# Context-Aware Dual-Attention Network for Natural Language Inference

Kun Zhang[1], Guangyi Lv[1], Enhong Chen[1(✉)], Le Wu[2], Qi Liu[1], and C. L. Philip Chen[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, Hefei, China
{zhkun,gylv}@mail.ustc.edu.cn, {cheneh,qiliuql}@ustc.edu.cn
[2] Hefei University of Technology, Hefei, China
lewu@hfut.edu.cn
[3] University of Macau, Macau, China
philip.chen@ieee.org

**Abstract.** Natural Language Inference (NLI) is a fundamental task in natural language understanding. In spite of the importance of existing research on NLI, the problem of how to exploit the contexts of sentences for more precisely capturing the inference relations (i.e. by addressing the issues such as polysemy and ambiguity) is still much open. In this paper, we introduce the corresponding image into inference process. Along this line, we design a novel *Context-Aware Dual-Attention Network (CADAN)* for tackling NLI task. To be specific, we first utilize the corresponding images as the *Image Attention* to construct an enriched representation for sentences. Then, we use the enriched representation as the *Sentence Attention* to analyze the inference relations from detailed perspectives. Finally, a sentence matching method is designed to determine the inference relation in sentence pairs. Experimental results on large-scale NLI corpora and real-world NLI alike corpus demonstrate the superior effectiveness of our *CADAN* model.

## 1 Introduction

Natural Language Inference (NLI), also named as Recognizing Textual Entailment (RTE), 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 [19], where the semantic of hypothesis can be concluded from the premise:

*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.*


Q. Yang et al. (Eds.): PAKDD 2019, LNAI 11441, pp. 185–198, 2019.
https://doi.org/10.1007/978-3-030-16142-2_15

*p* : People shopping at outside market

*h* : People are enjoying the sunny day at the market.

gold-label: Entailment

**Fig. 1.** Example from SNLI dataset.

Indeed, NLI not only is concerned with the key parts of natural language understanding, i.e. reasoning and inference [4], but also has broad applications, e.g. question answering [27] and automatic summarization [31]. Many research efforts have been conducted in this area. Generally, the main idea of these works can be summarized into two categories: sentence representation and words matching. Sentence representation models focus on extracting semantic representations for sentences by various network structures [3,9,21]. In contrast, words matching models express more concern about the interactions among aligned words between the premise and hypothesis, such as word-by-word matching model [34] and decomposable attention model [25].

To the best of our knowledge, most of existing research assumed that the hypothesis inference is independent of any context. The contexts (e.g. the corresponding images), however, are actually critical for natural language understanding [1]. Figure 1 gives an example. Both the premise and hypothesis sentences describe that people are shopping at the market. Without the image as context, we might conclude the inference relation is neutral since the weather in premise is unclear. However, when we know the context, it's easy to find out the relation is entailment, which indicates the importance of context. Non-literal contexts, like images, can be useful to clarify these issues such as polysemy, ambiguity, as well as fuzziness of words and sentences [39]. Therefore, it's urgent to take into consideration the image contexts for NLI.

In fact, researchers have converged that images convey important information about the associated sentences [14,18]. Much progress has been made on the image and sentence retrieval [13], image captioning [24], and visual question answer [28], e.g. m-RNN model [20] and NIC model [33]. However, these works focused more on the alignments between images and sentences rather than the interactions between sentences, which made it unsuitable for applying them to the conditional NLI task directly.

Inspired by these works, we introduce the corresponding image of the sentence pair as the context into inference process. The key challenge along this line is how to incorporate images into the inference processing effectively. Thus, in this paper, we propose a novel *Context-Aware Dual-Attention Network (CADAN)* to tackle NLI task. To be specific, we propose *Image Attention* layer to utilize the correlated image to enhance the sentence representations. The enhanced sentence representations are further sent to *Sentence Attention* layer to analyze the inference relations from detailed perspectives. With the help of this dual-attention,

*CADAN* can better evaluate sentence semantic and achieve better performance on NLI task. Finally, the extensive evaluations on the large-scale NLI corpus and real-world NLI alike corpus demonstrate the superior effectiveness of *CADAN*.

## 2 Related Work

In this section, we introduce the related works, which can be classified into two parts: methods about NLI and methods about image captioning.

**Natural Language Inference Methods.** With the help of large annotated datasets, such as Stanford Natural Language Inference (SNLI) [2] and Multi-Genre NLI [36], a variety of methods have been developed for NLI. These models can be classified into two frameworks: sentence representation framework and words matching framework.

The representation framework focused on the sentence representation and interaction. Bowman et al. [2] encoded the premise and hypothesis with different LSTMs. Munkhdalai et al. [22] proposed a memory augmented method, which understood the sentence through *read*, *compose* and *write* operation. In addition to network and sentence structures, inner information of sentences also attracted researchers' interests, such as TBCNN [21], bi-directional LSTM with inner-attention [16].

The second framework concentrated more on words matching. Rocktäschel et al. [29] proposed a word-by-word attention model to capture the attention information among words and sentences. Cheng et al. [5] proposed an LSTM with deep attention fusion model to process text incrementally from left to right. However, most of them assumed that the hypothesis inference was independent of any context, which is actually critical for natural language understanding and should be highly considered.

**Image Captioning Methods.** It has been observed that using the intermediate representation from Convolutional Neural Network (CNN) as an image descriptor significantly boosts subsequent tasks such as object detection, fine-grained recognition [6]. Moreover, researchers have found that using image descriptors from a pre-trained CNN benefited the image captioning [33]. For example, Karpathy et al. [10] proposed an alignment model to learn about the inter-model correspondences between images and texts. Then they utilize the alignments to learn to generate novel descriptions of images.

## 3 Problem Statement and Model Structure

In this section, we formulate the conditional NLI task as a supervised conditional classification problem and introduce the structure and technical details of the *Context-Aware Dual-Attention Network (CADAN)* for the task.

### 3.1 Problem Statement

The inputs of this problem are two sentences $\boldsymbol{s_a} = \{\boldsymbol{s}_1^a, \boldsymbol{s}_2^a, \ldots, \boldsymbol{s}_l^a\}$, $\boldsymbol{s_b} = \{\boldsymbol{s}_1^b, \boldsymbol{s}_2^b, \ldots, \boldsymbol{s}_l^b\}$, as well as one corresponding image $\boldsymbol{c}$ as the given context, where $\boldsymbol{s_a}$ and $\boldsymbol{s_b}$ denote the premise and hypothesis sentence. $l$ represents the length of sentences. Note that $\boldsymbol{s}_i^a$ or $\boldsymbol{s}_i^b$ here denotes the one-hot representation of the $i$th word in the premise or hypothesis sentence. $\boldsymbol{c}$ is the feature representation of the image. The goal is to predict a label $y$ that indicates the inference relation between the premise $\boldsymbol{a}$ and the hypothesis $\boldsymbol{b}$.

Our task in this paper is to learn an accurate classification model, to predict $y$ given a sentence pair with the associated image $(\boldsymbol{s_a}, \boldsymbol{s_b}, \boldsymbol{c})$. To this end, we propose the *Context-Aware Dual-Attention Network (CADAN)* to tackle this issue.

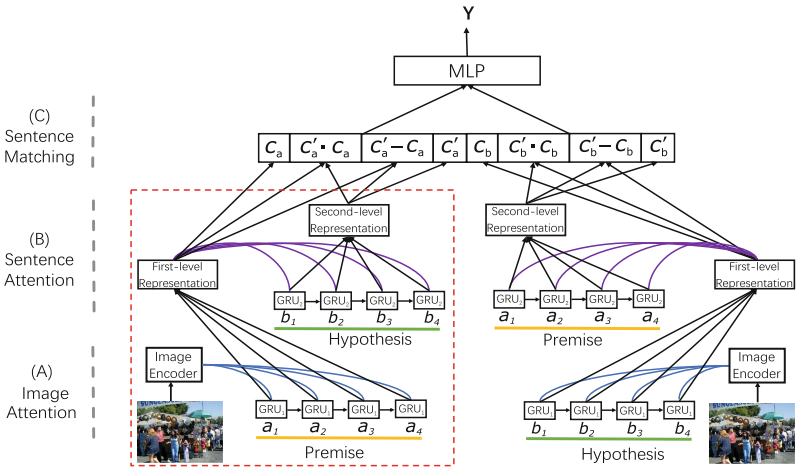


**Fig. 2.** Architecture of the *Context-Aware Dual-Attention Network (CADAN)*.

### 3.2 Context-Aware Dual-Attention Network

Our model can be divided into two parts; (1) The preprocessing part: generating the feature representations of sentences and images. (2) The inference part: utilizing the *Context-Aware Dual-Attention Network (CADAN)* to understand the sentences semantics and classify the inference relations between premise and hypothesis.

**The Preprocessing Part.** Since the inputs of the task are sentence pairs and corresponding images, we utilize different models to represent these different types of data.

For sentences, we utilize the concatenation of pre-trained word embedding (840B Glove) [26] and character feature for English words. The character feature

is obtained by applying a convolutional neural network and a max pooling to the learned character embeddings. For Chinese words, we utilize AutoEncoder [12] to perform the representations of words in sentences. Thus, we get the word embedding $\boldsymbol{E}$ for further use.

For images, we choose the pre-trained VGG19 [32] to process the images. Then we extract the outputs of the last convolution layer of VGG19 and send them to a fully-connected layer to get feature representations of images.

**The Inference Part.** Figure 2 shows the overall framework of *CADAN*, which consists of three components: (1) Image Attention layer; (2) Sentence Attention layer; (3) Sentence Matching layer. In the following part, we take the premise processing as an example to describe technical details of these three components. The same method will be applied to the hypothesis processing.

***(A) Image Attention Layer:*** The images contain the non-literal context of sentences. However, how to utilize the information effectively for sentence semantic is still challenging. Thus, we propose Image Attention layer to integrate them effectively.

In this layer, we first multiply the one-hot representations of the premise $\boldsymbol{s}_a = \{\boldsymbol{s}_1^a, \boldsymbol{s}_2^a, \ldots, \boldsymbol{s}_l^a\}$ and the hypothesis $\boldsymbol{s}_b = \{\boldsymbol{s}_1^b, \boldsymbol{s}_2^b, \ldots, \boldsymbol{s}_l^b\}$ by the word embedding $\boldsymbol{E}$ from the preprocessing part. Then we get the $\{\boldsymbol{a}\}_{j=1}^l$ for premise and $\{\boldsymbol{b}\}_{j=1}^l$ for hypothesis. Next, we leverage Gated Recurrent Units (GRU) [7] to encode these representations. The GRU hidden states below, i.e., $\{\bar{\boldsymbol{a}}\}_{i=1}^l$ and $\{\bar{\boldsymbol{b}}\}_{i=1}^l$ encode each word and sentence context around it:

$$\bar{\boldsymbol{a}}_i = \mathrm{GRU}_1(\{\boldsymbol{a}_{j=1}^i\}), \quad \bar{\boldsymbol{b}}_i = \mathrm{GRU}_1(\{\boldsymbol{b}_{j=1}^i\}), \quad i = 1, 2, \ldots, l. \tag{1}$$

After getting the hidden state of each word, we aim to identify the content of each sentence. Since sentences are both related to the image, the words that are more relevant to the image should get more attention. The attention mechanism can help the model focus on the most relevant part of the input [6,37]. Thus, we utilize VGG19 to get the feature representation $\boldsymbol{c}$ of the corresponding image and send it to the attention cell:

$$\begin{aligned} &\bar{\boldsymbol{A}} = [\bar{\boldsymbol{a}}_1, \bar{\boldsymbol{a}}_2, \ldots, \bar{\boldsymbol{a}}_l], \quad \boldsymbol{M} = \tanh(\boldsymbol{W}\bar{\boldsymbol{A}} + \boldsymbol{U}\boldsymbol{c} \otimes \boldsymbol{e}_l), \ \boldsymbol{W}, \boldsymbol{U} \in \mathbb{R}^{k*k}, \\ &\boldsymbol{\alpha} = \mathrm{softmax}(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{M}), \quad \boldsymbol{c}_a = \bar{\boldsymbol{A}}\boldsymbol{\alpha}^{\mathrm{T}}, \ \boldsymbol{\omega} \in \mathbb{R}^k, \end{aligned} \tag{2}$$

here $\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{\omega}$ are trained parameters. $k$ is the state size of GRU cell in Eq. (1), $\boldsymbol{\alpha}$ is the attention weights vector of hidden states for words, $\boldsymbol{c}_a$ is the first-level representation for premise, and $\boldsymbol{e}_l \in \mathbb{R}^l$ is a row vector of 1. The outer product $\boldsymbol{U}\boldsymbol{c} \otimes \boldsymbol{e}_l$ means repeating $\boldsymbol{U}\boldsymbol{c}$ as many times as the number of words in the premise (i.e. $l$ times).

To be specific, the Image Attention representation $\boldsymbol{m}_i$ ($i$-th column vector in $\boldsymbol{M}$) of the $i$-th word in the premise is obtained from a non-linear combination of the premise's hidden state $\bar{\boldsymbol{a}}_i$ and the transformation of image representation $\boldsymbol{c}$ [29]. With the guidance of the image, the relevant words are selected to form the first-level sentence representation $\boldsymbol{c}_a$. Therefore *CADAN* can understand

what the sentence is discussing under the image context information and model the inference relation in term of contents.

***(B) Sentence Attention Layer:*** However, knowing what exactly each sentence discusses is still not enough. What NLI is concerned with is the relations between two sentences. Thus, we also need to model the interaction between two sentences. Since sentence interaction can obtain mutual valued information of the premise and hypothesis, it will help to grasp the local relations in the premise and hypothesis. In order to further characterize the relationship between sentences, we propose Sentence Attention layer to analyze the interaction and local relations from detailed perspectives.

In this layer, we first send the $\{\boldsymbol{a}\}_{j=1}^{l}$ for premise sentence and $\{\boldsymbol{b}\}_{j=1}^{l}$ for hypothesis sentence to another GRU:

$$\bar{\boldsymbol{a}'}_i = \mathrm{GRU}_2(\{\boldsymbol{a}_{j=1}^{i}\}), \quad \bar{\boldsymbol{b}'}_i = \mathrm{GRU}_2(\{\boldsymbol{b}_{j=1}^{i}\}), \quad i = 1, 2, \ldots, l. \tag{3}$$

After getting hidden states $\{\bar{\boldsymbol{a}'}\}_{i=1}^{l}$ and $\{\bar{\boldsymbol{b}'}\}_{i=1}^{l}$, we utilize Sentence Attention to model the local relations between hypothesis and premise. Since the first-level sentence representation $\boldsymbol{c}_a$ contains the information that the image is concerned with, it can help to model the local interaction between the premise and hypothesis sentences on the same aspect. Therefore, we treat the first-level sentence representation as the input of Sentence Attention to figure out the local relations between two sentences in this layer.

In other words, with the help of Sentence Attention, the words in the hypothesis that are more important to the premise will get higher weights. We can use these concerned words to generate the second-level representation of premise, which contains enriched information from textual information and image information. We perform attention again and take the same mechanism like Image Attention as follows:

$$\begin{aligned} &\bar{\boldsymbol{B}'} = [\bar{\boldsymbol{b}'}_1, \bar{\boldsymbol{b}'}_2, \ldots, \bar{\boldsymbol{b}'}_l], \quad \boldsymbol{M}' = \tanh(\boldsymbol{W}'\bar{\boldsymbol{B}'} + \boldsymbol{U}'\boldsymbol{c}_a \otimes \boldsymbol{e}_l),\ \boldsymbol{W}', \boldsymbol{U}' \in \mathbb{R}^{k*k}, \\ &\boldsymbol{\alpha}' = \mathrm{softmax}(\boldsymbol{\omega}'^{\mathrm{T}}\boldsymbol{M}'), \quad \boldsymbol{c}'_a = \bar{\boldsymbol{B}}\boldsymbol{\alpha}'^{\mathrm{T}},\ \boldsymbol{\omega}' \in \mathbb{R}^k, \end{aligned} \tag{4}$$

Different from Image Attention, here we treat the hidden states $\{\bar{\boldsymbol{b}'}\}_{i=1}^{l}$ of hypothesis sentence and first-level premise representation $\boldsymbol{c}_a$ as the inputs. In this way, the content in $\{\bar{\boldsymbol{b}'}\}_{i=1}^{l}$ that is relevant to $\boldsymbol{c}_a$ will be selected and represented as the second-level premise representation $\boldsymbol{c}'_a$.

***(C) Sentence Matching Layer:*** In order to determine the overall inference between two sentences, we leverage heuristic matching [4] between first-level sentence representations $\boldsymbol{c}_a, \boldsymbol{c}_b$ and second-level sentence representations $\boldsymbol{c}'_a, \boldsymbol{c}'_b$ after attention operation. Specifically, we use the element-wise product, their difference, and concatenation. Then we concatenate two calculated vectors $\boldsymbol{v}_a$ and $\boldsymbol{v}_b$ and send the result $\boldsymbol{v}$ to multi-layer perceptron (MLP) to calculate the probability of inference relation's existence between these sentence pairs. The MLP has two hidden layers with ReLu activation and a softmax output layer.

$$\begin{aligned} &\boldsymbol{v}_a = (\boldsymbol{c}_a,\ \boldsymbol{c}'_a \odot \boldsymbol{c}_a,\ \boldsymbol{c}'_a - \boldsymbol{c}_a,\ \boldsymbol{c}'_a), \quad \boldsymbol{v}_b = (\boldsymbol{c}_b,\ \boldsymbol{c}'_b \odot \boldsymbol{c}_b,\ \boldsymbol{c}'_b - \boldsymbol{c}_b,\ \boldsymbol{c}'_b), \\ &\boldsymbol{v} = (\boldsymbol{v}_a, \boldsymbol{v}_b), \quad P(y|(\boldsymbol{s_a}, \boldsymbol{s_b}, \boldsymbol{c})) = \mathrm{MLP}(\boldsymbol{v}). \end{aligned} \tag{5}$$

In this layer, concatenation can retain all the information [38]. The element-wise product is a certain measure of "similarity" of premise and hypothesis [21]. Their difference can capture the degree of distributional inclusion on each dimension [35].

### 3.3 Model Learning

In this section, we introduce the details about the model learning. Recalling the model description, the training processing can also be divided into two parts: (1) The preprocessing part: We separately train the AutoEncoder and fine-tune VGG19. (2) The inference part: The loss function we use in this part is softmax cross-entropy function.

To be specific, in both stages, mini-batch gradient descent is utilized to optimize the models, where the batch size is 64. The dimensions of feature representation of the image and the words are all 300. The lengths of premise and hypothesis are all set as 15. The state sizes of two GRU cells are set as 200, the dimensions of the parameters $\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{w}$ are also set as 200. To initialize the model, we randomly set the weights $\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{w}$ following the uniform distribution in the range between $-\sqrt{6/(nin + nout)}$ and $\sqrt{6/(nin + nout)}$ as suggested by [23]. We use SGD with momentum [30], where the learning rate and momentum are separately set as 0.05 and 0.6, and gradient clipping is performed to constrain the L2 norm of the global gradients do not exceed 1.0.

## 4 Experiments

In this section, we provide empirical validation on the large-scale NLI corpus and real-world NLI alike corpus, and utilize the parameter size and accuracy on different test sets to evaluate the models.

### 4.1 Dataset Description

**SNLI.** Stanford Natural Language Inference (SNLI) [2] has 570k human annotated sentence pairs with labels "entailment", "neutral", "contradiction". The premise data is drawn from the captions of the Flickr30k corpus. Thus, we can treat the corresponding images as the context. Since the hypothesis data is manually composed, annotation artifacts will lead the model correctly classify the hypothesis alone, Gururangan et al. [8] proposed a challenging hard subset, in which the premise-oblivious model cannot classify accurately, to better evaluate the models' ability to understand sentences. We also evaluate the models' performance on this test set.

**DanMu.** Different from SNLI that has been synthesized specifically for NLI task [11], DanMu data comes from the real world with labels "entailed" and "not-entailed". Both the premise and hypothesis data are user-generated time-sync comments on videos. Therefore, the corresponding video frames can be treated as

the context information. Moreover, these sentences are highly diverse in various aspects (length, complexity, expression, etc.), posing linguistic challenges for the task. By the nature of its construction, DanMu focuses on what a good context-aware NLI system needs to find out inference relation between sentence pairs.

To be specific, DanMu contains 120,650 sentence pairs with associated video frames from more than 4,000 movie videos, including 42,527 positive and 78,123 negative pairs with the labels "entailed" and "not-entailed". Each item contains one premise sentence $p$, one hypothesis sentence $h$, and the corresponding video frame.

Following [15], we extract the premise and the corresponding image from a short period [17], the hypothesis sentence is a modified variant of one of the comments from either the same period or a random, unrelated one. The instances that have high word overlap are removed. Then, each remaining instance is modified by three annotators. The annotator was given the instance and asked "*whether he can conclude the hypothesis from the premise and the image*". The majority of the answers from annotators was treated as the label of the instance. Figure 3(A) show some examples of this dataset.

**Baselines.** In order to better verify the performance of *CADAN*, we choose some sentence encoding-based NLI models and image captioning models as baselines.

- **LSTM encoders** [2]: encoding the premise and hypothesis with two different LSTMs.
- **W-by-W Attention** [29]: checking for inference relations of word-pairs and phrase-pairs between the premise and hypothesis.
- **BiLSTM with Inner-Attention** [16]: using bidirectional LSTM with inner attention mechanism to generating sentence representation for NLI.
- **CENN** [38]: utilizing different sentence vectors to determine the inference relation.
- **Gated-Att BiLSTM** [3]: employing intra-sentence gated-attention component to encodes a sentence to a fixed-length vector for NLI.
- **m-RNN** [20]: utilizing a deep RNN for sentences and a CNN for images to model the probability distribution of words.
- **NIC** [33]: utilizing a vision CNN and a language RNN for image captioning.

For these two models, we add the premise and hypothesis as inputs to RNN module separately and treat the final state of models as sentence representations. After getting sentences representations, we use Sentence Matching layer in *CADAN* to determine the inference relation in sentences pairs. Note that all the models use the same pre-trained word and image representations.

### 4.2 Overall Performance

We evaluate the performance of models and baselines from the following aspects: (A) The parameter size (*#Para.*); (B) The accuracy in (1) SNLI Full test set (*SNLI Full*); (2) SNLI Hard test set (*SNLI Hard*); (3) DanMu test set (*DanMu Test*).

**Table 1.** Performance (accuracy) of models for NLI.

| Model | #Para. | SNLI full | SNLI hard | DanMu test |
|---|---|---|---|---|
| (1) LSTM encoders | 3.0M | 80.6 | 58.5 | 64.9 |
| (2) CENN | <700K | 82.1 | 60.4 | 65.2 |
| (3) W-by-W Attention model | 3.9M | 83.5 | 61.7 | 66.9 |
| (4) BiLSTM with Inner-Attention | 2.8M | 84.2 | 62.7 | 66.3 |
| (5) Gated-Att BiLSTM | 12m | 85.5 | 65.5 | 67.3 |
| (6) CENN with image | <700K | 83.1 | 61.7 | 66.6 |
| (7) NIC | - | 84.3 | 63.6 | 67.9 |
| (8) m-RNN | - | 84.9 | 64.9 | 68.2 |
| (9) *CADAN* | 2.6M | **85.7** | **67.9** | **71.8** |

The overall results are summarized in Table 1. We can observe that *CADAN* achieves comparable performance. To be specific, *CADAN* utilizes Image Attention layer to generate first-level sentence representation, thus it can understand sentences in terms of content accurately. Then Sentence Attention layer is employed to model the interaction and local relations from detailed perspectives. Therefore, our model achieved the best performance on SNLI full test. Since SNLI hard test remove those examples that premise-oblivious model can classify correctly, the performance on this test set can better evaluate the models' ability. We can observe that *CADAN* outperformed all the baselines by a large margin, e.g. Gated-Att BiLSTM (+2.4%), BiLSTM with Inner-Attention (+5.2%).

**Compared with NLI Models.** LSTM encoders [2] encode sentences with different LSTMs and lead many related works to use different neural networks as encoders. Thus, we choose it as one of the baselines. However, 58.5% in hard test and 64.9% in DanMu test prove that simply encoding a sentence with its own textual information is not enough. CENN and its variant have less than 700K parameters, but they achieve comparable results with Word-by-Word Attention model [29], which have 3.9M parameters. It proves that context is really helpful for sentence understanding and NLI indeed. BiLSTM with Inner-Attention [16] uses intra-attention on top of BiLSTM to generate sentence representation, and Gated-Att BiLSTM [3] leverages the gate information in LSTM to calculate the importance of states of words. Thus, they can understand sentences with a finer granularity. However, when sentence semantics become obscure, like the hard test, their performances is not so good, which proves the context is essential for sentence semantic understanding and NLI.

**Compared with Variants of Image Captioning Models.** Since *CADAN* introduces the image, we want to figure out whether image captioning works can have good performance on this task. We choose NIC [33] and m-RNN [20] as baselines. They can generate sentence representation and adapt to the NLI

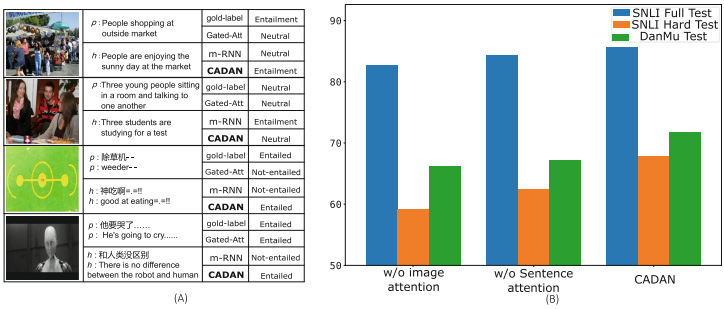


**Fig. 3.** Classification results of different models and ablation result of *CADAN*.

task through slight changes. From the results, we can conclude that they achieve comparable performance on the full test with the help of images. When sentences become complex, i.e. instances in hard test, their performances are still steady, which indicates the importance of images. However, their original purpose is generation rather than classification. They are good at aligning the images and sentence, but poor at modeling the interaction between sentences from detailed perspectives.

### 4.3 Ablation Performance

To investigate the effectiveness of the major components of *CADAN*, Fig. 3(B) provides additional analysis. From the best model, we remove the Image Attention layer, in which images are removed, to verify the performance of the model. We also remove the Sentence Attention layer to verify whether only Image Attention layer was enough. Without Image Attention layer, the performance drops to 59.2% (−8.7%) for hard test and 66.3% (−5.5%) for DanMu test, showing that incorporating image as the context is essential. Without Sentence Attention layer, the performance drops to 62.5% (−5.4%) for hard test and 67.2% (−4.6%) for DanMu test, proving that it's important to consider local relations between sentences from detailed perspectives. Based on these observations, we can summarize that contexts and sentence interaction are both very important for sentences semantic understanding.

### 4.4 Qualitative Evaluation

**Evaluation of the Results.** Here we choose the Gated-Att BiLSTM and m-RNN as they perform the best of the baselines in the NLI-related baselines and image captioning-related baselines for qualitative evaluation. The results are shown in Fig. 3(A). The first two examples come from SNLI hard test and the rest come from DanMu test. Taking the last instance as an example, this instance describes that a robot looks very sad like human beings. Without the image, the description of the premise will be ambiguous. We don't realize that

'He' is referring to an object rather than a person until we know the non-literal context. With the image, we could understand that the meaning of premise is that this robot may have the same emotions as humans since ordinary robots cannot be able to cry. Then it's easy to infer that this robot has no difference with humans. *CADAN* makes a right choice, while the other two misclassify it.

The rest examples also show the importance of images as contexts in Fig. 3(A). All of them indicate that context is essential for NLI.

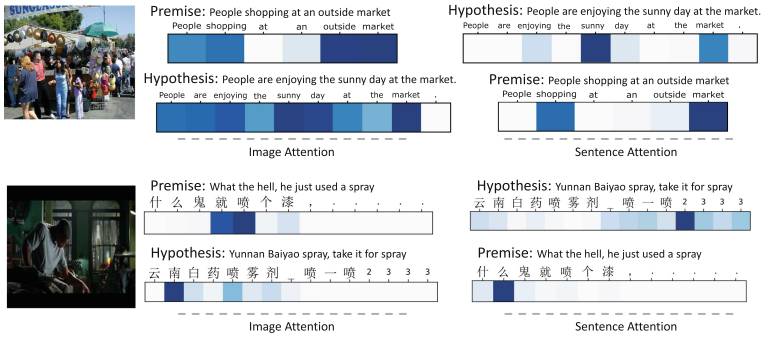


**Fig. 4.** Visualization of attention on two examples.

**Evaluation of the Attention.** Here we visualize the attention in our model. There are two kinds of attention: (1) The image's attention to each sentence: (2) The sentence's attention on each other. Figure 4 shows to what extent the Image Attention and Sentence Attention focus on the hidden states of two sentences respectively.

The example above may be confusing without the image. As described before, both premise and hypothesis describe that people are shopping at the market. However, the weather in the premise is unclear. We may conclude that the weather is sunny since the premise describes that the market is outside, which is hard for machines and we are not sure about the conclusion. However, with the image's help, it's easy for us to find out the relation is *entailment*. Moreover, *CADAN* focuses on the word "outside" in the premise and "sunny day" in hypothesis sentence. On this basis, *CADAN* also pay attention to "people shopping" in premise and "enjoying, market" in hypothesis. Therefore, our model not only makes the right classification, but also gives a clear explanation about the inference relation between the sentence pairs.

The example below, which comes from the movie "I, Robot"[1], also indicates that our model not only makes the right classification, but also gives a clear explanation.

With the information of the image, *CADAN* finds the alignment between "spray" in premise and "Yunnan Baiyao spray"[2] in hypothesis. Moreover,

[1] https://en.wikipedia.org/wiki/I,_Robot_(film).

[2] Yunnan Baiyao is a kind of healing spray.

*CADAN* finds that "2333" in hypothesis and "what the hell" in the premise both express the same feeling about the image. All these indicate "entailed" relation between the sentence pair.

In conclusion, when semantic meanings of sentences are clear, *CADAN* can make the right choice and give a detailed explanation about the inference relation. When sentence semantics are obscure, *CADAN* can utilize image as context to understand its meaning precisely and make the correct classification.

## 5 Conclusion and Future Work

In this paper, we argued that *context* is crucial for sentence understanding. We proposed a novel *Context-Aware Dual-Attention Network (CADAN)* to incorporate both textual and image information into the inference processing effectively. To be specific, we utilized Image Attention to incorporate image to understand the semantic meaning of sentences in terms of contents. Then Sentence Attention was employed to model the interaction and local relations of sentences from detailed perspectives. With the help of two-level representations and dual-attention mechanisms, our model could better understand sentence semantic and make correct decision. Experimental results demonstrated the superiority of our proposed model. In the future, we will explore more effective ways to process the context and finer grained methods to understand sentences semantics more precisely.

**Acknowledgements.** This research was partially supported by grants from the National Key Research and Development Program of China (No. 2016YFB1000904) and the National Natural Science Foundation of China (Grants No. 61727809, U1605251, 61572540, and 61751202).

## References

1. Altmann, G., Steedman, M.: Interaction with context during human sentence processing. Cognition **30**(3), 191–238 (1988)
2. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: EMNLP (2015)
3. Chen, Q., Zhu, X., Ling, Z.H., Wei, S., Jiang, H., Inkpen, D.: Recurrent neural network-based sentence encoder with gated attention for natural language inference. arXiv preprint arXiv:1708.01353 (2017)
4. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., Inkpen, D.: Enhanced LSTM for natural language inference. In: ACL. ACL, Vancouver, July 2017
5. Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading. In: EMNLP (2016)
6. Cho, K., Courville, A.C., Bengio, Y.: Describing multimedia content using attention-based encoder-decoder networks. IEEE Trans. Multimed. **17**, 1875–1886 (2015)
7. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555 (2014)

8. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. arXiv preprint arXiv:1803.02324 (2018)
9. Huang, Z., et al.: Question difficulty prediction for READING problems in standard tests. In: AAAI (2017)
10. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR, pp. 3128–3137 (2015)
11. Khot, T., Sabharwal, A., Clark, P.: SciTail: a textual entailment dataset from science question answering. In: AAAI (2018)
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. CoRR abs/1312.6114 (2013)
13. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using Fisher Vectors. In: CVPR, pp. 4437–4446 (2015)
14. Kun, Z., Guangyi, L., Le, W., Enhong, C., Qi, L., Han, W.: Image-enhanced multi-level sentence representation net for natural language inference. In: ICDM (2018)
15. Lai, A., Bisk, Y., Hockenmaier, J.: Natural language inference from multiple premises. In: IJCNLP (2017)
16. Liu, Y., Sun, C., Lin, L., Wang, X.: Learning natural language inference using bidirectional LSTM model and inner-attention. CoRR abs/1605.09090 (2016)
17. Lv, G., Xu, T., Chen, E., Liu, Q., Zheng, Y.: Reading the videos: temporal labeling for crowdsourced time-sync videos based on semantic embedding. In: AAAI (2016)
18. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: ICCV, pp. 2623–2631 (2015)
19. MacCartney, B.: Natural Language Inference. Stanford University, Stanford (2009)
20. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L.: Deep captioning with multimodal recurrent neural networks (m-RNN). CoRR abs/1412.6632 (2014)
21. Mou, L., et al.: Natural language inference by tree-based convolution and heuristic matching. In: ACL (2016)
22. Munkhdalai, T., Yu, H.: Neural tree indexers for text understanding. CoRR abs/1607.04492 (2016)
23. Orr, G.B., Müller, K.R.: Neural Networks: Tricks of the Trade. Springer, Heidelberg (2003)
24. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: CVPR, pp. 4594–4602 (2016)
25. Parikh, A.P., Täckström, O., Das, D., Uszkoreit, J.: A decomposable attention model for natural language inference. In: EMNLP (2016)
26. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)
27. Clark, P., et al.: Combining retrieval, statistics, and inference to answer elementary science questions. In: AAAI (2016)
28. Ren, M., Kiros, R., Zemel, R.S.: Exploring models and data for image question answering. In: NIPS (2015)
29. Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kociský, T., Blunsom, P.: Reasoning about entailment with neural attention. CoRR abs/1509.06664 (2015)
30. Ruder, S.: An overview of gradient descent optimization algorithms. CoRR abs/1609.04747 (2016)
31. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. In: EMNLP (2015)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)

33. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR, pp. 3156–3164 (2015)
34. Wang, S., Jiang, J.: Learning natural language inference with LSTM. In: HLT-NAACL (2016)
35. Weeds, J., Clarke, D., Reffin, J., Weir, D.J., Keller, B.: Learning to distinguish hypernyms and co-hyponyms. In: COLING, pp. 2249–2259 (2014)
36. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. CoRR abs/1704.05426 (2017)
37. Yin, Y., et al.: Transcribing content from structural images with spotlight mechanism. In: KDD (2018)
38. Zhang, K., Chen, E., Liu, Q., Liu, C., Lv, G.: A context-enriched neural network method for recognizing lexical entailment. In: AAAI (2017)
39. Zheng, X., Feng, J., Chen, Y., Peng, H., Zhang, W.: Learning context-specific word/character embeddings. In: AAAI (2017)