Resolving Word Vagueness with Scenario-guided Adapter for Natural Language Inference

Yonghao Liu[1] , **Mengyu Li**[1] , **Di Liang**[2] , **Ximing Li**[1] , **Fausto Giunchiglia**[3] ,
Lan Huang[1] , **Xiaoyue Feng**[1*] and **Renchu Guan**[1*]
[1]Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, College of Computer Science and Technology, Jilin University
[2]Fudan University
[3]University of Trento
{yonghao20, mengyul21}@mails.jlu.edu.cn,
liximing86@gmail.com, fausto.giunchiglia@unitn.it
{huanglan, fengxy, guanrenchu}@jlu.edu.cn

Abstract

Natural Language Inference (NLI) is a crucial task in natural language processing that involves determining the relationship between two sentences, typically referred to as the premise and the hypothesis. However, traditional NLI models solely rely on the semantic information inherent in independent sentences and lack relevant situational visual information, which can hinder a complete understanding of the intended meaning of the sentences due to the ambiguity and vagueness of language. To address this challenge, we propose an innovative **ScenaFuse** adapter that simultaneously integrates large-scale pre-trained linguistic knowledge and relevant visual information for NLI tasks. Specifically, we first design an image-sentence interaction module to incorporate visuals into the attention mechanism of the pre-trained model, allowing the two modalities to interact comprehensively. Furthermore, we introduce an image-sentence fusion module that can adaptively integrate visual information from images and semantic information from sentences. By incorporating relevant visual information and leveraging linguistic knowledge, our approach bridges the gap between language and vision, leading to improved understanding and inference capabilities in NLI tasks. Extensive benchmark experiments demonstrate that our proposed ScenaFuse, a scenario-guided approach, consistently boosts NLI performance.

1 Introduction

"*Humans rely less on words than on visual images, auditory images and propositions or rules of logic in order to think. Even commonly used metaphors and analogies employ visual and spatial attributes to provide us with a quick and easy context in which to communicate and build thoughts [Pinker, 2003],*" claimed a world-renowned linguist and evolutionary psychologist, *Steven Pinker*. This statement highlights the potential significance of visual information in human cognitive processing of real-world challenges, suggesting that it may, in fact, outweigh the importance of symbolic natural language information. This proposition warrants further investigation and consideration within the context of cognitive science and human-computer interaction research.

Natural language inference (NLI), as a vital task in NLP, has received widespread attention since it is closely related to human language comprehension and reasoning abilities. NLI involves knowledge from multiple fields, such as natural language understanding [Storks *et al.*, 2019; Liu *et al.*, 2024b] and logical reasoning [Li *et al.*, 2024; Liu *et al.*, 2024a], and is a challenging task. It has been widely used in various practical application scenarios, such as question answering [Liu *et al.*, 2023a; Liu *et al.*, 2023b], text summarization [Falke *et al.*, 2019; Guan *et al.*, 2021; Liu *et al.*, 2021], and information retrieval [Murty *et al.*, 2021; Liu *et al.*, 2022]. Generally, NLI consists of two sentences, referred to as the premise and the hypothesis. The premise is an assertive sentence that provides descriptive information, while the hypothesis is a sentence that requires inference based on the premise. The goal of NLI is to determine whether the hypothesis can be inferred from the premise (*i.e.*, entailment), or contradicts the premise (*i.e.*, contradiction), or has no relationship with it (*i.e.*, neutral).

Recently, modern NLI models [Chen *et al.*, 2017; Conneau *et al.*, 2017a] have made significant strides by leveraging the powerful feature extraction capabilities of deep learning to enhance their ability to understand and represent the semantic information of sentences. Successfully determining the semantic relationship in NLI tasks requires a deep comprehension of the semantics between sentences. Moreover, to further promote the development of NLI, researchers use human-labeling approaches to obtain large-scale high-quality annotated datasets, such as SNLI [Bowman *et al.*, 2015] and MultiNLI [Williams *et al.*, 2018], to evaluate the effectiveness of the proposed models. Despite the fruitful success achieved on these datasets, most previous models are still lim-

[*] Corresponding Author

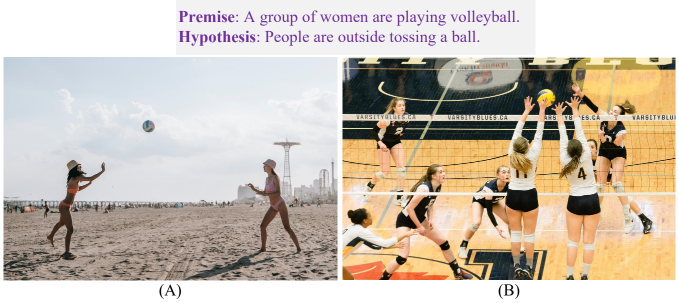
Premise: A group of women are playing volleyball.
Hypothesis: People are outside tossing a ball.
(A)
(B)

Figure 1: The two sentences have an entailment relationship when viewed in the context of image (A), but a contradiction relationship when viewed in the context of image (B).

ited in performance due to their focus on extracting textual information while ignoring other useful information, such as visuals, which are crucial for proper sentence semantics understanding that can be used to discern the semantic relations between sentences.

Previous studies [Gibson *et al.*, 2019] have highlighted the limitations of language due to its inherent ambiguity and vagueness. This often leads to challenges in interpretation, as evident in the SNLI dataset, where different annotators may perceive the same pair of sentences as either an entailment or a contradiction. In contrast, images are more concrete and can convey complementary information [Pan *et al.*, 2016] that is often overlooked by sentences such as facial expressions and motions, which can effectively solve textual ambiguity [Tong *et al.*, 2020]. Thus, exploring the complementary information conveyed by images can help gain better insights into actual sentence meanings. As shown in Fig. 1, both the premise and hypothesis describe the scenario of people playing with balls. However, while the hypothesis specifies an outdoor setting, the location in the premise remains ambiguous. This ambiguity indicates a neutral relationship between the premise and hypothesis. Once the image related to the premise is provided, the relationship between the premise and hypothesis can be determined with certainty. In Fig. 1 (A), the location information suggests that people are outdoors, creating an entailment relationship between the premise and hypothesis. Conversely, in Fig. 1 (B), the location information indicates people are indoors, resulting in a contradictory relationship between the two. From this example, we can conclude that scenario images further enrich the semantics of sentences and provide necessary supplementary information. Although many images are gathered alongside the sentences during dataset collection, they are not given much attention. Hence, it is necessary to leverage relevant scenario images to fully understand the sentence semantics for NLI tasks.

Our objective is to incorporate scenario information into our sentence-level representations in a precise manner, which poses a considerable challenge. Naive integration of the two modalities can introduce noise that hinder semantic understanding, so finding an effective way to merge images with sentences remains an open problem. Although extensive research on multimodal tasks, there is currently no widely accepted method for combining image modality with sentences.

To address the aforementioned challenges, we propose a novel scenario-guided adapter, coined **ScenaFuse**, for NLI tasks. Our approach involves designing an image-sentence interaction module that generates scenario-guided sentence semantic representations by deeply interacting images and sentences in the attention module of the pre-trained model. This module benefits the semantic understanding by explicitly considering the scenario images. Consequently, the text representations obtained from our approach incorporate valuable scenario information, are more comprehensive, and possess stronger representational power compared to those generated using a single modality. Moreover, we introduce an attention-based image-sentence fusion module with the gate mechanism to perform multimodal feature fusion related to sentence semantic and scenario-guided representations. This module selects informative features from both modalities while minimizing the noise impact, which facilitates the evaluation of sentence relationships. In summary, our contributions are listed as follows:

- We introduce a novel scenario-guided adapter named **ScenaFuse**, which explicitly incorporates visual information from the relevant scenarios. This is a valuable attempt at integrating scene information into large pre-trained language models for NLI tasks.
- We propose two modules: the image-sentence interaction module and the image-sentence fusion module. The former enables deep interaction between images and sentences, and the latter integrates two different modalities of features through specifically designed mechanisms. Moreover, we further demonstrate the effectiveness of the proposed approach on recent popular large language models.
- We conduct extensive experiments on benchmarks to evaluate the effectiveness of our approach. Our empirical results demonstrate the superiority of ScenaFuse compared to other competitive baselines.

2 Related Work

2.1 Natural Language Inference

NLI models can be mainly divided into three categories: symbolic, statistical, and neural network-based models [Storks *et al.*, 2019]. Symbolic models use logical forms to perform inference tasks and such models are heavily used in early NLI problems known as recognizing textual entailment (RTE) [MacCartney, 2009]. As data-driven feature training becomes more prevalent in NLP, statistical models based on feature engineering such as bag-of-words [Zhang *et al.*, 2010], become the mainstream approach. Some work [Haim *et al.*, 2006; Dagan *et al.*, 2006] also attempts to incorporate external knowledge as a supplement to training data features and achieve favorable results in NLI. The recent advancement in deep learning, along with the availability of larger annotated datasets, has led to the success of more complex neural network models in NLI tasks. These models can be categorized into two types: sentence-encoding-based and inter-sentence-attention-based. Sentence-encoding-based models [Conneau *et al.*, 2017b] utilize the same neural network architecture, such as LSTM [Bowman *et al.*, 2015] and GRU [Vendrov *et al.*, 2016] and their variants, to encode both the premise and hypothesis. A neural network classifier is then utilized to predict the relationship between the two sentences. On the other hand, models based on inter-sentence attention [Rocktäschel *et al.*, 2016; Wang and Jiang, 2016] introduce attention mechanisms between the premise and hypothesis. This not only alleviates the vanishing gradient problem but also provides alignment between inputs and outputs, which allows for a better understanding of their relationships.

2.2 Multimodal Learning

Multimodal data has garnered considerable attention from both academia and industry due to its potential applications in learning, analysis, and research [Ramachandram and Taylor, 2017]. Multimodal models aim to extract information from varying modalities to learn expressive features that can be used for downstream tasks. The core of multimodal learning is to map unimodal data into a multimodal space, and subsequently acquiring the joint representation. Classic methods involve concatenating features from each modality and training the model, while more advanced approaches utilize neural networks to input each modality into several separate neural layers in an end-to-end manner, and then project them into hidden layers in the joint space [Antol *et al.*, 2015]. This joint multimodal representation can then be processed through multiple hidden layers or directly used for prediction. Due to the incorporation of large quantities of data during neural network training, the resulting joint representation typically performs well on various tasks. Recently, many studies have introduced multimodality into NLP tasks, such as named entity recognition [Moon *et al.*, 2018] and machine translation [Nishihara *et al.*, 2020]. Generally, these methods enhance rough text understanding from non-textual perspectives by employing modality attention mechanisms to integrate information from heterogeneous sources. For example, MNMT [Nishihara *et al.*, 2020] incorporates a supervised visual attention mechanism, which is trained with constraints between manually aligned words in the sentence and corresponding regions in the image. This mechanism can more accurately capture the relationship between words and image regions. MNER [Moon *et al.*, 2018], on the other hand, proposes a universal modality-attention module that learns to reduce irrelevant modalities while amplifying the most informative ones for named entity recognition of data composed of text and images.

3 Problem Definition

In this section, we define the NLI problem in detail. Consider a scenario where we have been provided with a premise consisting of m words, represented by $P = (w_{p,1}, w_{p,2}, \cdots, w_{p,m})$, as well as a hypothesis with n words, depicted by $H = (w_{h,1}, w_{h,2}, \cdots, w_{h,n})$. Additionally, an associated image I is also available. The token embeddings $\mathbf{X}_{tex}$ in P and H are the sum of word, segment and position embeddings for each token, where the word embeddings can be initialized randomly or using word2vec or GloVe, *i.e.*, $\mathbf{X}_{tex,i} \in \mathbb{R}^d$. In general, we follow the experimental settings of large-scale pre-trained models such as BERT [Devlin *et al.*, 2019] or RoBERTa [Liu *et al.*, 2019b]. To formalize two sentences into this format, we use the sentence pair "[CLS] P [SEP] H [SEP]" as model inputs. Here, [CLS] is a special symbol added to the beginning of each input example, and [SEP] is used as a special separator to separate sentences. We extract the [CLS] token as the output of each example. Because it lacks clear semantic information and can better integrate the semantic information of each word in the sentence. As a result, it represents the sentence's semantics better. Subsequently, we feed it into a fully connected layer for classification tasks.

For the associated image I, we reshape it 224x224 pixels and input it into the pre-trained image encoder Enc$(\cdot)$, such as VGG [Simonyan and Zisserman, 2014] or ResNet [He *et al.*, 2016], which have already demonstrated the ability to extract meaningful representations from input images in deep layers. Subsequently, we obtain its visual representations by retaining the output of the last convolutional layer. This process essentially segments each input image into visual blocks of the same size represented by d'-dimensional vectors, denoted as $\mathbf{X}_{vis} = \text{Enc}(I) = (\mathbf{X}_{vis,1}, \mathbf{X}_{vis,2}, \cdots, \mathbf{X}_{vis,k}) \in \mathbb{R}^{k \times d'}$, where $\mathbf{X}_{vis,i}$ denotes the feature representation of the i-th visual block.

In summary, we expect to train a model $\mathcal{F}$ that can precisely determine the relationship between the premise and the hypothesis when relevant data is provided, *i.e.*, $y = \mathcal{F}(P, H, I) \in \{0, 1, 2\}$, where y=0 implies an entailment relationship, y=1 represents a neutral relationship, and y=2 represents a contradiction relationship.

4 Method

This section provides an in-depth explanation of the two main modules that constitute the scenario-guided adapter, namely the image-sentence interaction and the image-sentence fusion. To facilitate a better understanding of our model, we have depicted its overall framework in Fig. 2.

4.1 Image-sentence Interaction

As stated in the abstract, compared to language, images provide a more precise means of conveying information and may include important details that are difficult to express or neglected by language due to its inherent ambiguity and vagueness. Moreover, multimodal information has the potential to provide more discriminative input than a single modality, which can help address the issue of semantic ambiguity in text-only modality. Therefore, we explicitly incorporate visual contextual information to significantly enrich the semantic content of the sentence. To this end, we design an image-sentence interaction module. Specifically, we first adopt a linear transformation to project both modalities into the same representation space to achieve embedding space consistency and cross-modal semantic alignment between text representation modality $\mathbf{X}_{tex}$ and visual representation modality $\mathbf{X}_{vis}$. This can be expressed as follows:

$$\begin{aligned} \tilde{\mathbf{X}}_{tex} &= \varphi(\mathbf{X}_{tex}) \\ \tilde{\mathbf{X}}_{vis} &= \psi(\mathbf{X}_{vis}) \end{aligned} \tag{1}$$

where $\varphi(\cdot)$ and $\psi(\cdot)$ are linear projection functions that project the text representation and visual representation to the

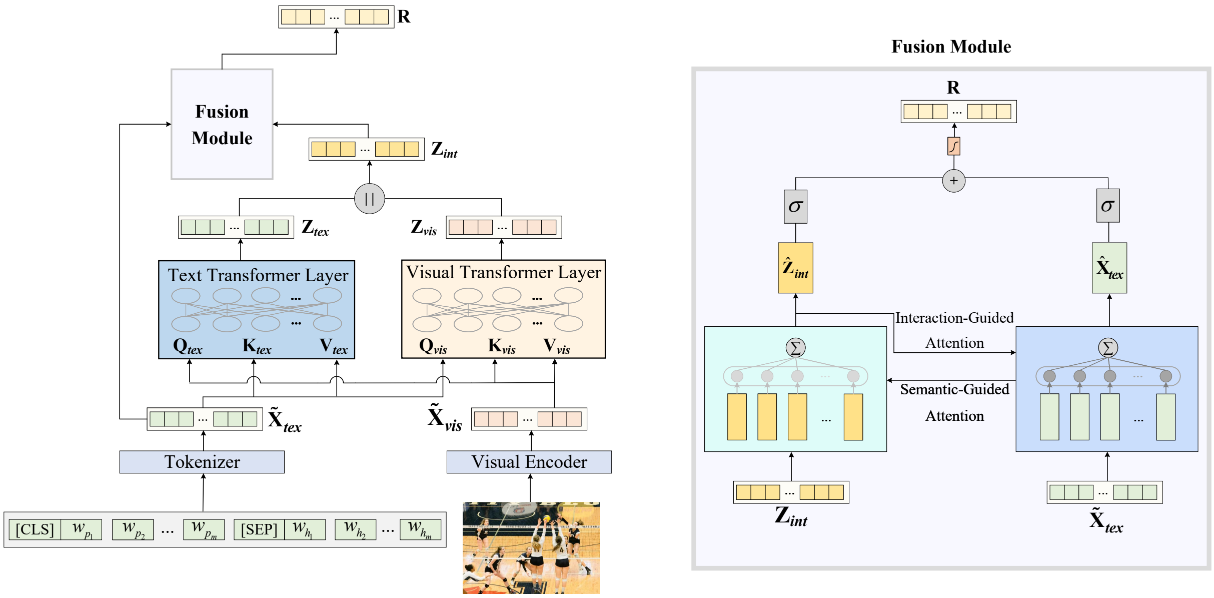

Figure 2: (Left) Overall architecture of our framework. (Right) The architecture of the fusion module. (Best viewed in color)

same generated space, *i.e.*, $\varphi : \mathbb{R}^d \to \mathbb{R}^f$ and $\psi : \mathbb{R}^{d'} \to \mathbb{R}^f$. $\tilde{\mathbf{X}}_{tex}$ and $\tilde{\mathbf{X}}_{vis}$ are generated sentence representations and visual representations, respectively.

Visual-enhanced Sentence Representation. Subsequently, we consider that the textual content in NLI tasks is limited. To address this limitation and enable the utilization of visual information to enhance word representation learning, we employ a multi-head attention mechanism akin to the Transformer architecture. Concretely, for each head, visual representations $\tilde{\mathbf{X}}_{vis}$ are treated as queries, while the textual representations $\tilde{\mathbf{X}}_{tex}$ are treated as keys and values, which can be formulated as:

$$\mathbf{Z}_{tex} = \text{softmax}(\frac{\mathbf{Q}_{tex}\mathbf{K}_{tex}^{\top}}{\sqrt{t}})\mathbf{V}_{tex} \qquad (2)$$

where $\mathbf{Q}_{tex}, \mathbf{K}_{tex}$, and $\mathbf{V}_{tex}$ are query vectors, key vectors, and value vectors obtained by applying the corresponding linear transformations. $\mathbf{Z}_{tex} \in \mathbb{R}^{k \times t}$ is the derived visual-enhanced sentence representation, where k is the length of all visual blocks. After performing Eq.2, the sentence explicitly interacts with the associated image.

Sentence-rectified Visual Representation. Furthermore, it is important to note that each visual block may have a close association with multiple input words. To gain a deeper understanding of the relationship between images and sentences, it is necessary to align the semantic information of each visual block with its relevant words. Additionally, we aim to impose certain constraints on the image through corresponding sentences to alleviate irrelevant associations. To achieve this, we use the following Eq.3 to obtain sentence-rectified visual representations, which can complement the previously obtained visual-enhanced sentence representations. Specifically, employing the multi-head attention mechanism, the sentence representation $\tilde{\mathbf{X}}_{tex}$ acts as queries, while the visual representation $\tilde{\mathbf{X}}_{vis}$ serves as keys and values. The procedure can be formulated as follows:

$$\mathbf{Z}_{vis} = \text{softmax}(\frac{\mathbf{Q}_{vis}\mathbf{K}_{vis}^{\top}}{\sqrt{t}})\mathbf{V}_{vis} \qquad (3)$$

where $\mathbf{Q}_{vis}, \mathbf{K}_{vis}$, and $\mathbf{V}_{vis}$ are also query vectors, key vectors, and value vectors derived by performing the corresponding linear transformations. $\mathbf{Z}_{vis} \in \mathbb{R}^{l \times t}$ is the obtained sentence-rectified representation, which can be interpreted as constraining visual semantics from a textual perspective, where l is the length of all the tokens involved in the process.

Afterwards, we merge $\mathbf{Z}_{tex}$ and $\mathbf{Z}_{vis}$ by concatenation and apply a fully connected layer to serve as the output of the image-sentence interaction module. This process can be denoted as:

$$\mathbf{Z}_{int} = \phi(\mathbf{Z}_{tex}||\mathbf{Z}_{vis}) \qquad (4)$$

where $\phi(\cdot)$ is another linear projector that projects visual-enhanced sentence representation and sentence-rectified visual representation into an interaction space, *i.e.*, $\phi : \mathbb{R}^{k+l} \to \mathbb{R}^l$. $\mathbf{Z}_{int} \in \mathbb{R}^{l \times t}$ denotes the generated final embeddings of the full interaction of the two modalities. By incorporating information from both modalities, the resulting embeddings offer a more comprehensive representation compared to single-modal input.

4.2 Image-sentence Fusion

Directly using the generated interactive embeddings may introduce potential bias and lead to suboptimal results. To alleviate this, we further design an image-sentence fusion module that applies an adaptive fusion strategy based on a gating mechanism to obtain more precise sentence-level representations and deduce the spread of errors caused by visual bias. This strategy requires simultaneously leveraging pure textual semantic representations $\tilde{\mathbf{X}}_{tex}$ and interaction representations $\mathbf{Z}_{int}$ obtained from the image-sentence interaction module. The specific architecture of the proposed module is illustrated in Fig. 2.

We first compute the attention coefficients α using pure semantic representation to update the interaction representations. Then, we obtain the attention coefficients β based on the updated interaction representations $\hat{\mathbf{Z}}_{int}$ and utilize them to obtain the updated semantic representation $\hat{\mathbf{X}}_{tex}$. This process can be formulated as follows:

$$\begin{aligned}
\alpha &= \tanh[(\tilde{\mathbf{X}}_{tex}||\mathbf{Z}_{int})\mathbf{W}_{\alpha} + b_{\alpha}] \\
\hat{\mathbf{Z}}_{int} &= \mathbf{Z}_{int} \odot \mathrm{softmax}(\alpha \mathbf{W}_z + b_z) \\
\beta &= \tanh[(\hat{\mathbf{Z}}_{int}||\tilde{\mathbf{X}}_{tex})\mathbf{W}_{\beta} + b_{\beta}] \\
\hat{\mathbf{X}}_{tex} &= \tilde{\mathbf{X}}_{tex} \odot \mathrm{softmax}(\beta \mathbf{W}_x + b_x)
\end{aligned} \tag{5}$$

where $\{\mathbf{W}_{\alpha},\ \mathbf{W}_{\beta}\} \in \mathbb{R}^{2t\times 1}$ and $\{\mathbf{W}_z,\ \mathbf{W}_x\} \in \mathbb{R}^{1\times t}$ are weight matrices. $\{b_{\alpha}, b_z, b_{\beta}, b_x\}$ are the corresponding biases. $||$ and $\odot$ denote the concatenation operation and element-wise multiplication, respectively.

Next, based on the gate mechanism, we design an approach that adaptively captures and fuses valuable and informative features from updated semantic and interaction representations, which is defined as follows:

$$\begin{aligned}
g &= \sigma((\hat{\mathbf{X}}_{tex}||\hat{\mathbf{Z}}_{int})\mathbf{W}_g + b_g) \\
\mathbf{U} &= g \cdot \hat{\mathbf{X}}_{tex} + (1-g) \cdot \hat{\mathbf{Z}}_{int}
\end{aligned} \tag{6}$$

where g is the calculated gating coefficient that dynamically controls the contribution of different representations. $\mathbf{W}_g \in \mathbb{R}^{2t\times 1}$ denotes the trainable parameter. b_g is the bias. $\sigma(\cdot)$ represents the sigmoid function. With Eq.6, we can obtain the fusion feature $\mathbf{U} \in \mathbb{R}^{l\times t}$, which adaptively integrates semantic and interaction representations.

Finally, as we mentioned before, considering the potential noise in an image, it is unreasonable to align such noise with visual blocks and sentences. To address this issue, a filtering mechanism is required to selectively leverage the fusion feature $\mathbf{U}$. If incorporating $\mathbf{U}$ enhances the model's ability to understand a sentence, then both the fusion and original semantic features are absorbed by the filter; otherwise, it filters out fusion information. Concretely, this filtering procedure can be expressed as follows:

$$\begin{aligned}
h &= \sigma((\mathbf{U}||\tilde{\mathbf{X}}_{tex})\mathbf{W}_h + b_h) \\
\mathbf{R} &= h \odot \tanh(\mathbf{U}\mathbf{W}_r + b_r)
\end{aligned} \tag{7}$$

where $\mathbf{W}_h \in \mathbb{R}^{2t\times 1}$ and $\mathbf{W}_r \in \mathbb{R}^{t\times t}$ are weight parameters. b_h and b_u are the biases. $h \in \mathbb{R}^{l}$ determines which fusion information needs to be retained. $\mathbf{R} \in \mathbb{R}^{l\times t}$ denotes the filtered representations obtained by the filtering mechanism for NLI tasks.

The classic pre-trained model based on the Transformer architecture adopts a multi-head attention mechanism. Each attention head eventually acquires an attention output $\hat{\mathbf{A}}$. However, in our designed scenario-guided adapter, we replace the filtered features $\mathbf{A}$ with the features $\hat{\mathbf{A}}$ obtained by the attention mechanism in pre-trained models. This ensures that the model can explicitly incorporate scenario information, thereby more comprehensively understanding sentences and alleviating word vagueness problems.

4.3 Model Training

After obtaining the filtered representation $\mathbf{R}$, we input it into the remaining parts of the pre-trained model based on Transformer architecture, such as layer regularization and feed-forward neural network layers, to obtain the final sentence semantic vector $\mathbf{F} \in \mathbb{R}^{\ell\times t}$ for NLI tasks, where ℓ represents the number of training examples.

Additionally, the objective function of this task is the cross-entropy function, denoted as:

$$\begin{aligned}
\hat{Y} &= \mathrm{softmax}(\mathbf{F}\mathbf{W}_f) \\
\mathcal{L} &= -\sum\nolimits_{i=1}^{\ell} y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)
\end{aligned} \tag{8}$$

where $\hat{Y}$ is the output probability derived by a linear classifier parameterized by $\mathbf{W}_f \in \mathbb{R}^{t\times 3}$.

5 Experiment

Dataset. We employ three datasets widely used by previous methods to validate the effectiveness of our proposed scenario-guided adapter. The statistical information of these datasets can be found in Table 1. The brief introduction to the datasets is provided below. (I) **SNLI** [Bowman *et al.*, 2015] is a large-scale dataset designed for NLI tasks, where premises are summarized from photo captions in the Flickr30k and hypotheses generated by humans. *Each instance in the dataset corresponds to an image, which provides scenario information.* (II) **SNLI-hard** [Gururangan *et al.*, 2018] is built upon the SNLI dataset, but it excludes examples from the original test set that have annotation artifacts. (III) **SNLI-lexical** [Glockner *et al.*, 2018] is also based on the SNLI. However, it creates a new test set that requires simple lexical inference. Note that SNLI-hard and SNLI-lexical have the same training and validation sets as SNLI.

Baselines. We introduce four selected categories of baseline models for comparison. (I) *Classical sentence-based neural models* include **LSTM** [Hochreiter and Schmidhuber, 1997], **ESIM** [Chen *et al.*, 2017], **CAFE** [Tay *et al.*, 2018b], and **CSRAN** [Tay *et al.*, 2018a]. (II) *Pre-trained language-based models* contain BERT [Devlin *et al.*, 2019], RoBERTa [Liu *et al.*, 2019b], **UERBERT** [Xia *et al.*, 2021], **SemBERT** [Zhang *et al.*, 2020], and **MT-DNN** [Liu *et al.*, 2019a]. Here, we use the base and large versions of BERT and RoBERTa models, and for the remaining models, we adopt BERT-base as the encoder. (III) *Large language models* include **Bloom-7.1B** [Muennighoff *et al.*, 2022] and **Llama2-7B** [Touvron *et al.*, 2023]. Due to our limited computational resources, we choose the approximately 7B version to the best of our ability. (IV) *Multimodal models* consist of **NIC** [Vinyals *et al.*, 2015], **m-RNN** [Mao *et al.*, 2015], **IEMLRN** [Zhang *et al.*, 2018], **MIESR** [Zhang *et al.*, 2019], **VisualBERT** [Li *et al.*, 2019], and **CLIP** [Radford *et al.*, 2021].

Implementation Details. In the default experimental setting, we use a pre-trained ResNet-50 as the image encoder to initialize the visual representation $\mathbf{X}_{vis}$, which is later frozen during training. The obtained image embeddings $\mathbf{X}_{vis}$ and sentence token embeddings $\mathbf{X}_{tex}$ are input into the bottom Transformer block of the pre-trained BERT-base model after

Dataset	Entailment	Neutral	Contradiction	Total	Avg.word	
					premise	hypothesis
SNLI	3,368	3,237	3,219	10,000	13.9	7.5
SNLI-hard	1,058	1,135	1,068	3,261	13.8	7.7
SNLI-lexical	982	7,164	47	8,193	11.4	11.6

Table 1: The statistics in the test set of the evaluated datasets.

linear projection. The BERT-base model is fine-tuned during training. Moreover, we use the AdamW optimizer with learning rate values of {1e-5, 2e-5, 3e-5, 5e-5}. The warm-up and weight decay are set as 0.1 and 1e-8, respectively. The batch size is determined by grid search in {16, 32, 64}. Additionally, the dropout is within the range of {0.1, 0.2, 0.3}. Meanwhile, we apply gradient clipping within {7.0, 10.0, 15.0} to prevent gradient explosion. We adopt three widely used evaluation metrics: accuracy (*abbr.* Acc), micro-precision (*abbr.* P), and micro-recall (*abbr.* R).

6 Result

Performance Comparison. We present the experimental results in Table 2. In-depth understanding based on the quantitative results are provided as follows.

• We observe that ScenaFuse outperforms all other models on all datasets, providing substantial evidence of its effectiveness in handling NLI tasks. This outcome could be attributed to the designed scene-guided adapter consisting of an image-sentence interaction module and an image-sentence fusion module. The model incorporates scene information, which previous models did not fully utilize, into the analysis, thereby alleviating the inherent vagueness of sentences and enabling a more comprehensive understanding of their semantic information. Furthermore, the introduction of the image-sentence interaction module allows visual representations from images to enrich sentence semantics, while textual representations from sentences can rectify visual information, benefiting each other. Next, the image-sentence fusion module deeply integrates the information from both modalities to obtain more precise sentence representations.

• We observe that ScenaFuse can effectively integrate with recent popular large language models (LLMs), such as Bloom and Llama2. This integration significantly improves the model performance compared to those based on other pretrained language models. This phenomenon is expected, as LLMs are trained on massive high-quality data and possess excellent text understanding capabilities. Furthermore, we find that our model with LLMs still exhibits certain improvements compared to using LLMs alone. Because they ignore scenario information, which can guide words to produce more informative representations.

• We find that sentence-based neural and multimodal models show varying degrees of decline in performance on the SNLI-hard and SNLI-lexical datasets. This finding indicates that the previously proposed NLI models' success is predominantly limited to simple examples, and their ability to recognize text entailment may not be as proficient as anticipated. When the dataset becomes challenging, *i.e.*, removing annotation artifacts in the hypothesis (SNLI-hard) and requiring identifying specific words in the sentences (SNLI-lexical), it poses difficulties for all models.

Further Discussion. It is possible that someone may question whether scenario could directly replace the premise. If so, we only need to leverage appropriate scenarios and hypotheses to assess inferential relations. To address this doubt, we conduct an experiment where we remove all premise data from the dataset and denote the model variant as ScenaFuse$_{\text{rep}}$. From Table 3, we find that the model performance on the removed premises shows a serious degradation. Although the premises summarize the content of the scenario, there are still significant differences between them. Scenarios may serve as reference information to enhance semantic understanding of sentences, but cannot replace premises. Because it may contain noise irrelevant to the core semantics of the sentences. Our work primarily focuses on inference relations between sentences and differs from the work on image and sentence retrieval.

Moreover, we further explore whether the scenario be substituted with the corresponding sentences. We note that the premises are extracted from photo captions in the Flickr30k corpus, indicating that the premise is a summary of the photo. Therefore, this is equivalent to generating a synonymous sentence of the original premise. We use the synonym substitution to generate semantically identical sentences for each premise. Then, we combine the original and the replaced premises into an extended premise and feed it into BERT for training. This model variant is denoted as ScenaFuse$_{\text{sub}}$. According to Table 3, when the corresponding sentences replace the scenario, there is some performance improvement compared to BERT, but it is far behind our model.

Ablation Study. We conduct a series of ablation studies on the evaluation dataset by designing different model variants to investigate the impact of various model components on experimental results. (I) *w/o ISI*: We remove the image-sentence interaction (ISI) module and replace it with the classic multi-head attention mechanism of Transformers, which is equivalent to a complete lack of image information. (II) *w/o VESR*: We eliminate the visual-enhanced sentence representation (VESR) and only retain the sentence-rectified representation in the ISI module. (III) *w/o SRVR*: We eliminate the sentence-rectified representation (SRVR) and only retain the visual-enhanced sentence representation in the ISI module. (IV) *w/o ISF*: We remove the image-sentence fusion (ISF) module and replace it with a simple concatenation. (V) *w/o GM*: We remove the gate mechanism (GM) of the ISF module and replace it with an average of updated semantic and interaction representations. (VI) *w/o FM*: We remove the filtering mechanism (FM) of the ISF module and replace it with the average of fusion features and sentence representations.

Model	SNLI			SNLI-hard			SNLI-lexical		
	Acc	P	R	Acc	P	R	Acc	P	R
LSTM	80.62	80.49	80.55	58.53	58.60	58.49	52.31	55.59	52.30
ESIM	88.02	88.06	88.05	71.32	71.66	71.29	65.60	68.19	65.59
CAFE	88.52	88.59	88.39	72.15	72.35	72.14	66.17	69.73	66.17
CSRAN	88.71	88.81	88.65	72.51	72.79	72.49	66.71	70.07	66.70
BERT$_{\text{base}}$	90.66	90.43	90.46	80.50	80.57	80.49	92.62	95.41	92.63
RoBERTa$_{\text{base}}$	90.69	90.41	90.52	80.53	80.59	80.46	92.66	95.45	92.66
BERT$_{\text{large}}$	91.06	90.81	90.83	80.90	80.62	80.55	92.92	95.91	92.71
RoBERTa$_{\text{large}}$	91.23	90.49	90.52	80.82	80.66	80.51	92.65	95.46	92.72
UERBERT	90.78	90.61	90.52	80.62	80.59	80.60	92.73	95.61	92.71
SemBERT	90.90	90.56	90.59	81.06	81.15	81.05	92.81	95.72	92.80
MT-DNN	91.06	90.86	90.78	81.19	81.22	81.19	92.95	95.59	93.05
Bloom-7.1B	91.20	91.16	91.18	82.02	81.92	82.01	95.26	97.79	95.15
Llama2-7B	91.58	91.86	91.78	82.15	82.12	82.15	96.05	97.90	96.00
NIC	84.75	84.59	84.62	63.59	63.52	63.57	67.12	72.15	67.15
m-RNN	85.16	84.92	84.95	64.92	64.77	64.90	69.41	74.57	69.42
IEMLRN	87.52	87.30	87.35	75.43	75.36	75.40	78.15	83.15	78.16
MIESR	87.83	87.52	87.69	76.81	76.75	76.72	78.72	83.62	78.69
VisualBERT	91.66	91.32	91.29	82.09	82.25	81.96	93.02	95.96	93.22
CLIP	92.02	91.92	92.06	82.49	82.95	82.93	93.15	96.65	93.62
ScenaFuse$_{\text{bert-base}}$	92.16	92.06	92.12	83.25	83.16	83.20	94.05	97.07	94.04
ScenaFuse$_{\text{bert-large}}$	92.69	92.62	92.66	83.96	83.76	83.60	94.75	97.65	94.52
ScenaFuse$_{\text{bloom-7.1B}}$	93.11	93.06	93.16	84.52	84.50	84.46	96.62	98.66	96.60
ScenaFuse$_{\text{llama2-7B}}$	**93.19**	**93.32**	**93.26**	**84.86**	**84.82**	**84.66**	**97.25**	**98.75**	**97.12**

Table 2: Evaluation performance (%) of various models on three datasets.

Model	SNLI			SNLI-hard			SNLI-lexical		
	Acc	P	R	Acc	P	R	Acc	P	R
BERT$_{\text{base}}$	90.66	90.43	90.46	80.50	80.57	80.49	92.62	95.41	92.63
ScenaFuse$_{\text{rep}}$	80.16	80.32	80.12	70.31	69.55	69.92	79.10	77.16	77.22
ScenaFuse$_{\text{sub}}$	90.72	90.49	90.52	80.82	80.66	80.51	92.65	95.46	92.72
ScenaFuse	92.16	92.06	92.12	83.25	83.16	83.20	94.05	97.07	94.04

Table 3: The results (%) of two model variants on all datasets.

Model	SNLI	SNLI-hard	SNLI-lexical
w/o ISI	90.66	80.50	92.62
w/o VESR	90.72	81.22	92.91
w/o SRVR	90.85	81.36	93.02
w/o ISF	91.19	81.95	93.25
w/o GM	91.39	82.26	93.52
w/o FM	91.62	82.66	93.69
ScenaFuse	**92.16**	**83.25**	**94.05**

Table 4: The ablation results (%) for accuracy on all datasets.

We reach the following conclusions based on the results presented in Table 4. *Firstly*, all the designed components make important contributions to the model, and removing any of them results in varying degrees of performance degradation. *Secondly*, removing the entire ISI leads to the most significant decline in performance, indicating that scenario information can assist in sentence semantic understanding. *Thirdly*, removing either VESR or SRVR leads to a performance decline, indicating that the two modules benefit each other and only appear together to learn better word embeddings. *Fourth*, when the whole ISF module is deleted, performance drops due to the gap between interaction vectors and sentence vectors. Besides, removing the GM or FM of the ISF module also results in performance degradation due to potential noise interference in the image.

7 Conclusion

We present ScenaFuse, which is a novel model that addresses the word ambiguity and vagueness problem in NLI tasks by incorporating scenario information explicitly through visual inputs. We utilize pre-trained language models and introduce an image-sentence interaction module to extract features from both modalities. These features are then made to interact using Transformer layers. Additionally, we incorporate an image-sentence fusion module that is specially designed to adaptively fuse the learned features from images and corresponding sentences. Extensive experiments demonstrate that our proposed ScenaFuse outperforms previous competitive methods on three datasets.

Acknowledgments

This work is supported in part by funds from the National Key Research and Development Program of China (No. 2021YFF1201200), the National Natural Science Foundation of China (No. 62172187 and No. 62372209). Fausto Giunchiglia's work is funded by European Union's Horizon 2020 FET Proactive project (No. 823783).

Contribution Statement

Y.H.L. and M.Y.L. designed and developed the method and analysed the data. Y.H.L., M.Y.L., and X.Y.F. drafted the paper. D.L., F.G., X.M.L., and L.H. revised the paper. X.Y.F., and R.C.G. supervised the project and contributed to the conception of the project.

References

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *CVPR*, 2015.

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

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

[Conneau *et al.*, 2017a] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 2017.

[Conneau *et al.*, 2017b] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 2017.

[Dagan *et al.*, 2006] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *First PASCAL Machine Learning Challenges Workshop*, 2006.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[Falke *et al.*, 2019] Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *ACL*, 2019.

[Gibson *et al.*, 2019] Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407, 2019.

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

[Guan *et al.*, 2021] Renchu Guan, Yonghao Liu, Xiaoyue Feng, and Ximing Li. Paper-publication prediction with graph neural networks. In *CIKM*, 2021.

[Gururangan *et al.*, 2018] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *NAACL*, 2018.

[Haim *et al.*, 2006] R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

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

[Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[Li *et al.*, 2024] Mengyu Li, Yonghao Liu, Fausto Giunchiglia, Xiaoyue Feng, and Renchu Guan. Simple-sampling and hard-mixup with prototypes to rebalance contrastive learning for text classification. *arxiv preprint arXiv:2405.11524*, 2024.

[Liu *et al.*, 2019a] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *ACL*, 2019.

[Liu *et al.*, 2019b] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2021] Yonghao Liu, Renchu Guan, Fausto Giunchiglia, Yanchun Liang, and Xiaoyue Feng. Deep attention diffusion graph neural networks for text classification. In *EMNLP*, 2021.

[Liu *et al.*, 2022] Yonghao Liu, Mengyu Li, Ximing Li, Fausto Giunchiglia, Xiaoyue Feng, and Renchu Guan. Few-shot node classification on attributed networks with graph meta-learning. In *SIGIR*, 2022.

[Liu *et al.*, 2023a] Yonghao Liu, Mengyu Li Di Liang, Fausto Giunchiglia, Ximing Li, Sirui Wang, Wei Wu, Lan Huang, Xiaoyue Feng, and Renchu Guan. Local and global: Temporal question answering via information fusion. In *IJCAI*, 2023.

[Liu *et al.*, 2023b] Yonghao Liu, Di Liang, Fang Fang, Sirui Wang, Wei Wu, and Rui Jiang. Time-aware multiway adaptive fusion network for temporal knowledge graph question answering. In *ICASSP*, 2023.

[Liu *et al.*, 2024a] Yonghao Liu, Lan Huang, Bowen Cao, Ximing Li, Fausto Giunchiglia, Xiaoyue Feng, and Renchu Guan. A simple but effective approach for unsupervised few-shot graph classification. In *WWW*, 2024.

[Liu *et al.*, 2024b] Yonghao Liu, Lan Huang, Fausto Giunchiglia, Xiaoyue Feng, and Renchu Guan. Improved graph contrastive learning for short text classification. In *AAAI*, 2024.

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

[Mao *et al.*, 2015] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015.

[Moon *et al.*, 2018] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity recognition for short social media posts. In *NAACL*, 2018.

[Muennighoff *et al.*, 2022] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

[Murty *et al.*, 2021] Shikhar Murty, Tatsunori B Hashimoto, and Christopher D Manning. Dreca: A general task augmentation strategy for few-shot natural language inference. In *NAACL*, 2021.

[Nishihara *et al.*, 2020] Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. Supervised visual attention for multimodal neural machine translation. In *COLING*, 2020.

[Pan *et al.*, 2016] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[Pinker, 2003] Steven Pinker. *How the mind works*. Penguin UK, 2003.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Ramachandram and Taylor, 2017] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE SPM*, 34(6):96–108, 2017.

[Rocktäschel *et al.*, 2016] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. Reasoning about entailment with neural attention. In *ICLR*, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Storks *et al.*, 2019] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.

[Tay *et al.*, 2018a] Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. Co-stack residual affinity networks with multi-level attention refinement for matching text sequences. In *EMNLP*, 2018.

[Tay *et al.*, 2018b] Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference. In *EMNLP*, 2018.

[Tong *et al.*, 2020] Meihan Tong, Shuai Wang, Yixin Cao, Bin Xu, Juanzi Li, Lei Hou, and Tat-Seng Chua. Image enhanced event detection in news articles. In *AAAI*, 2020.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Vendrov *et al.*, 2016] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[Wang and Jiang, 2016] Shuohang Wang and Jing Jiang. Learning natural language inference with lstm. In *NAACL*, 2016.

[Williams *et al.*, 2018] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*, 2018.

[Xia *et al.*, 2021] Tingyu Xia, Yue Wang, Yuan Tian, and Yi Chang. Using prior knowledge to guide bert's attention in semantic textual matching tasks. In *WWW*, 2021.

[Zhang *et al.*, 2010] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *IJMLC*, 1:43–52, 2010.

[Zhang *et al.*, 2018] Kun Zhang, Guangyi Lv, Le Wu, Enhong Chen, Qi Liu, Han Wu, and Fangzhao Wu. Image-enhanced multi-level sentence representation net for natural language inference. In *ICDM*, 2018.

[Zhang *et al.*, 2019] Kun Zhang, Guangyi Lv, Le Wu, Enhong Chen, Qi Liu, Han Wu, Xing Xie, and Fangzhao Wu. Multilevel image-enhanced sentence representation net for natural language inference. *IEEE TSMC*, 51(6):3781–3795, 2019.

[Zhang *et al.*, 2020] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware bert for language understanding. In *AAAI*, 2020.