RCE-HIL: Recognizing Cross-media Entailment with Heterogeneous Interactive Learning

XIN HUANG, YUXIN PENG, and ZHANG WEN, Peking University, China

Entailment recognition is an important paradigm of reasoning that judges if a hypothesis can be inferred from given premises. However, previous efforts mainly concentrate on text-based reasoning as *recognizing textual entailment (RTE)*, where the hypotheses and premises are both textual. In fact, humans' reasoning process has the characteristic of *cross-media reasoning*. It is naturally based on the joint inference with different sensory organs, which represent complementary reasoning cues from unique perspectives as language, vision, and audition. How to realize cross-media reasoning has been a significant challenge to achieve the breakthrough for width and depth of entailment recognition. Therefore, this article extends RTE to a novel reasoning paradigm: ***recognizing cross-media entailment (RCE)***, and proposes ***heterogeneous interactive learning (HIL)*** approach. Specifically, HIL recognizes entailment relationships via cross-media joint inference, from image-text premises to text hypotheses. It is an end-to-end architecture with two parts: (1) Cross-media hybrid embedding is proposed to perform cross embedding of premises and hypotheses for generating their fine-grained representations. It aims to achieve the alignment of cross-media inference cues via image-text and text-text interactive attention. (2) Heterogeneous joint inference is proposed to construct a heterogeneous interaction tensor space and extract semantic features for entailment recognition. It aims to simultaneously capture the interaction between cross-media premises and hypotheses and distinguish their entailment relationships. Experimental results on widely used Stanford natural language inference (SNLI) dataset with image premises from Flickr30K dataset verify the effectiveness of HIL and the intrinsic inter-media complementarity in reasoning.

CCS Concepts: • **Information storage systems → Multimedia information systems**; • **Computing methodologies → Knowledge representation and reasoning**;

Additional Key Words and Phrases: Cross-media reasoning, cross-media hybrid embedding, heterogeneous joint inference, recognizing cross-media entailment

ACM Reference format:
Xin Huang, Yuxin Peng, and Zhang Wen. 2019. RCE-HIL: Recognizing Cross-media Entailment with Heterogeneous Interactive Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 1, Article 5 (February 2020), 21 pages.
https://doi.org/10.1145/3365003

5

This work was supported by the National Natural Science Foundation of China under Grant 61925201 and Grant 61771025.
Authors' addresses: X. Huang, Y. Peng (corresponding author), and Z. Wen, Peking University, Wangxuan Institute of Computer Technology, Beijing, 100871, China; email: pengyuxin@pku.edu.cn.

1551-6857/2019/02-ART5 $15.00
https://doi.org/10.1145/3365003

1 INTRODUCTION

Reasoning is a vital ability of human being. An important and basic reasoning topic is entailment recognition, which aims to determine if a hypothesis (H) can be inferred from given premises (P). This can be viewed as the logic relationship of "$P \rightarrow H$." In the widely used setting of research, there exist three possible results when P is true: H is definitely true ("entailment"), definitely false ("contradiction"), and neither of them ("neutral"). It plays a key role for various applications, such as semantic search, question answering, as well as knowledge extraction from big data.

The challenge of entailment reasoning is how to understand the semantic of P and H and analyze the interaction of cues within them. It usually needs background or common sense knowledge. Existing efforts mainly concentrate on text-based reasoning with textual premises and hypotheses. This is known as the extensively studied topic ***recognizing textual entailment (RTE)***, or natural language inference (NLI) [22]. For example, *"the woman is very happy"* can be entailed from *"a woman with a green headscarf, blue shirt, and a very big grin,"* along with the knowledge that *"grin"* means a *"happy"* face. The research of RTE has lasted for a long period. Recently, inspired by the great progress of deep neural network (DNN) in a wide range of applications, DNN-based methods have become the mainstream of RTE [2, 3, 6, 25, 40]. They follow the idea of relationship modeling between sentence pairs with DNN and achieve accuracy improvement.

However, it is not enough to recognize only text-based entailment relationship. Instead of solely text-based reasoning, humans naturally have the ability of inference from cues of different sensory organs, such as language, vision, and audition. Such **cross-media reasoning mechanism** is very common in people's daily lives. For example, it is very common to see an image with its textual caption in the same website, where we can infer information from the content of them by joint reasoning. Figure 1(b) shows one such case. The text premise says the Bird's Nest is in Beijing, but there is no cue about the Water Cube. In the image, we can recognize the Bird's Nest and the Water Cube, so we know they are very near to each other. Therefore, the Water Cube is in the same city Beijing with the Bird's Nest. As Figure 1(a) shows, only by text reasoning we cannot verify the hypothesis and draw the conclusion. As shown in the above example, cross-media reasoning is very common and important in the entailment recognition for humans. Nevertheless, it is ignored by the existing work, leading to the limitation of width and depth of reasoning.

Therefore, this article extends RTE to entailment recognition based on premises with different media types (i.e., ***recognizing cross-media entailment, RCE***). Compared with RTE, RCE is more consistent with human beings to cognize the world and has significant advantages as follows: On the one hand, different media types provide unique and complementary cues, which can significantly improve the flexibility and comprehensiveness of reasoning. For example, images have the natural advantage in describing scene and appearance, while text can effectively describe abstract logical phenomena such as disjunction, conditional, and negation. On the other hand, with the rapid development of computer and digital transition technology, the era of big data has witnessed the explosive growth of multimedia data. A great deal of knowledge and common sense information exists in the form of multimedia data (such as websites with image and text). Only text-based reasoning cannot meet the demand of understanding big data for semantic and knowledge extraction.

Although RCE is promising to improve the width and depth of reasoning, it is also a challenging problem: First, how to find cues contained in different media types relevant to the hypotheses? RCE needs to align semantics of cross-media information, so faces the challenge of "heterogeneity gap," which leads to the representation inconsistency of different media types. Second, how to combine these cues for judging entailment relationships? RCE is naturally a heterogeneous multi-premise reasoning problem. It needs to jointly represent the interaction from multiple premises to hypotheses and distinguish the entailment relationships. For addressing the above issues, this

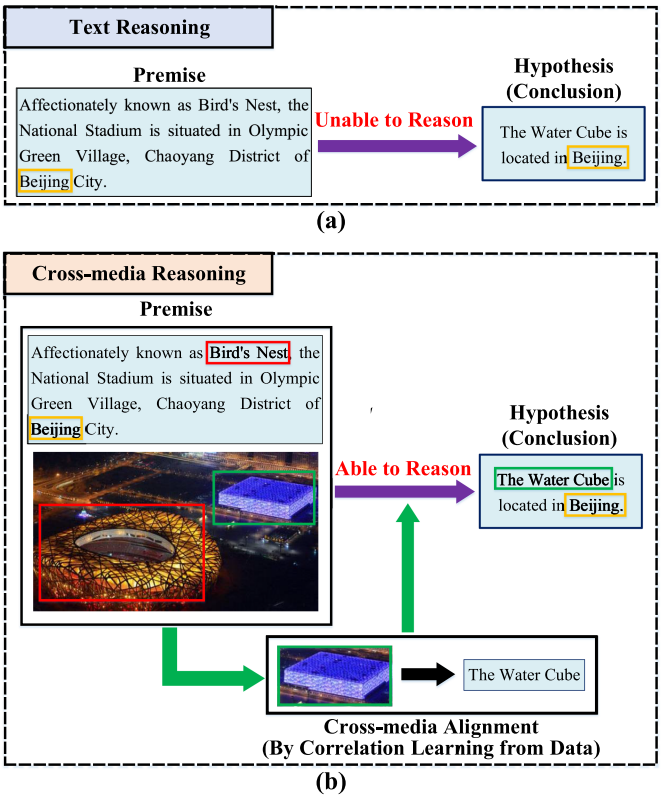


Fig. 1. Examples of text reasoning and cross-media reasoning. The boxes with the same color mean the correlated cues for reasoning. We can observe that this case needs to simultaneously realize the cross-media alignment and entailment inference. Note that such cross-media alignment can be based on correlation learning from data, where there exists information like image-text pairwise correlation.

article focuses on one representative and vital case of RCE—reasoning from image-text premises to text hypotheses—and proposes the ***heterogeneous interactive learning (HIL)*** approach. The main contributions of this article can be summarized as follows:

- ***Recognizing cross-media entailment (RCE)*** is proposed as a novel reasoning paradigm, where text and image premises are in equivalent positions to provide complementary reasoning cues. As the previous work mainly concentrates on text-based reasoning, this can improve the width and depth of reasoning in entailment recognition.
- ***Cross-media hybrid embedding*** is proposed to perform cross embedding of premises and hypotheses for generating their fine-grained representations. It aims to achieve the alignment of cross-media inference cues via image-text and text-text interactive attention.
- ***Heterogeneous joint inference*** is proposed to construct a heterogeneous interaction tensor space and extract semantic features with convolutional layers for entailment recognition. It can simultaneously reflect the interaction between cross-media premises and hypotheses and distinguish their entailment relationships.

For verifying the effectiveness of HIL, we adopt the textual premises and hypotheses from the widely used RTE dataset SNLI and their corresponding images from Flickr30K dataset to serve as the image premises. We not only compare HIL with 15 state-of-the-art RTE methods to show its

advantage of overall reasoning accuracies, but also show the experimental results of ablation study. Especially, we compare the results of text-based reasoning, image-based reasoning, and the collaboration of them, which show the intrinsic cross-media complementarity in reasoning process.

The remainder of this article is organized as follows: Section 2 presents a brief review of related work. Section 3 introduces our proposed HIL approach, which presents the architecture in detail. Section 4 presents the experiments and analyses, including comparison with state-of-the-art RTE methods, ablation study, and visualization results. Section 5 concludes this article, as well as presents the future work.

2 RELATED WORK

The central motivation of this article is based on RTE and cross-media analysis. So, in this section, we briefly review the above two aspects of related work.

2.1 Recognizing Textual Entailment

Basically, the goal of RTE is to build an intelligent system that can verify whether a textual hypothesis H stands based on a textual premise P, usually along with common sense knowledge. The paradigm of RTE can also be understood as the logic expression $P \rightarrow H$.

RTE is fundamental to a wide range of natural language processing (NLP) tasks such as question answering [12] and semantic search. Great efforts have been made to address RTE problem. One intuitive idea is to transform premises to hypotheses by reasoning rules [24], where the rules can include lexical containment relationships as "dog→animal" from WordNet [23] and causal relationships as "buy→own" [8]. However, because the rules are difficult to cover a wide range of reasoning phenomena, the current mainstream of RTE is to directly predict the entailment relationship with features extracted from premises and hypotheses.

Inspired by the successful application of DNN, the state-of-the-art RTE methods are mostly based on deep networks [2, 3, 6, 25, 40]. Bowman et al. [2] construct a large-scale dataset SNLI for RTE, which has been a widely used benchmark dataset to evaluate the performance of RTE methods. It has more than 570K sentence pairs, where each pair contains a premise and hypothesis. The work of Reference [2] also proposes to adopt two separate sentence recurrent models for extracting textual features, followed by several fully connected layers to predict the entailment relationship. Stack-augmented parser-interpreter neural network (SPINN) [3] is proposed as a single tree-sequence model for parsing and interpretation of sentences to perform entailment prediction. Bilateral multi-perspective matching (BiMPM) [40] views RTE as a text-to-text sequence analysis problem, which encodes the sentence pair in two directions $P \rightarrow H$ and $H \rightarrow P$ from multiple perspectives. Gong et al. [10] propose to construct an interaction tensor with word representations by self-attention of premises and hypotheses and then extract features by convolutional network to obtain the final entailment relationship.

However, the previous efforts of entailment relationship recognition are still limited to ***text-based reasoning***. Although some works like Reference [11] have attempted to incorporate visual information, the images are used for paraphrase acquisition to support text reasoning, instead of acting as reasoning premises. The joint inference of premises from different media types is ignored, resulting in the limitation of width and depth in reasoning. In this article, we aim to address the problem of ***cross-media reasoning*** from image-text premises to text hypotheses, thus improving the flexibility and comprehensiveness of entailment recognition.

2.2 Cross-media Analysis

To perform cross-media reasoning, a key challenge is how to establish the correlation between different media types. Cross-media analysis is the fundamental idea of a wide range of problems.

Here, we mainly introduce two representatives of them: cross-media retrieval and visual question answering (VQA).

Cross-media retrieval [30] aims to realize the information retrieval across different media types such as image, text, and audio. It is a fundamental problem in cross-media analysis, whose basic challenge is to bridge "heterogeneity gap" for measuring the cross-media semantic relevance. A mainstream and intuitive idea is called common representation learning, which represents data of different media types with the same "feature" type for distance metric. Existing methods can be divided as shallow learning methods [4, 13, 17, 20, 33] and DNN-based methods [1, 9, 16, 26, 29, 31, 41]. Shallow learning methods usually take linear projection to convert features of different media types to common representations with the same dimension. For example, as a classical and representative method, canonical correlation analysis (CCA) [13] learns common representation projections by maximizing pairwise cross-media correlation of cross-modal data.

Due to DNN's notable ability of learning non-linear correlations, DNN-based methods have been the current mainstream of cross-media retrieval. These methods construct DNN architectures to perform cross-media correlation learning and common representation generation. For example, Ngiam et al. propose Bimodal deep autoencoder [26], which consists of two deep autoencoders for the input of two media types, respectively. They share the same code layer to generate common representations. Cross-media multiple deep networks (CMDN) [29] are designed to jointly extract the intra-media and inter-media separate information, and then the common representation is obtained via hierarchical combination. For addressing the problem of insufficient cross-media training data, Huang et al. propose cross-media hybrid transfer network (CHTN) [16], which aims to extract knowledge from large-scale single-media dataset (e.g., ImageNet[19]) to promote the training performance and retrieval accuracy on cross-media data. Deep cross-media knowledge transfer [15] is proposed to transfer knowledge from a large-scale cross-media source domain to boost the retrieval accuracy on a small-scale cross-media target domain.

VQA is proposed to generate answers by given text questions based on visual content. The inputs of VQA include an image and a corresponding textual question, which is directly about the visual content, like the names, numbers, and colors of the objects. For example, one may ask a question such as "What is in the picture?" and expect the system to provide the names of objects in it. Co-attention model [21] is proposed to align the related image and text parts to guide the answer generation. The work of Reference [38] proposes to construct a knowledge base with RDF triplets and extract concepts from image to generate explainable question answering. Besides the introduced image-based QA, there are also some works focusing on video-based QA. For example, Xue at al. [42] attempt to address video-based QA with sequential video attention and temporal question attention model. The scenarios of VQA and RCE are different. VQA is an image-centric problem, where textual questions usually tell which areas in image/text to see. However, in RCE image and text, premises are equivalent to provide complementary cues for reasoning.

Cross-media analysis has become an active research area. However, the existing works have paid little attention to the challenge of entailment reasoning. The motivation of this article is to address the reasoning problem of RCE, which also aims to extend the concept width and depth of cross-media analysis.

3 HETEROGENEOUS INTERACTIVE LEARNING

Heterogeneous interactive learning (HIL) is proposed for recognizing entailment recognition from image-text premises to text hypotheses, which is a 3-pathway end-to-end architecture with two main parts, as shown in Figure 2: (1) **Cross-media hybrid embedding** generates fine-grained representations for premises and hypotheses via cross-media alignment of inference cues.

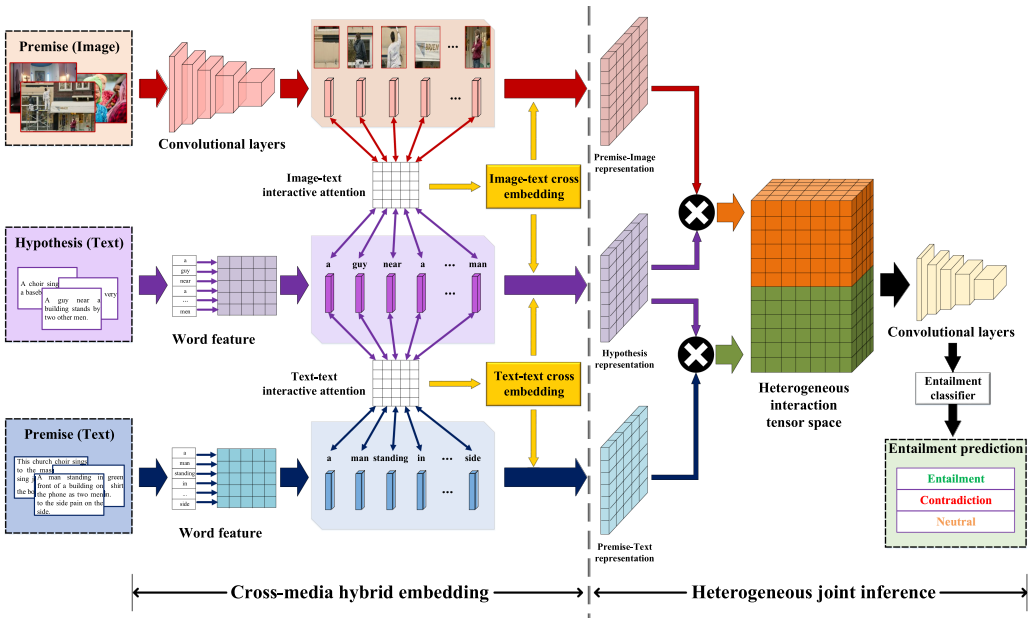


Fig. 2. The overview of our proposed Heterogeneous interactive learning (HIL) approach, which mainly consists of two parts: Cross-media hybrid embedding aims to generate fine-grained representations for premises and hypotheses via cross-media alignment of inference cues. Heterogeneous joint inference aims to obtain entailment prediction vectors via modeling the cue interaction between premises and hypotheses.

(2) **Heterogeneous joint inference** obtains entailment prediction vectors by analyzing the cue interaction between premises and hypotheses in a heterogeneous interaction tensor space.

Here, we give the formula definitions of the data and objective for HIL. The whole dataset is denoted as $Data = \{(s(I)_n, s(T)_n), h_n, e_n\}_{n=1}^{N+M+D}$, which includes three parts: N entries as the training set Tr, M entries as the testing set Te, and D entries as the development set Dev. $(s(I)_n, s(T)_n)$, and h_n form the nth premise-hypothesis pair with the entailment label e_n. In each pair, the premise includes one image $s(I)_n$ and one text $s(T)_n$, while the hypothesis is one text h_n. All e_n in Te are set to be unknown and are only used for evaluating the accuracy of entailment recognition. The aim of HIL is to predict the entailment relationship $\hat{e}_n$ for each premise-hypothesis pair in Te, which can be "entailment," "contradiction," and "neutral." The formula is as follows, where $P(\cdot)$ means the predicted possibility:

$$\hat{e}_n = \arg\max_{e} P(e|(s(I)_n, s(T)_n), h_n). \tag{1}$$

3.1 Cross-media Hybrid Embedding

HIL is a 3-pathway structure to simultaneously accept the input of image-text premises and text hypotheses. For convenience, we denote the three pathways in Figure 2 as Premise-Image, Hypothesis, and Premise-Text pathways, respectively, from top to bottom. The entailment relationship usually lies in the local information of image and text, such as the objects and their relations. Therefore, we get the fine-grained representations of input premises and hypotheses by first extracting their local features and then performing cross-media cue alignment with dual cross embedding.

3.1.1 Local Feature Extraction. For **Premise-Image pathway**, we adopt the convolutional neural network (CNN) as the local feature extractor. Specifically, we adopt VGG19 [36] as the basic model to generate feature maps, which naturally represent information of image regions. For an image $s(I)_n$, we can directly extract local features from pool5 layer as $\{s(I)_{n,1}, s(I)_{n,2}, \ldots, s(I)_{n,v}\}, s(I)_{n,i} \in R^{d_I}$, where v is the number of image regions and d_I denotes the feature dimension of each region. For **Hypothesis** and **Premise-Text pathways**, they share the same network structure. Taking Premise-Text pathway as an example, we follow Reference [10] to extract the feature of each word in $s(T)_n$ as $\{s(T)_{n,1}, s(T)_{n,2}, \ldots, s(T)_{n,w}\}, s(T)_{n,i} \in R^{d_T}$, where w is the number of words, and d_T denotes the dimension of word feature. Similarly, in Hypothesis pathway, we can obtain the local feature of h_n as $\{h_{n,1}, h_{n,2}, \ldots, h_{n,w}\}, h_{n,i} \in R^{d_T}$. For the balance of image and text premises in the subsequent stage of interaction tensor space construction, we set $w = v$, which will be discussed in Section 3.3 with details of implementation and optimization.

The feature representations of image and text are inconsistent, so they are represented in different spaces with different feature dimensions. This results in the problem that image and text reasoning cues cannot be directly aligned. Therefore, the image local features are projected from image space to text space, which adopts the idea of cross-media common representation learning [30]. This is achieved via a fully connected layer as:

$$\hat{s}(I)_{n,i} = \phi_I(s(I)_{n,i}; \theta_I) \in R^{d_T}, \tag{2}$$

where θ contains the parameter of network, $\hat{s}(I)_{n,i}$ denotes the projected image local feature, which is converted as the same dimension with text feature space. In the remainder of this article, $\hat{s}(I)_n = \{\hat{s}(I)_{n,1}, \hat{s}(I)_{n,2}, \ldots, \hat{s}(I)_{n,v}\}$, $\hat{s}(I)_{n,i} \in R^{d_T}$ is called "image local feature," which is corresponding to text local feature for convenience. The common representation projection of image is denoted as $\phi_I(x; \theta_I)$, which is achieved via a fully connected layer with d_T output dimension. For convenience, the similar notation $\phi(x; \theta)$ is used for all fully connected layers in this article, which can be optimized during the network training.

3.1.2 Dual Cross Embedding. Now that we have obtained the local features of cross-media premises and hypotheses, we need to establish the relevances between them. For example, assuming that the hypothesis contains the word "car," we expect the network to align it simultaneously to words like "vehicle" in text premise, as well as the regions of cars in the image premise. In this way, the trained network can find relevant cues from premises according to the hypotheses.

As shown in Figure 2, this is achieved by a symmetric structure with pairwise cross embedding of three pathways, based on image-text and text-text interactive attention learning. Attention mechanisms have been adopted in the existing methods of RTE as Reference [43], but they usually focus on only text-text relevance, i.e., identifying "what to read" in text. In this article, we propose to jointly consider "what to read" and "where to look" problems regarding cues from premises.

For the sake of brevity and readability, here, we take the cross embedding between Premise-Image and Hypothesis pathways as an example to describe. The basic idea is to encode $\hat{s}(I)_n$ and h_n with the relevant context from each other. For each $\hat{s}(I)_{n,i}$, we compute the relevance of each word in $\{h_{n,j}\}_{j=1}^{w}$ with the guidance of image-text attention for embedding and vice versa. Specifically, given the local features of $\hat{s}(I)_n$ and h_n, we define the image-text interactive attention matrix E, whose each element e_{ij} is calculated as:

$$e_{ij} = \phi_A(\hat{s}(I)_{n,i} \circ h_{n,j}; \theta_A), \tag{3}$$

where $\circ$ denotes element-wise product. We adopt a fully connected layer to implement $\phi_A(x; \theta_A)$ with one output dimension. Intuitively, The value of e_{ij} represents the relevance between the image

region $\hat{s}(I)_{n,i}$ and text word $h_{n,j}$. We get the cross embedding vector of $\hat{s}(I)_{n,i}$ as:

$$c_{\hat{s}(I)_{n,i}} = \sum_{j=1}^{w} \frac{exp(e_{ij})}{\sum_{k=1}^{w} exp\,(e_{ik})} h_{n,j}. \tag{4}$$

Conversely, we obtain the cross embedding vector of $h_{n,j}$ as:

$$c_{h_{n,j}} = \sum_{i=1}^{v} \frac{exp(e_{ij})}{\sum_{k=1}^{v} exp(e_{kj})} \hat{s}(I)_{n,i}. \tag{5}$$

The above cross embedding processing is also simultaneously performed between Hypothesis and Premise-Text pathways, which forms a dual cross embedding mechanism. The Hypothesis pathway serves as a bridge to not only jointly find reasoning cues from both image and text premises, but also allow the alignment information to be shared between Premise-Text and Premise-Image pathways to mutually boost. The *cross embedding vectors* and *local features* are both used to generate representations for each pathway, which will be introduced in Section 3.2.

3.2 Heterogeneous Joint Inference

In the part of cross-media hybrid embedding, we have obtained the fine-grained alignment of premises and hypotheses. However, it remains a problem how to obtain the final prediction of entailment relationship. The part of heterogeneous joint inference contains two main steps to achieve this goal. The first is to construct a heterogeneous interaction tensor space, which can simultaneously reflect the cue interaction between cross-media premises and hypotheses. The second is to extract semantic features from the constructed interaction tensor space by a CNN network and then generate the final prediction vectors.

The construction of interaction tensor space is inspired by Reference [10]. Different from Reference [10], which only considers text-text interaction, this article proposes to construct a heterogeneous interaction tensor space to exploit the cross-media collaboration of image-based and text-based reasoning. The interaction tensor space is constructed by image-text and text-text interaction tensors via a symmetric 3-pathway structure.

Figure 3 shows how to construct the heterogeneous interaction tensor space. Taking the processing between Premise-Image and Hypothesis-Text pathways as an example, we first compute the representations of $\hat{s}(I)_n$ and h_n as matrices $R_{\hat{s}(I)_n} \in R^{d_R \times v}$ and $R_{h_n} \in R^{d_R \times w}$, respectively. The columns of $R_{\hat{s}(I)_n}$ and R_{h_n} can be computed as follows:

$$r_{\hat{s}(I)_{n,i}} = \phi_{RI}([\hat{s}(I)_{n,i}; c_{\hat{s}(I)_{n,i}}; \hat{s}(I)_{n,i} \circ c_{\hat{s}(I)_{n,i}}]; \theta_{RI}) \in R^{d_R}, \tag{6}$$

$$r_{h_{n,i}} = \phi_{RH}([h_{n,i}; c_{h_{n,i}}; h_{n,i} \circ c_{h_{n,i}}]; \theta_{RH}) \in R^{d_R}, \tag{7}$$

where $[;]$ denotes the vector concatenation, which includes the local features, the cross embedding vectors, and their combination. For each of the two pathways, we adopt a fully connected layer as $\phi_{RI}(x; \theta_{RI})$ and $\phi_{RH}(x; \theta_{RH})$, with the output dimension of d_R. Note that we have set $w = v$, so we can build the image-text interaction tensor as follows:

$$\begin{aligned} Tensor(IT)_{ij} &= r_{\hat{s}(I)_{n,i}} \circ r_{h_{n,j}} \in R^{d_R}, \\ 1 \le i \le w, 1 \le j \le w. \end{aligned} \tag{8}$$

For each pair of i and j, there will be a d_R-d vector $r_{\hat{s}(I)_{n,i}} \circ r_{h_{n,j}}$. So the above image-text interaction tensor can be viewed as a 3-d space. Similarly, we can build the text-text interaction tensor as

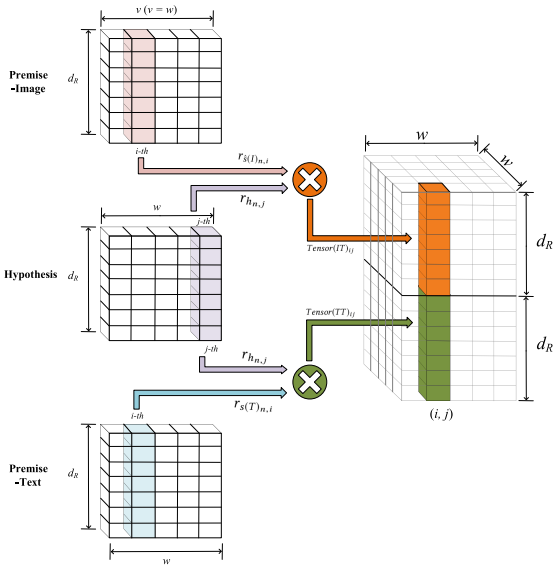


Fig. 3. Construction of the heterogeneous interaction tensor space. For demonstration purposes, we take jth hypothesis word with both ith premise image region and text word as an example, which forms (i, j)-th column in the tensor space.

follows:

$$Tensor(TT)_{ij} = r_{s(T)_{n,i}} \circ r_{h_{n,j}} \in R^{d_R}, \tag{9}$$

$$1 \le i \le w, 1 \le j \le w. \tag{9}$$

Then the heterogeneous interaction tensor space can be built as the concatenation of $Tensor(IT)$ and $Tensor(TT)$:

$$Tensor(HT) = [Tensor(IT); Tensor(TT)] \in R^{2d_R \times w \times w}. \tag{10}$$

$Tensor(HT)$ can jointly represent the fine-grained cue interaction between image-text premises and text hypotheses, where the cross-media entailment relationships can be fully captured.

After obtaining $Tensor(HT)$, we get the final entailment predictions through a CNN network with an entailment classifier. Specifically, we feed $Tensor(HT)$ into the CNN model DenseNet [14] with the channel number of $2d_R$ and extract the flattened feature maps from the last transition layer of DenseNet, which will be further fed into a following 3-way softmax classifier to predict $\hat{e}_n$. The entailment relationship loss is as follows:

$$Loss_{ER} = \sum_{n=1}^{N} f_{sm}(\hat{e}_n, e_n; \theta_{sm}), \tag{11}$$

where f_{sm} denotes the softmax function, $\hat{e}_n$ denotes the predicted entailment relationship, while e_n denotes the true entailment relationship.

Note that *cross-media hybrid embedding* and *heterogeneous joint inference* form an end-to-end architecture, which can be seen as a whole learning process of multimodal representation and entailment relationship. They can mutually boost and concentrate on the final goal of entailment recognition to achieve better reasoning accuracy.

3.3 Details of Implementation and Optimization

We give the details of our implementation based on TensorFlow.[1] In **cross-media hybrid embedding**, we first resize each image to 224×224 and adopt VGG19 [36] pre-trained with ImageNet dataset as the basic model and extract the feature maps of pool5 with the size of $7 \times 7 \times 512$. We treat them as descriptors of totally 7×7 regions with 512-d representations, so $v = 49, d_I = 512$. For text, we follow Reference [10] to generate 448-d word features so $d_T = 448$, and set $w = v = 49$. The 448-d word feature includes 300-d word embedding, 100-d char feature, 47-d speech tagging, and 1-d binary exact match feature. For Equation (2), there is a fully connected layer with 448 output units as ϕ_I, which converts 512-d vectors $s(I)_{n,i}$ to 448-d vectors $\hat{s}(I)_{n,i}$. For Equation (3), we adopt a fully connected layer with 1 output unit as ϕ_A, which converts 448-d vectors $\hat{s}(I)_{n,i} \circ h_{n,j}$ to real values.

In **heterogeneous joint inference**, d_R is set to be 300, which is further discussed in Section 4. Therefore, for Equations (6) and (7), we adopt a fully connected layer with 300 output units as ϕ_{RI} and ϕ_{RH}, respectively. The use of DenseNet is following Reference [10] without pre-training. HIL can be optimized by standard stochastic gradient descent (SGD). To avoid overfitting, L2 regularization is utilized to constrain network weights, and dropout layers are applied before all linear layers. The initial learning rate is set as 0.5 with a learning rate decay of 3e-4 after 30K iterations. The batch size is set to be 54 because of the memory limitation, and the model achieving the highest recognition accuracy on *Dev* will be recorded as the final model for testing.

Note that the constraint $w = v$ is necessary to achieve the balance of image and text premises. If $v \neq w$, then the sizes of $Tensor(IT)$ and $Tensor(TT)$ will be inconsistent, and the larger one will have more effect in the prediction of entailment relationship. The value of $w = 49$ is according to the feature maps from VGG19 pool5 as descriptors of naturally $7 \times 7 = 49$ regions. We follow previous works such as Reference [2] to set a fixed text length and complete the shorter ones with zero-padding strategy. In this strategy, w should be large enough to cover almost all sentences corresponding to images, while a small w may lead to information loss for long sentences. Because of the interactive attention, the parts of zero-padding can be filtered and will not bring negative effect.

4 EXPERIMENT

This section presents the experiments for verifying the effectiveness of HIL. The details of dataset and evaluation task will first be introduced, along with the compared state-of-the-art methods. Then, the results of HIL and compared methods will be presented. Furthermore, we will show the ablation study experiments, parameter analyses, as well as the visualized results.

4.1 Dataset

As introduced above, HIL is proposed to perform RCE from image-text premises to text hypotheses. However, there is no existing dataset for such a challenging problem, so we need to construct a dataset for evaluation. Fortunately, the widely used and large-scale RTE dataset ***SNLI*** [2] is constructed from the captions in Flickr30K corpus [32]. Therefore, we can incorporate the corresponding images as premises and establish the dataset that we call ***SNLI-RCE***.

We first introduce the SNLI dataset. It consists of totally about 570K textual premise-hypothesis pairs. These pairs are divided into three parts: a training set of 549,367 pairs, a development set of 9,842 pairs, and a testing set of 9,824 pairs. All the premises are directly from the textual captions of images in Flickr30K corpus (except for only about 4K from VisualGenome corpus [18]), and the hypotheses are newly written by the constructor of SNLI dataset [18]. Every pair is manually labeled with one of the three entailment relationships: "entailment," "contradiction," and "neutral."

[1] www.tensorflow.org.

Premise	Hypothesis	Label
 People of all ages stroll carelessly through aisles and aisles of kiosks , browsing the variety of random merchandise from vendors at a large flea market.	There are many people shopping.	Entailment
 The Irish setter with the safety vest is running ahead of the Rottwieler and the Dalmation.	Three dogs are running.	Entailment
 A woman is dressed up like playing cards for some type of parade.	A woman is dressed up like a duck.	Contradiction
 A track event held by J. P. Morgan Chase with security.	Security guards are searching for weapons at a track event.	Neutral

Fig. 4. Examples of the image-text premises and text hypotheses in SNLI-RCE dataset.

SNLI dataset itself provides a large-scale corpus of textual premise-hypothesis pairs, but we also need corresponding image premises. Because the premises are mostly directly from the textual caption of images in Flickr30K dataset, we can adopt the corresponding image for each $P(text) \rightarrow H(text)$ pair to form cross-media pairs as $P(image - text) \rightarrow H(text)$. Now, we have a large-scale dataset for RCE, where each premise contains a pair of corresponding image and text. This is similar to the common scenario where we can see images with their titles in the same website. We call the constructed dataset as ***SNLI-RCE***, and some of the examples are shown in Figure 4. From the examples, we can see that the image and text are generally consistent for semantic expression. However, the information from image and text is not equal. Taking the first premise-hypothesis pair for example, it is not easy to identify that the people are "shopping" from the image premise, while from text we cannot clearly know whether there are "many" people.

Note that for the pairs from VisualGenome corpus, for convenience, we omit them because they are very minor (only about 4K) compared with the whole dataset (about 570K) and only appear in the training data. Except for the above modification, we adopt exactly ***the original SNLI dataset***. The entailment relationships in SNLI dataset *will not be changed by incorporating the image premise.* This is because of the construction of SNLI dataset: The text premises are just from the image captions, and the ambiguous premise-hypothesis pairs are removed to make the final SNLI corpus. This makes the entailment relationships also applicable for the images and the combination of image and text. The advantage of such case is that it establishes a unified standard to perform objective comparison with the state-of-the-art RTE method, as well as comparison among text-based, image-based, and cross-media joint entailment recognition.

4.2 Evaluation Metric and Compared Methods

We conduct the task of entailment recognition to verify the effectiveness of HIL. The aim is to judge each pair of premise and hypothesis to be "entailment," "contradiction," and "neutral." The

evaluation metric is recognition accuracy, which intuitively is the ratio of correctly recognized pairs.

Because there is no previous work involving RCE, we adopt the state-of-the-art methods of RTE for comparison, which only use the text premises and hypotheses. Note that in SNLI-RCE dataset, the text premises, text hypotheses, and entailment relationship labels are exactly the same with SNLI dataset. Therefore, the comparison with existing RTE methods can objectively show the effectiveness of HIL and the benefit gained from cross-media collaborative reasoning.

The compared methods include KIM [5], 300D CAFE [37], DIIN [10], ESIM [6], btree-LSTM [27], Re-read LSTM [34], NTI-SLSTM-LSTM [44], BiMPM [40], Decomposable attention model [28], LSTMN [7], mLSTM [39], DiSAN [35], SPINN-PI [3], Tree-based CNN encoders [25], and 100-d LSTM encoders [2]. Brief introductions of the above compared methods are presented below:

- **KIM** [5] uses WordNet 3.0 [23] to provide external knowledge of the relationships of words, which can guide the encoding of premises and hypotheses by attention learning.
- **300D CAFE** [37] proposes to construct a compare-propagate architecture to propagate alignment vectors for representation learning and adopts novel factorization layers for compression of alignment vectors.
- **DIIN** [10] constructs an interaction tensor with word representations by interaction attention between the premises and hypotheses and obtains the final reasoning results by convolutional networks.
- **ESIM** [6] builds a chain LSTM-based sequential inference model, which explicitly encodes parsing information with recursive networks for local inference modeling and inference composition.
- **btree-LSTM** [27] first uses a btree-LSTM for premises and hypotheses, respectively, then models the interaction between the two sequential with attention and finally obtains the recognition results by a learned tree.
- **Re-read LSTM** [34] proposes to adopt the intensive reading mechanic to re-read the sentences with the sentence memory for obtaining a better understanding of sentence pairs.
- **NTI-SLSTM-LSTM** [44] is a syntactic parsing-independent tree structured model. A middle ground is generated between the sequential RNNs and syntactic tree-based recursive models.
- **BiMPM** [40] views RTE as a sequence matching problem, which encodes the sentence pairs in two directions $P \rightarrow H$ and $H \rightarrow P$ from multiple perspectives.
- **Decomposable attention model** [28] uses attention mechanism to decompose the entailment recognition problem into subproblems that can be solved in parallel.
- **LSTMN** [7] constructs a simulator of machine reading to process text from left to right incrementally and uses memory and attention mechanism for shallow reasoning.
- **mLSTM** [39] uses an LSTM model for matching each word in the hypothesis with an attention-weighted representation of the premise, which emphasizes the critical matching results for final prediction.
- **DiSAN** [35] proposes a lightweight neural net for attention mechanism, where the attention between elements from input sequence is directional and multi-dimensional to learn sentence embedding.
- **SPINN-PI** [3] proposes a single tree-sequence model for parsing and interpretation of sentence to perform entailment prediction.
- **Tree-based CNN encoders** [25] model each individual sentence with a tree-based CNN model shared by premises and hypotheses and then adopt a sentence matching layer for information aggregation.

Table 1. Experimental Results Compared with State-of-the-art Methods, under **single-model** Setting

Method	Accuracy (%)
Our HIL	**90.3**
KIM [5]	88.6
300D CAFE [37]	88.5
DIIN [10]	88.0
ESIM [6]	88.0
btree-LSTM [27]	87.6
Re-read LSTM [34]	87.5
NTI-SLSTM-LSTM [44]	87.3
BiMPM [40]	86.9
Decomposable attention model [28]	86.8
LSTMN [7]	86.3
mLSTM [39]	86.1
DiSAN [35]	85.6
SPINN-PI [3]	83.2
Tree-based CNN encoders [25]	82.1
100-d LSTM encoders [2]	77.6

- **100-d LSTM encoders** [2] adopt two separate sentence recurrent models for extracting textual features, followed by several fully connected layers to predict the entailment relationship.

Some state-of-the-art methods also report results under the *ensemble setting*, which means to combine the results from several individual models. On the contrary, *single-model setting* means the result is from an individual model. We conduct experiments under both single-model and ensemble settings. Specifically, the prediction probabilities of 10 individual HIL models are fused for ensemble setting, with different network initializations such as in Reference [5].

4.3 Comparison with State-of-the-art Methods

Table 1 shows the accuracies of entailment recognition, under *single-model* setting. We can see that the performance of 100-d LSTM encoders is relatively low, because the simple architecture cannot fully capture the semantic and interaction of sentences: only one LSTM encoder for each sentence and several fully connected layers to generate prediction vectors. Methods such as DiSAN and BiMPM adopt more complex attention and sequential models, which obtain accuracy improvement. For exploiting more common sense information, KIM utilizes a large-scale external knowledge base WordNet 3.0 [23] to indicate the word relationships, which is a key factor for it to outperform other compared methods. The above results show the importance of effectively exploiting information within text (e.g., by attention and sequential models) and the guidance of external source of common sense. However, HIL achieves the highest accuracy of 90.3% among the compared methods. This shows that the incorporation of multiple media types as premises can provide rich and complementary cues to improve the accuracy.

Besides, as shown in Table 2 of *ensemble setting*, HIL also achieves the highest accuracy of 90.8%. The above advantages not only come from the complementary information of different media types in premises, but also the effective exploitation of cross-media collaboration by HIL: Cross-media hybrid embedding can achieve fine-grained alignment of cross-media inference cues for exploiting

Table 2. Experimental Results Compared with State-of-the-art Methods, under **Ensemble** Setting

Method	Accuracy (%)
Our HIL (Ensemble)	**90.8**
300D CAFE (Ensemble) [37]	89.3
DIIN (Ensemble) [10]	88.9
BiMPM (Ensemble) [40]	88.8
ESIM+Syntactic TreeLSTM [6]	88.6

Table 3. Experimental Results of the Ablation Study for HIL

Method	Accuracy (%)
HIL	**90.3**
HIL (Text Premise)	88.4
HIL (Image Premise)	86.4
HIL (Late Fusion)	89.0
HIL (w/o DCE)	87.9
HIL (w/o HIT)	86.9

complementary information from image and text; heterogeneous joint inference can fully exploit the interaction between cross-media premises and hypotheses and improve the accuracy of entailment recognition.

4.4 Experiments of Ablation Study

HIL consists of several network pathways and components. Therefore, we conduct the ablation study to show the impacts of them on the accuracy. We first show the accuracies by only text or image premises, then present the performance with late fusion of them and discuss the impacts of dual cross embedding (abbreviated as DCE) as well as heterogeneous interaction tensor (abbreviated as HIT). The experimental results of the ablation study are shown as Table 3 under *single-model* setting, which are discussed as follows:

- **HIL (Text Premise)**: This aims to show the performance that is obtained by only text premises, i.e., $text \rightarrow text$ entailment recognition. We remove the Premise-Image pathway. Accordingly, there are also no parts of image-text interactive-attention, image-text cross-embedding, and $Tensor(IT)$ in Equation (8). By comparing HIL (Text Premise) with HIL, we can see that the incorporation of image premise can obtain an accuracy improvement of **1.9%**. It shows the effectiveness of cross-media collaboration and our proposed HIL approach.
- **HIL (Image Premise)**: Contrary to HIL (Text Premise), HIL (Image Premise) means that we only use image premises. We can see that HIL (Image Premise) obtains a little bit lower accuracy than HIL (Text Premise), which is reasonable, because $image \rightarrow text$ entailment recognition is more challenging due to the "heterogeneity gap." However, HIL (Image Premise) still achieves the accuracy of 86.4%, and it even outperforms some text-based methods. It can be observed that the visual information can be effectively captured for entailment recognition by HIL.

- **HIL (Late Fusion)**: This means the late fusion of predicted possibilities by HIL (Text Premise) and HIL (Image Premise), and we can see that HIL (Late Fusion) fails to obtain a considerable advantage. This occurs because text-based reasoning and image-based reasoning have unique advantages for different cases, and straightforward combination strategies cannot fulfill their complementarity. On the contrary, HIL can achieve an inspiring improvement than HIL (Image Premise) and HIL (Text Premise), which shows the proposed cross-media hybrid embedding and heterogeneous joint inference can effectively fulfill the complementarity of image-based and text-based reasoning.
- **HIL (w/o DCE)**: Dual cross embedding is the key component in the part of cross-media hybrid embedding, which can reflect the main idea of cross-media interaction. Here, we aim to separately verify its effectiveness. That is to say, there are no parts of image-text and text-text interaction attention, as well as Image-text and Text-text cross embedding. Accordingly, the tensors are constructed directly with the local features of each pathway. Comparing with HIL, we can see an accuracy drop of **2.4%**, which shows that dual cross embedding is important for accuracy improvement.
- **HIL (w/o HIT)**: Heterogeneous interaction tensor is the key component in the part of heterogeneous joint inference, which simultaneously considers image and text cues. For verifying its effectiveness, we remove the part of heterogeneous interaction tensor and corresponding convolutional layers. Specifically, we follow Reference [6] for the d_R-dimensional $r_{\hat{s}(I)_{n,i}}$, $r_{h_{n,i}}$, and $r_{s(T)_{n,i}}$ to convert them as a fixed-length vector with average and max pooling strategy. Taking $r_{\hat{s}(I)_{n,i}}$ as an example, we have:

$$R(ave)_I = \sum_{i=1}^{v} \frac{r_{\hat{s}(I)_{n,i}}}{v}, \quad R(max)_I = \max_{1 \le i \le v} r_{\hat{s}(I)_{n,i}}. \tag{12}$$

  Similarly, we have $R(ave)_H$, $R(max)_H$, $R(ave)_T$, and $R(max)_T$. So, the vector to feed into 3-way classifier is defined as $[R(ave)_I; R(max)_I; R(ave)_H; R(max)_H; R(ave)_T; R(max)_T]$. Comparing HIL (w/o HIT) with HIL, we can observe a clear drop of **3.4%** in accuracy. This shows that the interaction tensor has the considerable ability to represent cross-media cue interaction, so the accuracy can be improved.

4.5 Example Analysis

To intuitively show the effectiveness of cross-media joint reasoning by HIL, we show some concrete examples such as Figure 5. These examples are all wrongly recognized without consideration of image information, i.e., by HIL (Text Premise), but correctly recognized by HIL. Although humans can analyze the entailment relationships only from text, these cases are difficult for computers, because they are quite indirect. For example, in the first shown case, the text premise "A small, pale bird bends down to examine a crumb" has no clear indication of place to judge the assertion "in the air," unless we know that a bird cannot bend down while flying and a crumb is unlikely to be in the air. However, the image has clear cue of scene: It is not the air. So the entailment relationship is "Contradiction."

Although the entailment relationship can be analyzed from text or image premises, the experimental evaluation aims to cover a common reasoning phenomenon in our daily life: The cue contained in one media type can be quite indirect for a computer to perform reasoning, while another media type can provide complementary and clear cues. We can also see HIL achieves better effective accuracy improvement than HIL (Text Premise) and HIL (Image Premise). This indicates

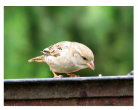
Premise	Hypothesis	True label	Predicted label
 A small, pale bird bends down to examine a crumb.	The bird is in the air.	Contradiction	HIL: Contradiction HIL (Text Premise): Neutral
 A demolition crane tearing down what is left of a building with a man standing on a platform.	The building is being painted.	Contradiction	HIL: Contradiction HIL (Text Premise): Entailment
 A scientist studies a slide in order to work on her new creation.	A woman is reading.	Contradiction	HIL: Contradiction HIL (Text Premise): Entailment
 An older balding man with glasses wearing a pink and white shirt with a sweater vest is playing a musical instrument while looking at the sheet music.	The man has little hair.	Entailment	HIL: Entailment HIL (Text Premise): Contradiction

Fig. 5. Some results that are correctly recognized by HIL, but wrongly recognized by HIL (Text Premise).

that text and image premises in the dataset actually provide complementary information, which verifies the motivation of cross-media joint reasoning.

4.6 Visualization of Cross-media Alignment

For performing RCE, it is very important to find supporting cues for a decision on entailment relationships. That is to say, we need to align the hypothesis word to both words and image regions in premises. In HIL, the attention matrices are the key components to align the text and image cues. For intuitive description, we have provided examples of alignment such as Figure 6. There are two kinds of interactive attention in HIL: **text-text** and **image-text** attention, which compute the alignment between each word in hypotheses and each text word or image region in premises, respectively.

For demonstration purposes, we select one keyword in each hypothesis and show its interactive attention to text and image premises. For instance, in the first example, the word "woman" and image region of a woman in premise are simultaneously aligned with "lady" in the hypothesis, with high attention weight. We can observe that our HIL can align important parts in hypotheses with both image and text premises to achieve effective accuracy improvement of cross-media collaborative reasoning.

4.7 Failure Cases

We also show some failure cases of HIL approach, as well as some visualization results of interactive attention, such as Figure 7. We can see that these failure cases are mainly because the hypotheses need imagination or contain ambiguous semantics. In the first example, although the label is "entailment," there are no exact cues to indicate what the firefighters are "ready" to do. In the second example, it is hard to clearly define what "vigorously" means. In such cases, the entailment recognition will be challenging to obtain the true labels. From the visualization of interactive attention, it can be seen that such indirect cues are difficult to align to meaningful image regions and specific words.

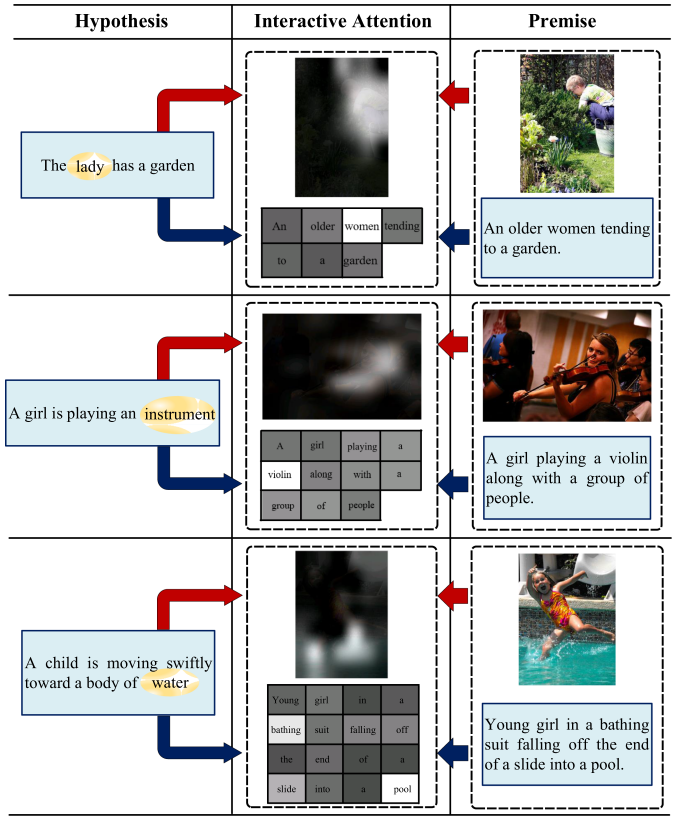


Fig. 6. Examples of cross-media alignment. The highlighted words in hypotheses are selected keywords to be aligned, and the brightness of image regions and text words indicates the attention weights. High brightness means high attention weights to align.

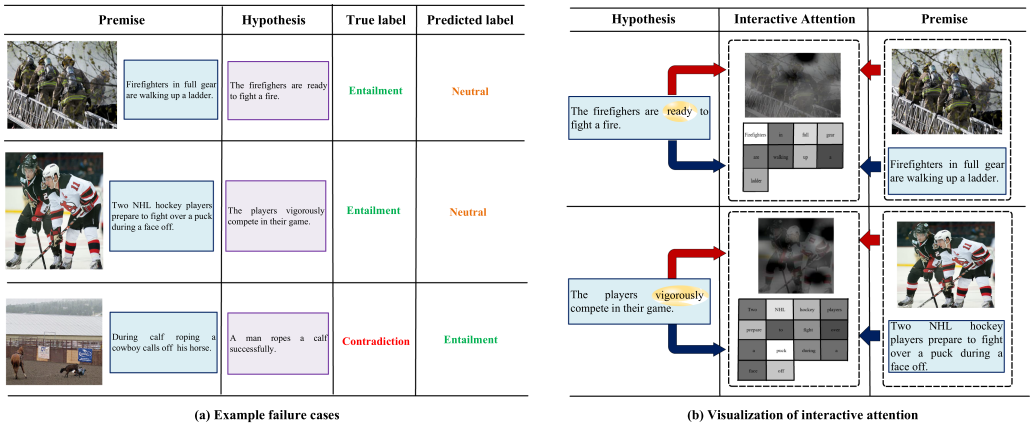


Fig. 7. Some failure cases of HIL, as well as the visualization of interactive attention.

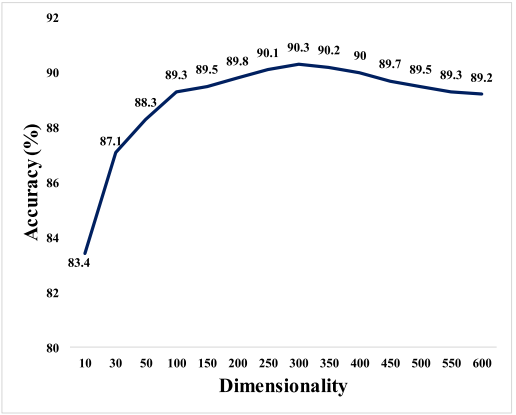


Fig. 8. Impact of d_R, which determines the size of heterogeneous interaction tensor.

Table 4. Experimental Results with Different Values of w for HIL

w value	Accuracy (%)
49	90.3
36	89.5
16	88.4
9	83.2

4.8 Impact of Parameters

4.8.1 Impact of d_R. d_R is an important parameter in HIL, which determines the size of heterogeneous interaction tensor. d_R is actually the unit number of a fully connected layer in each pathway, which outputs $r_{\hat{s}(I)_{n,i}}$, $r_{h_{n,i}}$, and $r_{s(T)_{n,i}}$, respectively. By adjusting the value of d_R, we have the following observations from the result of Figure 8: If d_R is very small, then the representation ability of the constructed tensor is not enough, which results in a low accuracy. As d_R increases, the accuracy will be improved until it reaches a peak when d_R is 300. However, a higher d_R does not bring improvement. This may occur because an unnecessarily large d_R results in uncertainty of training.

4.8.2 Impact of w. We follow previous works such as Reference [2] to set a fixed w, and complete the shorter ones with zero-padding strategy. w should be large enough to cover almost all sentences corresponding to images, while a small w may lead to information loss for long sentences. We find that $w = 49$ is safe to cover 99.9% text, except very few long sentences for compromise. An experiment is conducted with different w values; Table 4 shows that a small w leads to accuracy drop.

4.8.3 Impact of Learning Rate. Figure 9 shows the accuracies with different initial learning rate. From the result of Figure 9, we can observe that the highest accuracy is obtained at the learning rate of 0.5. Very small or large learning rates are both not optimal.

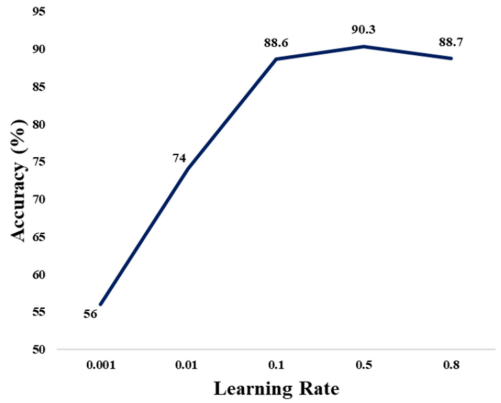


Fig. 9. Impact of different learning rates on accuracy for HIL.

5 CONCLUSION

This article has extended RTE to ***recognizing cross-media entailment (RCE)*** and proposed heterogeneous interactive learning (HIL) approach for reasoning from image-text premises to text hypotheses, which is an end-to-end architecture with two parts: ***Cross-media hybrid embedding*** performs cross embedding via interactive attention, which achieves fine-grained alignment of cross-media inference cues. ***Heterogeneous joint inference*** constructs a heterogeneous interaction tensor space and extracts semantic features to obtain prediction from it, which captures cross-media cue interaction for entailment recognition. To the best of our knowledge, HIL is the first work for entailment recognition under cross-media settings, whose effectiveness is verified on SNLI dataset with image premises from Flickr30K dataset. The experimental results also show the complementarity of image and text in entailment recognition, which indicates the promising potential of cross-media collaboration in reasoning.

The future work lies in mainly two aspects: First, it is important to achieve flexibility of media types in reasoning, so we will incorporate more media types for extending the scope of entailment recognition, such as audio and video. Second, we will also utilize existing knowledge bases to provide external information, including the rich sources of images and text on the Internet.

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2010. Deep canonical correlation analysis. In *Proceedings of the International Conference on Machine Learning (ICML'10)*. 3408–3415.

[2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP'15)*. 632–642.

[3] Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL'16)*. 1466–1477.

[4] Herng-Yow Chen and Sheng-Wei Li. 2007. Exploring many-to-one speech-to-text correlation for web-based language learning. *ACM Trans. Multim. Comput. Commun. Applic.* 3, 3 (2007), 13.

[5] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL'18)*. 2406–2417.

[6] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL'17)*. 1657–1668.

[7] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP'16)*. 551–561.

[8] Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of languagevariability. In *Proceedings of the Learning Methods for Text Understanding and Mining Workshop*.

[9] Fangxiang Feng, Xiaojie Wang, Ruifan Li, and Ibrar Ahmad. 2015. Correspondence autoencoders for cross-modal retrieval. *ACM Trans. Multim. Comput. Commun. Applic.* 12, 1s (2015), 26:1–26:22.

[10] Yichen Gong, Heng Luo, and Jian Zhang. 2017. Natural language inference over interaction space. *arXiv preprint arXiv: abs/1709.04348* (2017).

[11] Dan Han, Pascual Martínez-Gómez, and Koji Mineshima. 2017. Visual denotations for recognizing textual entailment. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP'17)*. 2853–2859.

[12] Sanda M. Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Meeting of the Association for Computational Linguistics*. 905–912.

[13] Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28, 3–4 (1936), 321–377.

[14] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 2261–2269.

[15] Xin Huang and Yuxin Peng. 2018. Deep cross-media knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 8837–8846.

[16] Xin Huang, Yuxin Peng, and Mingkuan Yuan. 2017. Cross-modal common representation learning by hybrid transfer network. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'17)*. 1893–1900.

[17] Cuicui Kang, Shiming Xiang, Shengcai Liao, Changsheng Xu, and Chunhong Pan. 2015. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Trans. Multim.* 17, 3 (2015), 370–381.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* 123, 1 (2017), 32–73.

[19] G. Hinton, A. Krizhevsky, and I. Sutskever. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS'12)*. 1106–1114.

[20] Kai Li, Guo-Jun Qi, and Kien A. Hua. 2018. Learning label preserving binary codes for multimedia retrieval: A general approach. *ACM Trans. Multim. Comput. Commun. Applic.* 14, 1 (2018), 2:1–2:23.

[21] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS'16)*. 289–297.

[22] Bill MacCartney. 2009. *Natural Language Inference*. Ph.D. thesis. Stanford University.

[23] George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[24] Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of the 47th Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 791–799.

[25] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL'16)*. 1466–1477.

[26] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning (ICML'11)*. 689–696.

[27] Biswajit Paria, K. M. Annervaz, Ambedkar Dukkipati, Ankush Chatterjee, and Sanjay Podder. 2016. A neural architecture mimicking humans end-to-end for natural language inference. *arXiv preprint arXiv: abs/1611.04741* (2016).

[28] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP'16)*. 2249–2255.

[29] Yuxin Peng, Xin Huang, and Jinwei Qi. 2016. Cross-media shared representation by hierarchical learning with multiple deep networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'16)*. 3846–3853.

[30] Yuxin Peng, Xin Huang, and Yunzhen Zhao. 2018. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Trans. Circ. Syst. Vid. Technol.* 28, 9 (2018), 2372–2385.

[31] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. 2018. CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Trans. Multim.* 20, 2 (2018), 405–420.

[32] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vis.* 123, 1 (2017), 74–93.

[33] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the ACM International Conference on Multimedia (ACM MM'10)*. 251–260.

[34] Lei Sha, Baobao Chang, Zhifang Sui, and Sujian Li. 2016. Reading and thinking: Re-read LSTM unit for textual entailment recognition. In *Proceedings of the International Conference on Computational Linguistics (COLING'16)*. 2870–2879.

[35] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. DiSAN: Directional self-attention network for RNN/CNN-free language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'18)*.

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

[37] Yi Tay, Luu Anh Tuan, and Siu Cheung. 2018. Compare, compress, and propagate: Enhancing neural architectures with alignment factorization for natural language inference. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP'18)*. 1565–1575.

[38] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'17)*. 1290–1296.

[39] Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL'16)*. 1442–1451.

[40] Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'17)*. 4144–4150.

[41] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. 2017. Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Trans. Cyber.* 47, 2 (2017), 449–460.

[42] Hongyang Xue, Zhou Zhao, and Deng Cai. 2017. Unifying the video and question attentions for open-ended video question answering. *IEEE Trans. Image Proc.* 26, 12 (2017), 5656–5666.

[43] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *Trans. Assoc. Computat. Ling.* 4 (2016), 259–272.

[44] Hong Yu and Tsendsuren Munkhdalai. 2017. Neural tree indexers for text understanding. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*. 11–21.

Received January 2019; revised July 2019; accepted September 2019