of *Image-Enhanced Multi-Level Sentence Representation Net (IEMLRN)* for natural language inference.

### A. Problem Statement

First, we define our task in a formal way. Given a premise sentence $\boldsymbol{s}^p = \{\boldsymbol{w}_1^p, \boldsymbol{w}_2^p, ..., \boldsymbol{w}_{l_p}^p\}$, a hypothesis sentence $\boldsymbol{s}^h = \{\boldsymbol{w}_1^h, \boldsymbol{w}_2^h, ..., \boldsymbol{w}_{l_h}^h\}$ and the corresponding image $\boldsymbol{I}$, our goal is to learn a classifier $\xi$ 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 $i$ th and $j$ th 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, two challenges should be considered:

1) Sentences may have various meanings within different contexts. How to ensure the semantic meanings captured from both the premise and the hypothesis sentences match the context given by the image?
2) The key feature related to the given context (image) can exist in different scales such as a key word, a specific phrase or the whole sentence. How to model the sentences in a multi-scale manner to understand the sentences from lexical view to global view?

To this end, we propose an Image-Enhanced Multi-Level Sentence Representation Net (*IEMLRN*). In the following subsections, we will show our proposed model dealing with these issues.

### B. Technic Solution

The overall architecture is shown in Figure 2. In order to understand sentences at multiple scales, we utilize three networks, i.e., lexical-level network, phrase-level network, as well as sentence-level network, as it is shown in the dashed boxes in Figure 2.

Each network consists of four layers: 1) Embedding Layer: encoding the text inputs with different granularity, i.e. lexical-level, phrase-level, and sentence-level; 2) Image-Enhanced Unit Layer: generating the comprehensive sentence representation with the image reference information; 3) Matching Layer: modeling the inference relation between two sentences and getting the inference relation vector; 4) Classification Layer: classifying the inference relation with different granularity. Next, we will introduce the technical detail of each layer.

*1) **Embedding Layer**:* The image input of *IEMLRN* are the feature representations. We select the pre-trained VGG19 [40] 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 $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_{l_c}\}, \boldsymbol{c}_i \in \mathbb{R}^{d^{'}}$, where $d^{'}$ represents the dimension of each feature.

The text inputs of *IEMLRN* are one-hot representation sequences $\boldsymbol{s}^p = \{\boldsymbol{w}_1^p, \boldsymbol{w}_2^p, ..., \boldsymbol{w}_{l_p}^p\}$ for premise sentence, $\boldsymbol{s}^h = \{\boldsymbol{w}_1^h, \boldsymbol{w}_2^h, ..., \boldsymbol{w}_{l_h}^h\}$ for hypothesis sentence. In order to better represent each word, we utilize the concatenation of pre-trained word embedding [41], character feature [42], and syntactical features [38], [43] to represent each word in sentences. The character feature is obtained by applying a convolutional

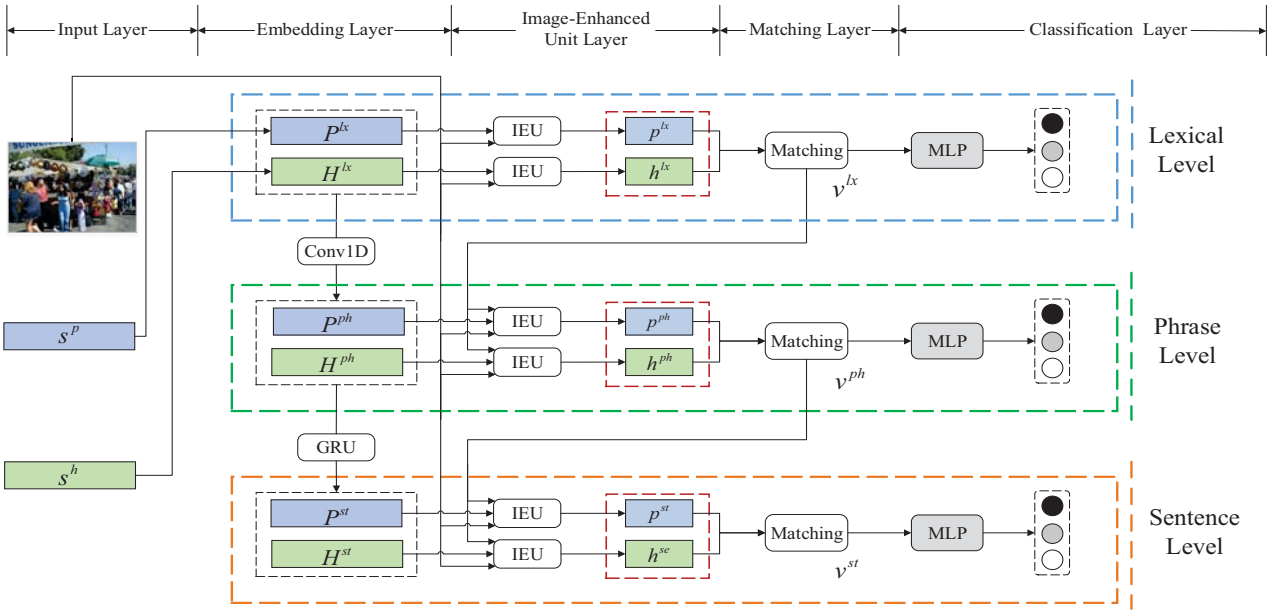


Fig. 2. Architecture of Image-Enhanced Multi-Level sentence Representation Net (*IEMLRN*). (1):The lexical-level embedding $\boldsymbol{P}^{lx}, \boldsymbol{H}^{lx}$ are the concatenation of text features. The phrase-level embedding $\boldsymbol{P}^{ph}, \boldsymbol{H}^{ph}$ are the results of 1-D convolution with lexical-level embedding. The sentence-level embedding $\boldsymbol{P}^{st}, \boldsymbol{H}^{st}$ are the outputs of GRU based on phrase-level embedding. (2): $\boldsymbol{C}$ are the image features from pre-trained VGG19. (3): Each output in the classification layer is the probability distribution on three categories $P(y|\boldsymbol{s}^p, \boldsymbol{s}^h, \boldsymbol{I})$.

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 pre-trained 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 has been proved useful for sentence semantic understanding [38], [43]. Then we get the extravagant representations $\{\boldsymbol{p}_i^{lx}|i = 1, 2, ..., l_p\}$ and $\{\boldsymbol{h}_j^{lx}|j = 1, 2, ..., l_h\}$ 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 subsection III-C.

However, these text representations 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 dependencies among sentences. Thus, multi-level embedding methods are employed to encode the necessary information from different granularity.

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 those $\boldsymbol{p}_i^{lx}$ 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, 1-D convolutions with different filter sizes (unigram, bigram, and trigram) [4] 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 follow:

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

Furthermore, 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 [44] 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:

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

Therefore, we have access to embedding layer for three levels of sentence representations. We have to note that each representation at different levels will then be passed to the Image-Enhanced Unit (IEU) Layer to make a deeper fusion with image reference information.

*2) **Image-Enhanced Unit Layer**:* As mentioned before, reference information is essential for sentence semantic understanding and helpful for evaluating two sentences with the same standard. However, how to make full 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 model the dependency regardless of sequence length [42]. 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 [15], [45]. 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.

Figure 3 shows the structure of Image-Enhanced Unit (IEU). As shown in the figure, the inputs are one embedding sequence $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_{l_s}\}$, one image feature sequence $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_{l_c}\}$, as well as the inference relation vector $\boldsymbol{v}$ which we will introduce next. Please note that the embedding sequence $\boldsymbol{P}$ can be the premise vectors $\boldsymbol{P}^{lx}$, $\boldsymbol{P}^{ph}$ $\boldsymbol{P}^{st}$ or

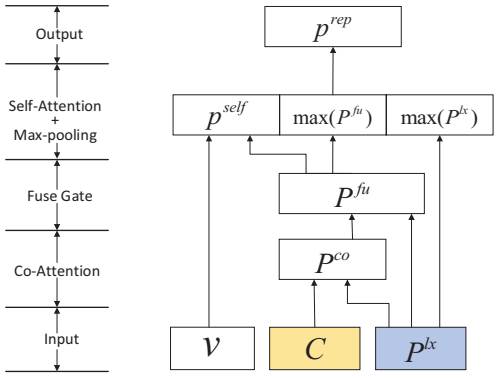


Fig. 3. Architecture of Image-Enhanced Unit.

the hypothesis vectors $\boldsymbol{H}^{lx}$, $\boldsymbol{H}^{ph}$ $\boldsymbol{H}^{st}$ 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 [42] to model the relevance of each word in the premise sentence and the image features, which can be formulated as follows:

$$\begin{aligned} \alpha_{ij} &= \tanh(\boldsymbol{p}_i^{lx}\boldsymbol{W}^{co}\boldsymbol{c}_j) \in \mathbb{R}, \\ \boldsymbol{p}_i^{co} &= \sum_{j=1}^{l} \frac{\exp(\alpha_{ij})}{\sum_{k=1}^{l}\exp(\alpha_{kj})}\boldsymbol{c}_j, \quad i = 1, 2, ..., l_p, \end{aligned} \tag{3}$$

where $\boldsymbol{p}_i^{co}$ is actually a weight summation of the image context $\boldsymbol{c}_j$ for $i$-th word in the premise. $\boldsymbol{W}^{co} \in \mathbb{R}^{d*d'}$ is the trainable parameters. According to *Similarity differences hypothesis* [6], 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 for semantic understanding of each word in sentences.

After getting the representation $\{\boldsymbol{p}_i^{co}\}$ from reference information, it's natural to consider integrating this representation and the original representation $\{\boldsymbol{p}_i^{lx}\}$. Inspired by GRU architecture, we introduce the fuse gate to integrate two type of representations [42], which can be formalized as follows:

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

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 $\sigma$ are activation functions, while $\odot$ is element-wise product. By utilizing fusion gate operation, we can integrate textual information as well as reference information. Thus, the semantic of each word is 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 words' dependency and interaction among the sentence. In order to capture the dependency between words and significant properties in each sentence, we perform a self-attention, a max-pooling on each fusion result, as well as max-pooling on the text input sequence. Then, we concatenate them together:

$$\begin{aligned} \beta_i &= \boldsymbol{w}^T\sigma(\boldsymbol{W}_\beta\boldsymbol{p}_i^{fu} + \boldsymbol{U}_\beta\boldsymbol{v} + \boldsymbol{b}_\beta), \\ \boldsymbol{p}^{self} &= \sum_{i=1}^{p} \frac{\exp(\beta_i)}{\sum_{k=1}^{p}\exp(\beta_k)}\boldsymbol{p}^{fu}, \quad i = 1, 2, ..., l_p, \\ \boldsymbol{p}^{rep} &= [\boldsymbol{p}^{self}; \max_i^p(\boldsymbol{p}^{fu}); \max_i^p(\boldsymbol{p}^{lx})]. \end{aligned} \tag{5}$$

Here, $\boldsymbol{v}$ in Eq. (5) is 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, its $\boldsymbol{v}$ is 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 are aimed at the same sentence, we aim to inform the current level of the classification reason of the previous level. By utilizing this operation, the model can grasp the most relevant parts for inference relation precisely and make the correct decision. What's more, 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 Figure 2, the sentence vector $\boldsymbol{p}^{rep}$ for the premise and $\boldsymbol{h}^{rep}$ for hypothesis represent the sentence semantics in a comprehensive way and guarantee the models' ability of sentence understanding and inference relation classification.

*3) **Matching Layer**:* In order to better evaluate the overall inference relation between two sentences. we employ matching layer to integrate the information among the premise representation $\boldsymbol{p}^{rep}$ and hypothesis representation $\boldsymbol{h}^{rep}$. To be specific, we leverage heuristic matching methods to modify these representations, which can be formulated as follows:

$$\boldsymbol{v} = \mathrm{relu}([\boldsymbol{p}^{rep}; \boldsymbol{h}^{rep}; \boldsymbol{h}^{rep} - \boldsymbol{p}^{rep}; \boldsymbol{h}^{rep} \odot \boldsymbol{p}^{rep}]), \tag{6}$$

where $[.;.]$ represents the concatenation operation, $\odot$ means element-wise product and $\mathrm{relu}$ is the non-linear activation function. $\boldsymbol{v}$ is the inference relation vector of two sentences. To be specific, concatenation can retain all the information [26]. The element-wise product is a certain measure of "similarity" of premise and hypothesis [9]. Their differences can capture the degree of distributional inclusion on each dimension [46]. The output $\boldsymbol{v}$ will be used as the input of the classification layer. Besides, as mentioned in section III-B2, 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.

*4) **Classification Layer**:* After getting the inference relation vector $\boldsymbol{v}$, we utilize a multi-layer perceptron (MLP) and one $\mathrm{softmax}$ output layer to classify the inference relation of two sentences. The output of this layer is the probability

distribution of the inference relation between these sentence pairs the reference information. The formulation is as follows:

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

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 Figure 2, we utilize $P^{lx}, P^{ph}, P^{st}$ to represent the outputs of lexical-level network, phrase-level network, as well as sentence-level network. The final classification result we use is the output of sentence-level network.

### C. Model Learning

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

*1) **Loss Function**:* Since it's 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:

$$L = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \log P(y_i|\boldsymbol{s}_i^p, \boldsymbol{s}_i^h, \boldsymbol{I}_i) \tag{8}$$

$\boldsymbol{y}_i$ is the one-hot representation for the true class of $i$-th example, and $P(y_i|\boldsymbol{s}_i^p, \boldsymbol{s}_i^h, \boldsymbol{I}_i)$ is the probability distribution over the classes that *IEMLRN* outputs. As mentioned in section III-B4, each network in our model has an output. We intend that each network in our model should make the right classification. Therefore, we apply cross-entropy function to each-level output. Considering the model complexity, we also add the l2-norm of all parameters $\boldsymbol{\theta} = \{\boldsymbol{W}^{co}, \boldsymbol{W}_z, \boldsymbol{b}_z, \boldsymbol{W}_r, \boldsymbol{b}_r, \boldsymbol{W}_f, \boldsymbol{b}_f, \boldsymbol{W}_\beta, \boldsymbol{U}_\beta, \boldsymbol{b}_\beta\}$ in Image-Enhanced Unit Layer to the entire loss function. Then we get the loss function for the whole model as follows:

$$L = L^{lx} + L^{ph} + L^{st} + \epsilon \left\|\boldsymbol{\theta}\right\|_2 \tag{9}$$

*2) **Model Initialization**:* We set the word embedding dimension as 300, character-level embedding level as 100, the dropout as 0.6, and $\epsilon$ as 0.01. The word embedding we use are obtained from a pre-trained word vectors (840B GloVe) [41]. The hidden state size of GRU is 512. To initialize the model, we randomly set the all weights such as $\boldsymbol{W}$ following the uniform distribution in the range between $-\sqrt{6/(nin + nout)}$ and $\sqrt{6/(nin + nout)}$ as suggested by [47]. All biases such as $\boldsymbol{b}$ are set as zeros. We use Adam optimizer with learning rate $10^{-4}$. During implementation, we utilize *Photinia*[1] to build our entire model.

## IV. Experiment

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

### A. Data Description

In this subsection we introduce two datasets we evaluate the models on, and we utilize parameter size and accuracy on different test sets to evaluate the performance of all models.

**SNLI.** Standford Natural Language Inference (SNLI) [13] has 570k human annotated sentence pairs. The premise sentences are drawn from the captions of the Flickr30k corpus [48], and the hypothesis sentences are manually composed. Thus each instance has one corresponding image treated as the reference information. The labels we use are "*entailment*", "*neutral*", and "*contradiction*".

We use the same data split as in [13]. In order to reduce the impact of annotate artifacts and better evaluate models' ability of sentence understanding, we also select the challenging hard subset from [38], in which the premise-oblivious model cannot classify accurately, as one of the test set. What's more, the challenging lexical subset from [39], which require lexical and world knowledge, is also selected to evaluate models' generalization ability. Table I gives some examples from different SNLI test sets.

**DanMu.** This dataset comes from the user-generated time-sync comments about the videos[2], which has 12k sentence pairs with corresponding images. Following the idea [49], the premise and the corresponding image are extracted from a short period [50], [51], the hypothesis sentence is a modified variant of one of the comments from either the same period or a random, unrelated one. The labels we use here are "*entailment*" and "*not entailment*". Table III reports some key statistics about these test sets.

### B. Baselines

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

- **LSTM encoders** [13]: encoding the premise and hypothesis with two different LSTMs.
- **CENN** [26]: utilizing different context to generating sentence representation for NLI.
- **BiLSTM with Inner-Attention** [29]: using bidirectional LSTM with inner attention mechanism to generating sentence representation for NLI.
- **Gated-Att BiLSTM** [10]: employing intra-sentence gated-attention component to encodes a sentence to a fixed-length vector for NLI.
- **Distance-based Self-Attention** [30]: utilizing self-attention and distance mask to model the local and global dependency for NLI.

We also select two image captioning models to better verify the performance of *IEMLRN*. Since these works aim to generate the description of the images, we add the premise and hypothesis as inputs to RNN module separately and treat the final state of models as sentence representations. Then, we employ the same matching unit and classification layer as *IEMLRN*, shown in Figure 2 for NLI.

[1]https://github.com/XoriieInpottn/photinia

[2]www.bilibili.com

TABLE II
PERFORMANCE (ACCURACY) OF MODELS ON SNLI AND DANMU DATASETS.

| Model | #Paras | Full test | Hard test | Lexical test | DanMu test |
|---|---|---|---|---|---|
| (1)LSTM encoders [27] | 3.0m | 80.6% | 58.5% | 52.3% | 64.9% |
| (2)BiLSTM with Inner-Attention [29] | 2.8m | 84.5% | 62.7% | 58.6% | 66.3% |
| (3)CENN [26] | ≈700k | 82.1% | 60.4% | 51.9% | 65.2% |
| (4)Gated-Att BiLSTM [10] | 12m | 85.5% | 65.5% | 65.6% | 67.3% |
| (5)Distance-based Self-Attention [30] | 4.7m | 86.3% | 67.4% | 68.5% | 69.7% |
| (6)CENN with image [26] | ≈700k | 83.1% | 61.7% | 66.8% | 66.6% |
| (7)NIC [21] | - | 84.7% | 63.6% | 67.1% | 67.9% |
| (8)m-RNN [29] | - | 85.1% | 64.9% | 69.4% | 68.1% |
| (9)*IEMLRN* | 3.9m | **87.5**% | **75.4**% | **78.1**% | **79.5**% |
| (10)*IEMLRN*(*ensemble*) | - | 88.3% | 78.7% | 80.5% | 82.4% |

TABLE III
STATISTICAL INFORMATION OF EACH TEST SET.

| Test set | Data Size | | | Average Token Count | |
|---|---|---|---|---|---|
| | E | C | N | premise | hypothesis |
| SNLI Full Test | 3368 | 3237 | 3219 | 13.91 | 7.48 |
| Hard Test | 1058 | 1135 | 1068 | 13.81 | 7.71 |
| Lexical Test | 982 | 7164 | 47 | 11.42 | 11.60 |
| DanMu Test | 2127 | - | 3906 | 10.07 | 9.26 |

- **NIC** [21]: a neural network consisting of a vision CNN followed by a language RNN.
- **m-RNN** [29]: utilizing a deep RNN for sentences and a deep CNN for images to model the probability distribution of words.

### C. Experimental Results

We evaluate the performance of all models from the following aspects: A) The parameter size (#Para) of models; B) The accuracy in 1) SNLI full test set(Full test); 2) SNLI Hard test (Hard test); 3) SNLI lexical test (Lexical test); D) DanMu test set (DanMu test).

*1) **Overall Performance**:* The overall results are summarized in Table II. We can conclude that *IEMLRN* achieves state-of-the-art performance with respect to parameter size and accuracy on all test sets. To be specific, *IEMLRN* introduces the corresponding image as reference information. Thus, the inference relation between the premise and hypothesis sentences can be evaluated with the same standard. With attention mechanism, reference information can provide necessary help for sentence semantic. As shown in Figure 1, the image can be helpful for distinguishing the exact weather in premise sentence. What's more, our model integrates the reference information and evaluates the inference relation between two sentences with different granularity, which means *IEMLRN* can not only understand the sentence semantic with lexical knowledge, but also model local and global semantic interactions between sentences. Thus, *IEMLRN* can understand sentence from a more comprehensive perspective and achieve state-of-the-art performance compared with these baselines.

*2) **Experiments on SNLI full test**:* LSTM encoder [27] utilizes different LSTMs to encode sentences and leads many related works to employ different network structure as encoders, such as BiLSTM with Inner-Attention [29], CENN [26]. However, these models encode each sentence separately. The interactions between two sentences, which are essential for NLI, have not been utilized effectively. The results also demonstrate that just separated textual information is insufficient for NLI. Gated-Att BiLSTM [10] and Distance-based Self-Attention [30] both utilize attention mechanism to evaluate the important parts of sentences, which leads a better performance. The former utilizes the gate information in LSTM to represent the importance of words. The dependencies among sentences are evaluated effectively. However, if words in the sentences have high overlap, it might have a bad influence on its performance. The latter uses multi-head attention to consider as many aspects of sentences as possible. Thus it can understand the sentence semantic and evaluate the inference relation effectively. However, They take into consideration only the text information, which is insufficient for solving the issues, such as ambiguity and fuzziness, that sentence semantics suffer. *IEMLRN* takes advantage of image reference information to understand the sentence semantic more precisely and avoid the issues that sentence semantic suffers. Thus, our proposed model significantly outperforms other NLI baselines.

*3) **Experiments on SNLI hard test set**:* Gururangan et al. [38] suspects that annotation artifacts inflate model performance. Thus, they propose a challenging hard subset of SNLI to better evaluate the models' ability on sentence semantic understanding. Since the examples that premise-oblivious model classified accurately are removed, this test set can focus the model on sentence semantic rather than the annotate artifacts and better evaluate the models' performance.

From the results in Table II, we could conclude that *IEMLRN* outperforms all the baselines by a large margin, e.g. Distance-based Self-Attention model (+8.0%), Gated-Att BiLSTM model (+9.9%) and CENN (+15.0%), which indicates that our proposed model has better adaptability. Since *IEMLRN* introduce the corresponding image as reference information into inference processing and evaluate the sentence semantics from different granularity, it can still capture the sentence semantic even if the sentences became obscure.

*4) **Experiments on SNLI lexical test set**:* Recently, many end-to-end neural models pay more attention to network structures and make few efforts to incorporate external lexical knowledge, which is simple but important for sentence semantics, into inference processing. Whether learning from large-scale training data can help the model grasp the explicit lexical knowledge still attracts researchers' attention. Based on

TABLE IV
ABLATION PERFORMANCE (ACCURACY) OF MODELS ON SNLI AND DANMU DATASETS.

| Model | #Paras | Full test | Hard test | Lexical test | DanMu test |
|---|---|---|---|---|---|
| (1)IEMLRN-lexical feature | 1.6m | 34.3% | 42.1% | 67.5% | 51.7% |
| (2)IEMLRN-phrase feature | 1.6m | 52.7% | 45.2% | 66.2% | 57.5% |
| (3)IEMLRN-sentence feature | 1.8m | 76.2% | 57.9% | 65.0% | 62.3% |
| (4)IEMLRN-lexical-gram feature | 2.7m | 65.5% | 64.5% | 69.4% | 63.3% |
| (5)IEMLRN-lexical-sentence feature | 2.96m | 82.9% | 66.8% | 74.6% | 70.2% |
| (6)IEMLRN-gram-sentence feature | 2.96m | 83.2% | 65.7% | 73.2% | 71.5% |
| (7)*IEMLRN* | 3.9m | **87.5**% | **75.4**% | **78.1**% | **79.5**% |
| (8)*IEMLRN*(*ensemble*) | - | 88.3% | 78.7% | 80.5% | 82.4% |

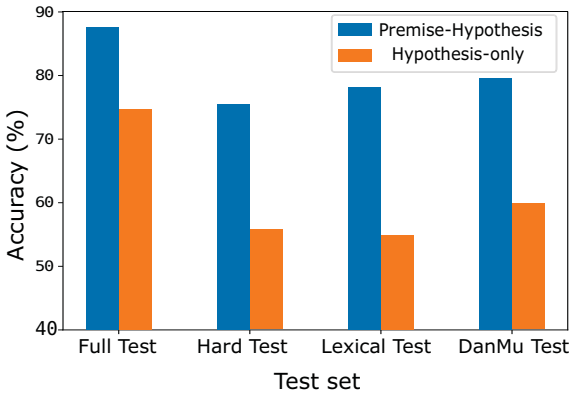


Fig. 4. Comparison between premise-hypothesis and hypothesis-only.

SNLI dataset, Glockner et al. [39] creates a new challenging SNLI lexical text set to evaluate models' ability to make inferences that require simple lexical knowledge. The lexical test examples in Table I show us that the premise and hypothesis sentences have high overlap on words, even the different words between these two sentences may be in the same category. Thus, this test set requires NLI models to pay close attention to the specific words and their lexical semantics in sentences. Due to the phenomenon that is not included in training data and the high similarity between two sentences, this test set make great difficulties for these NLI models, though it is simple for us human. As shown in Table II, most of the models' performances are worse than their own performances on the former two test set.

However, *IEMLRN* achieves a 78.1% accuracy, which is 2.7% higher than its own performance on the hard test. This indicates our proposed model has better generalization ability and grasps the lexical knowledge indeed. In order to better integrate the image reference information and text information, *IEMLRN* considers the sentences semantic with different granularity. Thus, it has the ability to concern the difference not only between the whole sentences, but also between the lexical-level or phrase-level semantic. What's more, the image reference information can provide necessary support for identifying this lexical knowledge in sentences. Therefore, our proposed model can outperform these state-of-the-art sentence-encoding baselines.

*5)* ***Experiments on DanMu test set:*** This NLI alike dataset is collected from user-generated time-sync comments about videos, which was highly diverse in various aspects (length, complexity, expression, etc), posing linguistic challenges for NLI task. They might present the same meaning with different forms of expression due to different audiences. *IEMLRN* takes into consideration the corresponding image, which helps to grasp the true meaning of sentences and evaluate the sentence semantics with the same standard. Moreover, multi-level representations and understanding help the model effectively utilize the image reference information. Thus, *IEMLRN* achieves the best performance. What's more, we can find that the model with images will perform better than most of those without images, which also proves that reference information is important for sentence semantics and NLI.

### *D. Ablation Performance*

We conduct an ablation study on our model to examine the effectiveness of each component. The results are shown in Table IV. Experiments (1)-(3) are single granularity. When considering global semantic, i.e. full test, hard test, and DanMu test, we can conclude that sequence information is necessary for sentence semantics. Without sequence information, the performance of lexical feature and phrase feature are very bad. When local semantic matters more, i.e. lexical test, we find that lexical feature and phrase feature achieve better performance than sentence feature, proving different granularity information is all useful for sentence semantic. Experiments (4)-(6) consider two of three granularity. We can draw the same conclusion as experiments (1)-(3). What's more, these results are a little better than the previous ones, indicating that considering the semantics of sentences from different granularity is very important and necessary for semantic understanding. In other words, Lexical-level information considers more about the local information, while sentence-level information concerns more about the global information and sequence information. They all should be considered for better sentence understanding and inference relation classification.

### *E. Analysis on the function of images*

We introduce the images as reference information into NLI. In SNLI data, the premise sentences are drawn from the captions of these images. In DanMu data, the premise sentences are comments about these images. Thus, it's urgent to distinguish the difference between our work and those about image and sentence retrieval. If the images have the ability to replace the premise sentence, we can just evaluate the relation between images and hypothesis sentences like the latter ones. Therefore, we remove all the premise data in *IEMLRN* and evaluate its performance on the test sets. The results are shown in Figure 4.

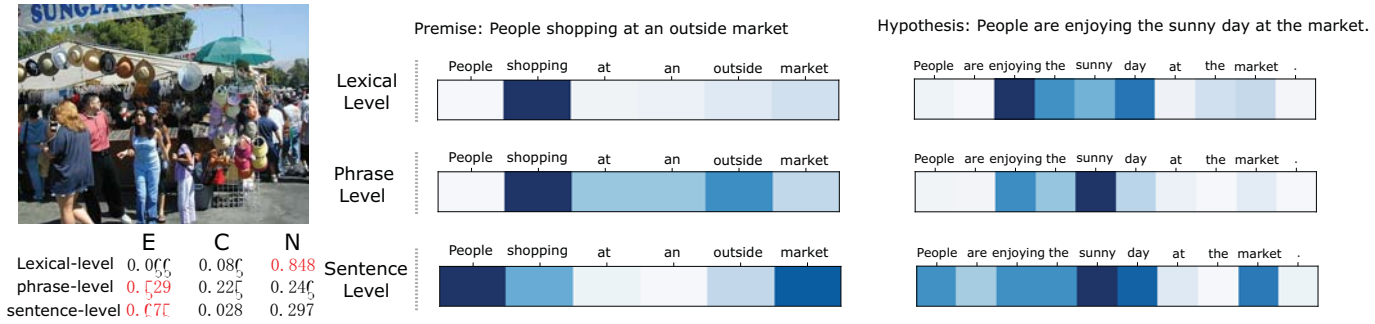


Fig. 5. Visualization of self-attention with different granularity.

From the results, we conclude that there is a big gap between the performance of complete data and the hypothesis-only data. What's more, the performances on these challenging test sets, such as hard test and Lexical test, become a lot worse. This indicates that the images cannot replace the premise sentences. Even though the premise sentences describe the information in corresponding images, there is still a big difference between them. The images can be treated as reference information to assist the sentence semantic understanding and inference relation classification, but they cannot be treated as the replacement of the premise sentences. In other words, our work still focuses on the inference relation between sentences and has a great difference with image and sentence retrieval alike work. *IEMLRN* is a novel NLI model indeed.

### F. Case Study

Here, we visualize the self-attention with different granularity to validate the introduced reference information and the multi-level representations. Figure 5 shows the attention distribution and classification probability distribution of each level over the example shown in Figure 1.

With the help of image reference information, *IEMLRN* focuses on the most relevant parts of the premise and hypothesis sentences. i.e. "*shopping, market*" in premise and "*enjoying the sunny day, the market*" in hypothesis sentences at lexical-level. Since *IEMLRN* only know the lexical information at this level, it selects the words that are critical to sentence semantics under the same standard according to the reference information. From the classification result, we find that the model considers the inference relation is "*neutral*".

When it comes to phrase-level self-attention, *IEMLRN* is able to consider not only the sentence semantic more comprehensive, but also the lexical knowledge from the previous level. Therefore, the classification result turns to "*entailment*". What's more, we can figure out from the attention visualization that the model started to concern the "*outside market*" in premise sentence and "*sunny day*" in hypothesis sentence, which indicates that the model evaluated the inference relation between these two phrases. As mentioned in Section I, there is a high correlation between the "*sunny day*" and "*outside market*", which is consistent with our model.

The following is Sentence-level self-attention. In this level, *IEMLRN* takes into consideration the global semantic and dependencies among sentences. The attention distribution at this level also indicates the same phenomenon. We find that *IEMLRN* pays attention to not only the weather information previous level found, i.e. "*outside market*" in premise sentence and "*sunny day*" in hypothesis sentence, but also people's activation, i.e. "*People shopping*" in premise sentence and "*People enjoying market*" in hypothesis sentence. In other words, *IEMLRN* evaluates sentence semantics and relations from lexical knowledge to global semantic, which in favor of tackling the NLI task. The classification probability also indicates that *IEMLRN* become more confident to classify this instance to "*entailment*". With respect to the information from different levels, our proposed model makes a confident and solid decision.

## V. Conclusion and Future Work

In this paper, we proposed an Image-Enhanced Multi-Level Sentence Representation Net (*IEMLRN*) for natural language inference, a novel architecture that allowed the model to utilize the image reference information and understand sentence semantic from lexical knowledge to global semantics. To be specific, we introduced the corresponding image of two sentences as the reference information, which could be helpful for sentence semantic understanding and evaluate the inference relation with the same standard. In order to better integrate the image reference information and text information, we proposed a multi-level networks, i.e. lexical-level, phrase-level, as well as sentence-level, to utilize the image for enhancing the sentence understanding and representation. Thus, sentence semantic and inference relations could be evaluated from a more comprehensive perspective. Experimental results on different SNLI test sets and a real-world NLI alike corpus demonstrated that *IEMLRN* had the ability to understand sentence semantic, generate sentence representation, and evaluate the inference relation between sentences at different scales. In the future, we will consider more different reference information and more efficient processing methods for more precise sentence semantic understanding and representation.

## VI. Acknowledgements

This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. U1605251, 61727809, 61672483 and 61602147), the Science Foundation of Ministry of Education of China & China Mobile (No. MCM20170507), and CCF-Tencent Open Fund.

## REFERENCES


[1] Bill MacCartney. *Natural language inference*. Stanford University, 2009.

[2] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. pages 1112–1122, 2018.

[3] Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. Combining retrieval, statistics, and inference to answer elementary science questions. In *AAAI*, 2016.

[4] Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *Proc. CVPR*, 2017.

[5] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784, 2016.

[6] Peter D Turney and Saif M Mohammad. Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering*, 21(3):437–476, 2015.

[7] Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. Entailment above the word level in distributional semantics. In *EACL*, pages 23–32, 2012.

[8] Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389, 2010.

[9] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. Natural language inference by tree-based convolution and heuristic matching. In *ACL*, volume 2, pages 130–136, 2016.

[10] Qian Chen, Xiao-Dan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Recurrent neural network-based sentence encoder with gated attention for natural language inference. 2017.

[11] Tsendsuren Munkhdalai and Hong Yu. Neural tree indexers for text understanding. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, page 11, 2017.

[12] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

[13] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

[14] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. 2016.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[16] Hao Chen, Keli Xiao, Jinwen Sun, and Song Wu. A double-layer neural network framework for high-frequency forecasting. *TMIS*, 7(4):11, 2017.

[17] Qi Liu, Zai Huang, Zhenya Huang, Chuanren Liu, Enhong Chen, Yu Su, and Guoping Hu. Finding similar exercises in online education systems. In *SIGKDD*, pages 1821–1830. ACM, 2018.

[18] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. *IEEE CVPR*, pages 4594–4602, 2016.

[19] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632, 2014.

[20] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. *IEEE CVPR*, pages 1473–1482, 2015.

[21] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *IEEE CVPR*, pages 3156–3164, 2015.

[22] Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B. Gouza. Fake news detection with deep diffusive network model. *CoRR*, abs/1805.08751, 2018.

[23] Limeng Cui, Zhensong Chen, Jiawei Zhang, Lifang He, Yong Shi, and Philip S Yu. Multi-view collective tensor decomposition for cross-modal hashing. In *ICMR*, pages 73–81. ACM, 2018.

[24] Kyunghyun Cho, Aaron C. Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Trans. Multimedia*, 17:1875–1886, 2015.

[25] Charles Kay Ogden. *Basic English: A general introduction with rules and grammar*, volume 29. K. Paul, Trench, Trubner, 1944.

[26] Kun Zhang, Enhong Chen, Qi Liu, Chuanren Liu, and Guangyi Lv. A context-enriched neural network method for recognizing lexical entailment. In *AAAI*, pages 3127–3134, 2017.

[27] Samuel R Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D Manning, and Christopher Potts. A fast unified model for parsing and sentence understanding. In *ACL*, 2016.

[28] Tsendsuren Munkhdalai and Hong Yu. Neural tree indexers for text understanding. *CoRR*, abs/1607.04492, 2016.

[29] Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. Learning natural language inference using bidirectional lstm model and inner-attention. *CoRR*, abs/1605.09090, 2016.

[30] Jinbae Im and Sungzoon Cho. Distance-based self-attention network for natural language inference. *CoRR*, abs/1712.02047, 2017.

[31] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664, 2015.

[32] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *EMNLP*, 2016.

[33] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *EMNLP*, 2016.

[34] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. *CoRR*, abs/1702.03814, 2017.

[35] Pierre Sermanet, David Eigen, Xiang Zhang, and Michaël Mathieu and Rob Fergus and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

[36] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. *2014 IEEE CVPR Workshops*, pages 512–519, 2014.

[37] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE CVPR*, pages 3128–3137, 2015.

[38] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. pages 107–112, 2018.

[39] Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking nli systems with sentences that require simple lexical inferences. In *ACL*, Melbourne, Australia, July 2018.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[41] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[42] Yichen Gong, Heng Luo, and Jian Zhang. Natural language inference over interaction space. *CoRR*, abs/1709.04348, 2017.

[43] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *ACL*, 2017.

[44] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[45] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. *CoRR*, abs/1709.04696, 2017.

[46] Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. Learning to distinguish hypernyms and co-hyponyms. In *COLING*, pages 2249–2259, 2014.

[47] Genevieve B Orr and Klaus-Robert Müller. *Neural networks: tricks of the trade*. Springer, 2003.

[48] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

[49] Alice Lai, Yonatan Bisk, and Julia Hockenmaier. Natural language inference from multiple premises. *IJCNLP*, 2017.

[50] Guangyi Lv, Tong Xu, Enhong Chen, Qi Liu, and Yi Zheng. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding. In *AAAI*, pages 3000–3006, 2016.

[51] Ming He, Yong Ge, Enhong Chen, Qi Liu, and Xuesong Wang. Exploring the emerging type of comment for online videos: Danmu. *ACM Transactions on the Web (TWEB)*, 12(1):1, 2018.