

# Webly Supervised Knowledge Embedding Model for Visual Reasoning

Wenbo Zheng<sup>1,2</sup> Lan Yan<sup>2,4</sup> Chao Gou<sup>3\*</sup> Fei-Yue Wang<sup>2,4</sup>

<sup>1</sup> School of Software Engineering, Xi'an Jiaotong University

<sup>2</sup> The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences

<sup>3</sup> School of Intelligent Systems Engineering, Sun Yat-sen University

<sup>4</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences

zwb2017@stu.xjtu.edu.cn; yanlan2017@ia.ac.cn; gouchao@mail.sysu.edu.cn; feiyue.wang@ia.ac.cn

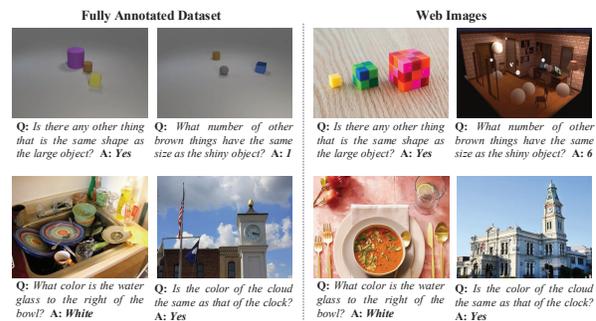
## Abstract

Visual reasoning between visual image and natural language description is a long-standing challenge in computer vision. While recent approaches offer a great promise by compositionality or relational computing, most of them are oppressed by the challenge of training with datasets containing only a limited number of images with ground-truth texts. Besides, it is extremely time-consuming and difficult to build a larger dataset by annotating millions of images with text descriptions that may very likely lead to a biased model. Inspired by the majority success of webly supervised learning, we utilize readily-available web images with its noisy annotations for learning a robust representation. Our key idea is to presume on web images and corresponding tags along with fully annotated datasets in learning with knowledge embedding. We present a two-stage approach for the task that can augment knowledge through an effective embedding model with weakly supervised web data. This approach learns not only knowledge-based embeddings derived from key-value memory networks to make joint and full use of textual and visual information but also exploits the knowledge to improve the performance with knowledge-based representation learning for applying other general reasoning tasks. Experimental results on two benchmarks show that the proposed approach significantly improves performance compared with the state-of-the-art methods and guarantees the robustness of our model against visual reasoning tasks and other reasoning tasks.

## 1. Introduction

Visual reasoning demands a strong model to learn relational computing and the ability of compositionality and generalization, i.e., understanding and answering compositional or relational questions without having seen similar semantic compositions before [10, 20, 21, 31]. Besides, visual reasoning tasks, the general task of asking questions about images, having its own line of datasets, which generally focus on asking a series of simple questions on an image, is usually answerable in a single glance. The illustrated example is shown in Figure 1.

The success in visual reasoning tasks with image-text (question, answer) pairs from hand-labeled image datasets (e.g., GQA [16], CLEVR [19]) has been achieved by training the joint embedding model in the form of supervised learning. While these datasets cover a large number of images (e.g., about 20M in GQA and 100K in CLEVR), it is labor-intensive and difficult for using image-text pairs to create more larger datasets [17]. In addition, it is usually feasible for only a limited number of users to annotate the training images, which may cause the model to be biased [25, 35]. Therefore, although these datasets provide convenient modeling assumptions, they are very limited given the large number of rich descriptions that humans can make



\*Chao Gou is the corresponding author.

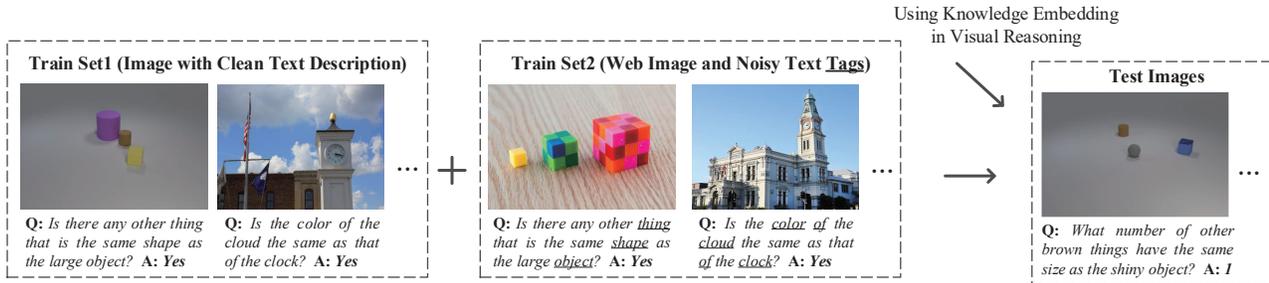


Figure 2. The Open-World Setting of Our Paper. We focus on the learning of robust knowledge embedding using clean images with ground-truth text, and update this learning by utilizing web images and its noisy associated tags. During this process, the latent space is learned and tested by images and text descriptions from our web datasets.

up. Thus, applying the trained model which has excellent performance on benchmark datasets in the open-world setting, may not meet the requirement of good generalization for other visual reasoning tasks.

Image streams with noisy labels are easily obtained from the dataset, such as GQA [16] and OK-VQA [22], and can also be obtained on the web in almost unlimited numbers. Considering a large number of web images, developing an effective visual reasoning system is likely to be robust. However, it may increase ambiguity and reduce performance for using weakly annotated images.

Motivated by the above observation, we put forward an essential question in this paper: *Can abundant noisy annotated web images be leveraged upon with fully annotated images to learn better joint knowledge embedding for visual reasoning?* Figure 2 shows an illustration of this scenario.

In this work, we focus on how to advisably and prudently use web images for developing a robust visual reasoning system. We present a novel mechanism and framework that can enhance knowledge through a useful embedding model with weakly supervised web data. In visual reasoning tasks, our method is always better than previous methods. It reveals the importance of efficiently learning with large-scale web data for more comprehensive representation. We hope and believe our work can be provided with insights for researchers.

## 1.1. Overview of Our Approach

In the visual reasoning task, we propose a novel, effective and robust knowledge memory embedding model with mutual modulation for visual reasoning, which exploits knowledge in the whole process. In this work, we aim to improve joint embeddings, which is trained by images and text (question, ground-truth answer) descriptions, using web images annotated with noisy tags. However, during the embedding training process, it is non-trivial that we combine the web image-tag pairs with image-text (question, ground-truth answer) pairs, due to the differences between

text descriptions and tags representations.

To bridge this gap, we present a two-stage approach for learning the representation of joint image-text. In Stage I, we take advantage of the available clean image-text (question, ground-truth answer) pairs from a dataset in a supervised formulation. Specially, we first design a key-value memory network that can learn the prior knowledge-based representations of textual and visual information, and then we obtain the embeddings of knowledge-based question information. Next, we update the mutual modulation to get network-based question information. Finally, our framework associates the embeddings of knowledge-based representations with network-based question information. In Stage II, we update the previously learned knowledge-based representation using weakly-annotated image-tags pairs from the web (e.g., Google Photo). By this stage, we can transfer the knowledge of weakly annotated images from our better visual reasoning system.

## 1.2. Our Contributions

We present a novel and pragmatic problem in this paper—can we utilize large-scale web data to learn an effective knowledge embedding without a lot of hand-crafted annotated training data? To address the above problem, our main contributions are as follows:

- ✧ We propose a webly-supervised approach for learning robust knowledge-based representations, where we make use of images-text descriptions from clean datasets and web images with its noisy tags from the web.

- ✧ We propose an effective and robust knowledge embedding memory model with mutual modulation for visual reasoning tasks.

- ✧ We design knowledge-based representation learning to make our model has the ability to generalize to other reasoning tasks.

- ✧ Experimental results show that the proposed approach has strong robustness and outperforms existing methods in visual reasoning on two benchmarks, especially demon-

strating an average accuracy of 99.7 percent points on the CLEVR dataset, and achieving 14.8% higher in Test-P accuracy over the best baseline on NLVR datasets.

## 2. Related Work

**Visual Reasoning.** The majority of approaches have been recently proposed to solve visual reasoning tasks. Multi-step models (e.g., MAC [14], Neural Module Networks (NMNs) [1]) are performed the visual reasoning tasks. These kind models create the layouts of image and question, and execute these layouts to get the answers. In particular, variants of this method build memory network to record information. These methods have also been applied to REF, e.g., CMN [13] and Stack-NMN [11]. FiLM [29] modulates the representation of image and question using conditional batch normalization, where both modalities can modulate each other. The FiLM can be extended in with multi-step reasoning. These models can perform complex relational reasoning, but their reasoning representations are built on visual appearance features that do not contain much knowledge-based information and information about the relationships between text and visual features. On the contrary, in order to reason about relationships, they focus on the heavily on manually designed inference structures or modules, and are appropriate for specific tasks. *To tackle these problems, in this paper, we propose a novel, effective and robust knowledge memory embedding model with mutual modulation for visual reasoning, which exploits prior knowledge in the whole process.*

**Webly Supervised Computer Vision.** The idea of utilizing web images for supervising computer vision algorithms has been explored in several tasks, such as object classification [41], object detection [7], object parts localization [26] and object segmentation [32]. The motivation of our work inspired by these efforts is to learn more powerful models by realizing the feasibility of web data. We believe that it is exceptionally significant and pragmatic for improving the generalization of image-text based knowledge embedding models that we supplement scarce clean image text (question, answer) data with web images to our model, as the largest CLEVR [19] dataset for visual reasoning has only 100K training images. *To the best of our knowledge, this is the first attempt to propose webly supervised model for visual reasoning.*

**Knowledge-Based Reasoning over Knowledge Bases.** Lots of knowledge bases have been built taking advantage of image-text pairs or for visual reasoning tasks [49–51]. These knowledge bases are potentially helpful resources for answering questions in our dataset. In the field of natural language processing (NLP), knowledge-based question answering has been brought into focus (e.g., [45, 47]).

## 3. Webly Supervised Approach

In this section, firstly, we describe the network structure. Then, based on our network structure, we propose knowledge-based representation learning for visual reasoning. Finally, we present our strategy to incorporate the noisy tags into our framework for learning an improved embedding. The training procedure of our approach is shown in Figure 3.

### 3.1. The Network Structure

Figure 4 shows the overview of the proposed network. First, an image  $I$  and a visual question  $Q$  is fed into our designed **Mutual Modulation**, which produces the final representation of network-based question information  $q^{network}_n$  and network-based visual information  $\{V_i, i = 1, 2, 3, \dots, n\}$ . Then, using the given knowledge bases, we can obtain the final representation of knowledge-based question information  $q^{knowledge}_{H+1}$  by our **Knowledge-Based Key-Value Memory Network**, after  $H$  iterations. From the final representation of knowledge-based question information and network-based question information, knowledge-attention representation  $q$  can be obtained using knowledge-based representation learning mentioned in next subsection, which is used to predict the answer  $\hat{a}$  of the visual question.

**Mutual Modulation** In order to better integrate visual modalities and linguistic modalities to solve visual reasoning problems, we have redesigned the mutual modulation model, according to the work of Yao et al. [46]. In each step  $i$  ( $i = 1, 2, 3, \dots, n$ ), we cascade the visual modulation with the language modulation. Specifically, we feed  $V_{i-1}$  into the visual modulation by parameters from  $q^{network}_{i-1}$  to compute  $V_i$ , and then control the process of the language modulation with parameters from  $V_i$ , to compute the new question vector  $q^{network}_i$ .

**Knowledge-Based Key-Value Memory Network** We have designed Key-Value Memory Networks based on Memory Network architecture of Miller et al. [24] and Sukhbaatar et al. [39]. First, we design a memory, which is a possibly vast array of slots. We can use the memory to encode both short-term and long-term context. We define the memory slots as key-value pairs of  $M$ -dimensional vectors and denote the question  $Q$ . Then, we use the iterative process of the key addressing and value reading from memory to look for concerned information to answer  $Q$ . Note that these iterations are also called “hops”. At each step, the received information from memory is accumulatively added to the original question to construct the representation of knowledge for the next round. After a fixed number  $H$  hops, we can get the final representation of knowledge-based question information  $q^{knowledge}_{H+1}$ . Also, we use Pezeshkpour et al.’s work [30] to build a knowledge base.

### Train Initial Knowledge Embedding Using Fully Annotated Dataset

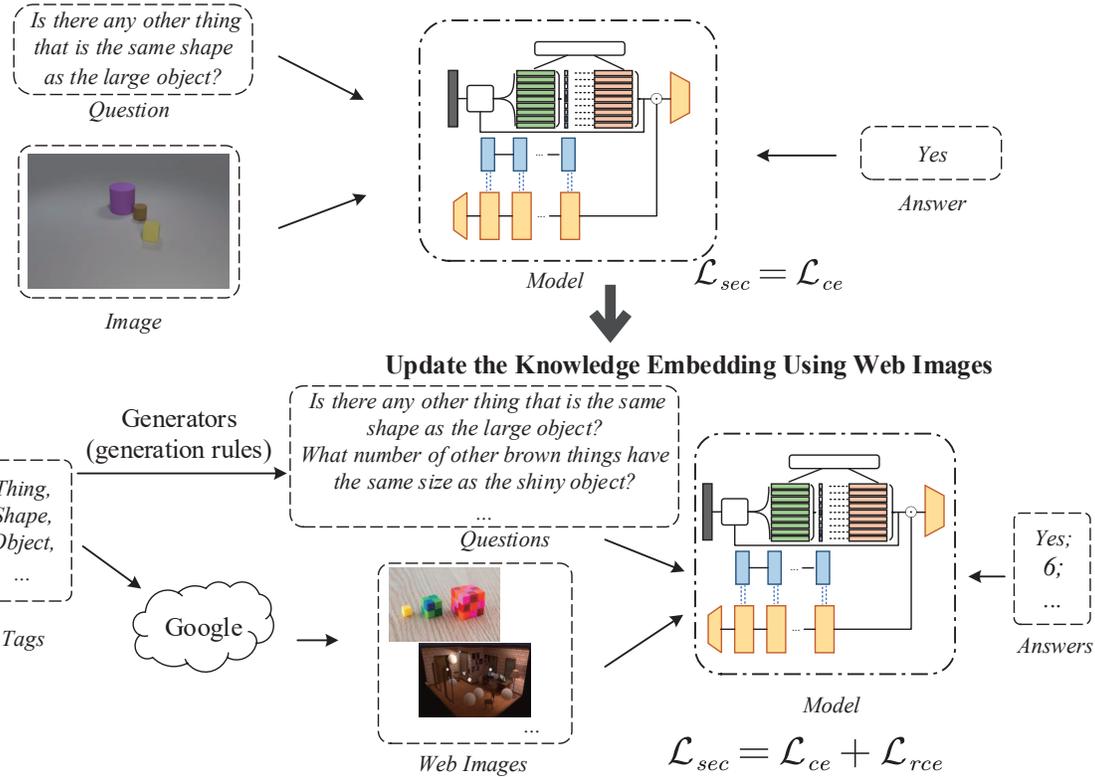


Figure 3. A Brief Illustration of the Proposed Framework. We use the image-text pairs from the clean dataset and image-tag pairs from the web to learn the knowledge embedding model. Firstly, we use the image from the clean and their text descriptions for learning our model. Then, we update our model using web images and their tags.

### 3.2. Knowledge-Based Representation Learning

We introduce the gated mechanism that embeds the knowledge representation to enhance the representation learning, considering suppressing non-informative features and allowing informational features to pass under the guidance of the knowledge-based key-value memory network. Similar to [5,6], we introduce a gated mechanism expressed as

$$q = \sigma(g(q^{network}_n, q^{knowledge}_{H+1})) \odot q^{network}_n \quad (1)$$

where  $\sigma$  is the logistic sigmoid,  $\odot$  denotes the element-wise multiplication operation,  $g$  is a neural network that takes the concatenation of the final representation of network-based question information by using mutual modulation and knowledge-based question information.

We use a fully connected layer with 1024 ReLU hidden units [27] as our answer generator. It takes  $q$  and  $\{\mathbf{V}_i, i = 1, 2, 3, \dots, n\}$  as input, and predicts the most probable answer  $\hat{a}$ :

$$\hat{a} = \arg \max_{i=1,2,3,\dots,n} softmax(q^T \times B \times \mathbf{V}_i) \quad (2)$$

where the  $d \times D$  matrix  $B$  is able to be identical and be constrained to  $A$ .

We minimize a standard cross-entropy loss [48] between  $\hat{a}$  and the correct answer  $a$  to train the end-to-end network, which learns to perform the iterative accesses to output the desired target  $a$ .

### 3.3. Training with Noisy Web Images

In this subsection, we take advantage of image-tag pairs from the web to improve trained knowledge embeddings using the clean datasets with image-text (question, answer) pairs. Our goal is to get an excellent representation of image-text knowledge embedding, which can generalize well and ideally be capable of data-dependent noise resistance. This approach is essentially an implicit data augmentation, as the effective use of web data increases the sample size used for training the model. However, we cannot directly apply the web data to update our trained model using image-tag pairs. Besides, considering the representation of tags, the traditional NLP approach cannot deal with any semantic context as in the text (questions).

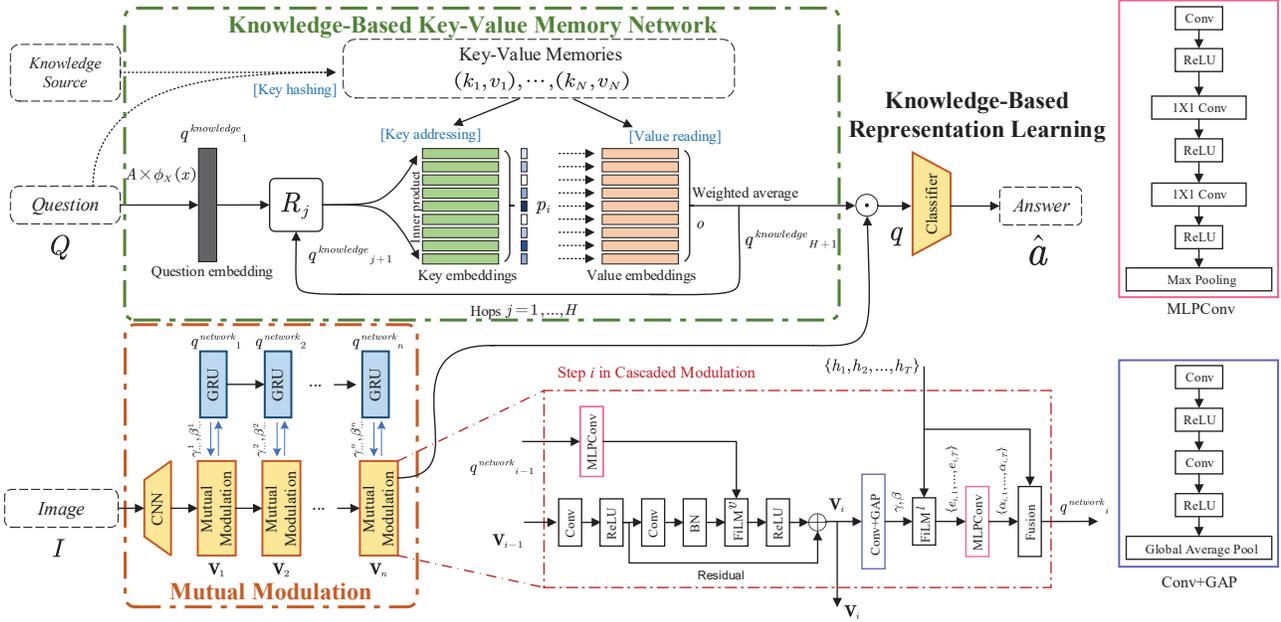


Figure 4. Our Network Model. We first design a key-value memory network that can learn the prior knowledge-based representations of textual and visual information, and then we obtain the embeddings of knowledge-based question information. Next, we update the mutual modulation to get network-based question information. Finally, our framework associates the embeddings of knowledge-based representations with network-based question information.

In the setting of our paper, we can use an additional tag modality during the training process and take advantage of this easily available information for training a more outstanding model. The clean datasets (e.g., CLEVR [19], NLVR [38]) only give the image-text (question, answer) pairs and cannot give more the information of tags. On the contrary, the web resources always give the images and their tags but do not give any text descriptions.

To bridge this gap, we present a two-stage approach for obtaining the excellent representation of image-text pairs. In the first stage, we take advantage of image-text pairs from the clean datasets for learning an initial representation. In the second stage, we update a trained model in the first stage using image-text pairs from the web datasets.

### 3.3.1 Stage I: Training with Clean Dataset

We take advantage of image-text pairs from the annotated datasets to learn a knowledge embedding. For the learning of knowledge representation, we use the **symmetric cross entropy**, which provides its effectiveness against various types and rates of label noise.

$$\mathcal{L}_{sce} = \tau \times \mathcal{L}_{ce} + \upsilon \times \mathcal{L}_{rce} \quad (3)$$

where  $\tau$  and  $\upsilon$  are two hyperparameters,  $\mathcal{L}_{ce}$  means a standard cross-entropy loss [48], and  $\mathcal{L}_{rce}$  means reverse cross entropy loss [42]. Details about  $\mathcal{L}_{sce}$  are shown in Ref [42].

In Eq. 3,  $\tau$  and  $\upsilon$  are predefined weights for different losses. In our first stage, the reverse cross-entropy loss is not used ( $\tau = 1$  and  $\upsilon = 0$ ) while in the second stage, both losses are used ( $\tau = 1$  and  $\upsilon = 1$ ).

### 3.3.2 Noisy Web Image

**Web Image Sets** We use Google Photo API [2] to retrieve web images via inputting tags from NUS-WIDE dataset [8]. We would like to utilize this web image without any hand-crafted labels. We build the list of 1000 most recurrent keywords using GQA [16] and CLEVR [19] dataset text (question, answer) descriptions. We sort these keywords in descending order based on the frequency and remove stop-words. Then, we group similar words after performing lemmatization. We utilize this list of keywords to query and retrieve around two hundred images per query, recording with their tags. All in all, we use the above way to collect about more than two hundred thousand images with their tags. During this process, we only retain the image which has at least 2 English tags and not more than four images from the same website source. We also use the first five tags to remove repetitive images.

**Questions Generation** Similarly to other synthetic benchmark datasets (e.g., CLEVR [19], GQA [16], EQA [9], TextVQA [36]), we choose to generate questions according to functional templatestyle representations (e.g.

“How many  $\langle attr \rangle \langle obj\_type - pl \rangle$  are in the  $\langle room\_type \rangle$ ?”). This facilitates the instantiation of ground-truth tags, once the image for the corresponding trajectory has been generated and analyzed. Moreover, we can easily execute the corresponding program to determine the answer this amounts to performing a series of basic operations, such as *input()*, *filter()*, *count()*, *unique()*, *get\_attr()* on the ground-truth.

The question generation process starts by randomly choosing one of the 28 templates to be instantiated. A valid question will always have tags instantiated with ground-truth values. For example, if there is a  $\langle room\_type \rangle$  tag and we have only seen a kitchen and a living room on our trajectory, then the set of possible instantiations is  $\{kitchen, living\ room\}$ . Using this principle, we build sets of possible values for each tag in the template. In order to generate a valid (question, answer) pair, we randomly assign each tag a value from its set, then run the template functional program to compute whether the question is valid and can be answered using the ground-truth. To illustrate the process, consider the template “What color is the  $\langle attr \rangle \langle obj\_type \rangle$ ?” with the associated program:

```
input(objs) → filter(obj_type) → filter(attr) →
unique() → get_attr(color)
```

We filter by the instantiated object type, then by the instantiated attribute (enforced not to be a color during the tag value assignment). Then, we ensure that the result is unique (i.e., that the question is unambiguous) and retrieve the color of the object as the answer.

### 3.3.3 Stage II: Training with Web Images

While the first stage (i.e., Stage I) has achieved, we get the representation of image and text (question, answer) description and the learned knowledge embedding model. During the second training stage (i.e., Stage II), we update the trained learned knowledge embedding model from Stage I, using weakly-annotated image-text(question, answer) pairs from noisy web images. This allows us to transfer knowledge from thousands of easily available webly-annotated images to the learned model. We set a lower learning rate in this way, since the network obtains outstanding performance after the first stage, and adjusting our network with a high learning rate from the webly-annotated image may cause catastrophic forgetting.

Since web data is straightforward to get, and their labels are noisy, it is challenging to learn good representations for the task of visual reasoning in many cases. Thus, during the second stage, we employ the strategy of curriculum learning [3] for training. Curriculum learning enables our model to learn from easy cases to complex cases. In other words, we can learn from simpler examples first, so they can be

used as a basis for learning more complex examples, resulting in better performance in the final task. Many previous works have shown that appropriate curriculum strategies can guide learners to better master local knowledge [23,42]. We gradually inject difficult information into our network, and the feature of outputs of the network is related to frequently occurring knowledge in the clean training set, in the early stages of training. The features related to rarely occurring knowledge are shown at a later stage. Due to the trained network in the first stage have outstanding representations about frequently occurring knowledge, the noisy label of web images may not go down the performance of our network.

## 4. Experiments and Results

In this section, we experimentally evaluate the performance of the proposed model on two benchmark datasets, and compare its performance with other state-of-the-art deep representation learning models.

### 4.1. Dataset Description

#### 4.1.1 The CLEVR Dataset

CLEVR, consisting of 700,000 (image, question, answer, program) tuples [19], is a synthetic dataset. Programs, such as *filter shape[cube]*, *relate[right]*, and *count*, consists of step-by-step instructions. Besides, on the way how to answer the question, they are an additional supervisory signal. Answers are each one word from a set of 28 possible answers. Questions are multi-step and compositional in nature. Images contain 3D-rendered objects of various shapes, materials, colors, and sizes. They range from counting questions to comparison questions and can be more than 40 words long.

#### 4.1.2 The NLVR Dataset

NLVR [38] is a visual reasoning dataset proposed by researchers in the NLP field. NLVR has 74,460 samples for training, 5,940 for validation and 5,934 for public test. In each sample, there is a human-posed natural language description on an image with 3 sub-images, and requires a false/true response.

## 4.2. Experiment Setup

In this subsection, we outline the criteria used for evaluation and then we describe the implementation details.

### 4.2.1 Evaluation Criteria

**Experiments on The CLEVR Dataset** We can use the program representation of questions to analyze model performance on different forms of reasoning. We use one eval-

uation metric, accuracy(%), on six question types, including Overall, Exist, Count, Compare Integer, Query Attribute and Compare Attribute. This is a traditional way to evaluate following the work of Li et al. [19].

**Experiments on The NLVR Dataset** NLVR is split into training, development, and test sets. The test set is public (Test-P) and available with the data. For both datasets, we use one evaluation metrics: accuracy. Accuracy (Acc) is computed as the proportion of examples (sentence-image pairs) for which a model correctly predicted a truth value.

### 4.2.2 Implementation Details

All experiments are conducted using a 4-core PC with a 12 GB NVIDIA TITAN XP GPU, 16GB of RAM, and Ubuntu 16. We continue training Stage I for an initial 120 epochs. Then we start updating the learned model in Stage I with web images in Stage II for another 120 epochs. The detailed implementations for mutual modulation and knowledge-based key-value memory network are as follows:

**Mutual Modulation** We embed the question words into a 200-dim continuous space, and use a GRU with 4096 hidden units to generate 1024-dim question representations. Questions are padded with the NULL token to a maximum length  $T = 50$ . The feature map number  $C$  is set to 128. Images are pre-processed with a ResNet101 network pre-trained on ImageNet [33] to extract  $1024 \times 14 \times 14$  visual features. We use a trainable one-layer CNN with 128 kernels ( $3 \times 3$ ) to encode the extracted features into  $V_0$  ( $128 \times 14 \times 14$ ). We train the model with an SGD [18] using a learning rate of  $1e - 5$  and a batch-size of 64 and 0.9 momentum, fine-tuning for 120 epochs.

**Knowledge-Based Key-Value Memory Network** We use Pezeshkpour et al.’s work [30] to build the knowledge bases. Our models were trained using an SGD [18] with a learning rate of  $\eta = 0.001$ , with anneals every 25 epochs by  $\eta/2$  until 120 epochs were reached. No momentum or weight decay was used. The weights were initialized randomly from a Gaussian distribution with zero mean and  $\sigma = 0.1$ . All training uses a batch size of 32 (but the cost is not averaged over a batch).

### 4.3. Comparison with State-of-The-Art Methods

We compare the state-of-the-art approaches with our model on two benchmarks, including the CLEVR [19] dataset, and the NLVR [38] dataset respectively. In this subsection, “Ours w/o Web Images” means a variant of Ours, which only using clear datasets and not using web images.

#### 4.3.1 Comparison on The CLEVR Dataset

On the CLEVR dataset, we compare ours with the state-of-the-art approaches, including Q-type baseline [19], LSTM [19], CNN+LSTM [19], CNN+LSTM+SA

Table 1. Comparison Results on The CLEVR Dataset

Model	Overall	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
Human	92.6	86.7	96.6	86.5	95.0	96.0
Q-type baseline [19]	41.8	34.6	50.2	51.0	36.0	51.3
LSTM [19]	46.8	41.7	61.1	69.8	36.8	51.8
CNN+LSTM [19]	52.3	43.7	65.2	67.1	49.3	53.0
CNN+LSTM+SA [34]	68.5	52.2	71.1	73.5	85.3	52.3
CNN+LSTM+RN [34]	95.5	90.1	97.8	93.6	97.9	97.1
CNN+LSTM+RN+ [34]	90.9	86.7	97.4	90.0	90.2	93.5
SAN [44]	76.7	64.4	82.7	77.4	82.6	75.4
N2NMN [12]	83.7	68.5	85.7	84.9	90.0	88.7
PG+EE-9K [19]	88.6	79.7	89.7	79.1	92.6	96.0
CNN+LSTM+multiRN [4]	92.3	85.2	96.5	93.6	95.1	92.9
CNNh+LSTM+multiRN [4]	97.2	94.1	98.9	98.3	98.6	97.6
CNNh+LSTM+multiRN+ [4]	97.7	94.9	99.2	97.2	98.7	98.3
PG+EE-700K [19]	96.9	92.7	97.1	98.7	98.1	98.9
RN [34]	95.5	90.1	97.8	93.6	97.9	97.1
COG-model [43]	96.8	91.7	99.0	95.5	98.5	98.8
FiLM [29]	97.7	94.3	99.1	96.8	99.1	99.1
FiLM-raw [29]	97.6	94.3	99.3	93.4	99.3	99.3
DDRprog [37]	98.3	96.5	98.8	98.4	99.1	99.0
CAN [15]	98.9	97.1	99.5	99.1	99.5	99.5
CMM-single [46]	98.6	96.8	99.2	97.7	99.4	99.1
CMM-ensemble [46]	99.0	97.6	99.5	98.5	99.6	99.4
Ours w/o Web Images	<b>99.3</b>	<b>98.5</b>	<b>99.9</b>	<b>99.2</b>	<b>99.7</b>	<b>99.5</b>
Ours	<b>99.8</b>	<b>99.7</b>	<b>99.9</b>	<b>99.9</b>	<b>99.7</b>	<b>99.7</b>

[34], CNN+LSTM+RN [34], CNN+LSTM+RN+ [34], SAN [44], N2NMN [12], PG+EE-9K [19], CNN+LSTM+multiRN [4], CNNh+LSTM+multiRN [4], CNNh+LSTM+multiRN+ [4], PG+EE-700K [19], RN [34], COG-model [43], FiLM [29], FiLM-raw [29], DDRprog [37], CAN [15], CMM-single [46], and CMM-ensemble [46]. The results are shown in Table 1.

**Effect of Proposed Webly Supervised Training.** For evaluating the performance of our approach, we compare results reported in row-“Ours w/o Web Images” and row-“Ours” from Table 1. Our approach leverages the same loss functions and features in row-“Ours w/o Web Images” for a fair comparison. From Table 1, we find that our approach improves performance consistently in all the cases. *It is evident that using webly supervised training can enhance the effectiveness of our approach.*

**Effect of Our Approach.** From Table 1, it is evident that our approach is better than others. Specifically, ours is 58.0%, 53.0%, 47.5%, 31.3%, 4.3%, 8.9%, 23.1%, 16.1%, 11.2%, 7.5%, 2.6%, 2.1%, 2.9%, 4.3%, 3%, 2.1%, 2.2%, 1.5%, 0.9%, 1.2%, and 0.8% higher than Q-type baseline, LSTM, CNN+LSTM, CNN+LSTM+SA, CNN+LSTM+RN, CNN+LSTM+RN+, SAN, N2NMN, PG+EE-9K, CNN+LSTM+multiRN, CNNh+LSTM+multiRN, CNNh+LSTM+multiRN+, PG+EE-700K, RN, COG-model, FiLM, FiLM-raw, DDRprog, CAN, CMM-single, and CMM-ensemble, in term of overall, respectively. In term of Count type, Exist type, Compare Numbers type, Query Attribute type and Compare Attribute type, there are similar scenarios as the above. Besides, ours is better than human performance. *From above, our approach is more effective and robust than the state-of-the-arts approaches on the CLEVR dataset.*

#### 4.3.2 Comparison on The NLVR Dataset

On the NLVR dataset, we compare ours with the state-of-the-art approaches, including CNN-BiATT [40], N2NMN

Table 2. Comparison Results on The NLVR Dataset

Model	Dev. (Acc/%)	Test-P (Acc/%)
Human Performance	94.6	95.4
CNN-BiATT [40]	66.9	69.7
N2NMN [12]	65.3	69.1
Neural Module Networks [1]	63.1	66.1
FiLM [29]	60.1	62.2
Majority Class [38]	55.3	56.2
MAC-Network [14],	55.4	57.6
CMM [46]	68	69.9
W-MemNN [28]	65.6	65.8
Ours w/o Web Images	<b>72.4</b>	<b>74.3</b>
Ours	<b>81.3</b>	<b>80.6</b>

[12], Neural Module Networks [1], FiLM [29], Majority Class [38], MAC-Network [14], CMM [46], and W-MemNN [28].

#### Effect of Proposed Webly Supervised Training.

“Ours” is 8.9% and 6.3% higher than “Ours w/o Web Images”. These improvements once again show that learning by utilizing large scale web data covering a wide variety of knowledge lead to a robust knowledge embedding for visual tasks.

**Effect of Our Approach.** From Table 2, it is visible that our approach is better than others. Specifically, our approach is 10.9%, 11.5%, 14.5%, 18.4%, 24.4%, 23%, 10.7%, and 14.8% higher than CNN-BiATT, N2NMN, Neural Module Networks, FiLM, Majority Class, MAC-Network, CMM, and W-MemNN, in term of Test-P, respectively. From above, our approach is more effective and robust than the state-of-the-arts on the NLVR dataset.

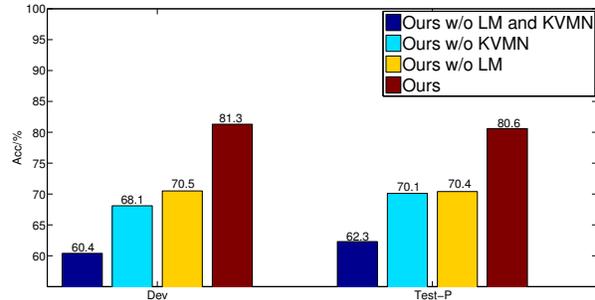
#### 4.4. Ablation Study

In order to verify the reasonableness and effectiveness of each part of our attention machine, we design the ablation experiment. In Figure 5(a), and Figure 5(b), “Ours w/o LM and KVMN” means a variant of Ours, which removes language modulation and key-value memory network; “Ours w/o LM” means a variant of Ours, which removes language modulation; “Ours w/o KVMN” means a variant of Ours, which removes key-value memory network. We analyze the following two aspects:

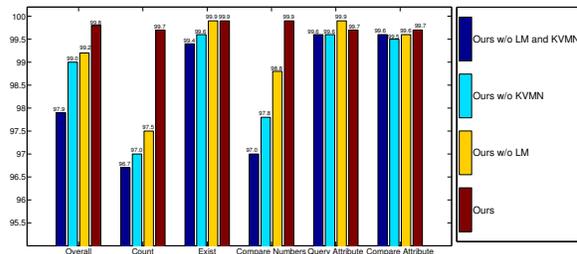
**Compared with “Ours”** From Figure 5(a), ours is 18.3%, 10.5%, and 10.2% higher than “Ours w/o LM and KVMN”, “Ours w/o KVMN”, and “Ours w/o LM”, respectively. As we can see, “Ours” is better than others. These suggest making joint use of textual and visual information helps us to improve the task of visual reasoning.

**Compared with “Ours w/o KVMN”** “Ours w/o KVMN” is 0.3% and 10.5% lower than “Ours w/o LM”, and ours, respectively. As we can see, “Ours w/o KVMN” is worse than “our without LM”. These suggest the importance of making full use of textual and visual information.

From the above, we get the conclusion in the following two aspects:



(a)



(b)

Figure 5. The Results of Ablation Study; (a) Ablation Results on The NLVR Dataset; (b) Ablation Results on The CLEVR Dataset.

(1) It is apparent that the design of language modulation and key-value memory network improves visual reasoning.

(2) It is manifest that the design of the key-value memory network is better than our language modulation. This suggests that the design of key-value memory network is more robust and effective.

Moreover, by analyzing ablation results shown in Figure 5(b) on CLEVR dataset, we can get similar conclusions.

## 5. Conclusion and Future Work

In this work, we show how to take advantage of web images with tags to assist in building strong and effective knowledge embedding models for the task of visual reasoning with limited labeled data. To address this challenge, we present a two-stage approach that can enhance knowledge through effective embedding model with weakly supervised web data. Experimental results demonstrate that our approach significantly improves the performance in the visual reasoning task in two benchmark datasets. Following this way, we will improve our approach by exploiting other types of meta-data (e.g., medical data, sensor data, social media data) in the future.

**Acknowledgments** We would like to thank the anonymous reviewers for their useful feedback. This work is supported in part by MOST and NNSF of China (2008AAA0101502, 61533019, 61806198, U1811463), and Squirrel AI Learning.

## References

- [1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [2] Google Photo API, <https://developers.google.com/photos>.
- [3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pages 41–48, New York, NY, USA, 2009. ACM.
- [4] Simyung Chang, John Yang, SeongUk Park, and Nojun Kwak. Broadcasting convolutional network for visual relational reasoning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 780–796, Cham, 2018. Springer International Publishing.
- [5] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *Proceedings of the 26th ACM International Conference on Multimedia, MM '18*, pages 2023–2031, New York, NY, USA, 2018. ACM.
- [6] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Conference on Computer Vision and Pattern Recognition*, 2019.
- [7] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [8] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '09*, pages 48:1–48:9, New York, NY, USA, 2009. ACM.
- [9] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [10] Justin Halberda, Michèle M. M. Mazzocco, and Lisa Feigenson. Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213):665–668, 2008.
- [11] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [12] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [13] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [14] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. *CoRR*, abs/1803.03067, 2018.
- [15] Drew Arad Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018.
- [16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [17] B. Jin, M. V. O. Segovia, and S. Ssstrunk. Webly supervised semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1705–1714, July 2017.
- [18] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. Stochastic gradient descent escapes saddle points efficiently. *CoRR*, abs/1902.04811, 2019.
- [19] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.
- [20] Guohao Li, Xin Wang, and Wenwu Zhu. Perceptual visual reasoning with knowledge propagation. In *Proceedings of the 27th ACM International Conference on Multimedia, MM '19*, pages 530–538, New York, NY, USA, 2019. ACM.
- [21] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [22] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [23] T. Matiisen, A. Oliver, T. Cohen, and J. Schulman. Teacher-student curriculum learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–9, 2019.
- [24] Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. *CoRR*, abs/1606.03126, 2016.
- [25] L. Niu, A. Veeraraghavan, and A. Sabharwal. Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7171–7180, June 2018.
- [26] David Novotný, Diane Larlus, and Andrea Vedaldi. Learning the semantic structure of objects from web supervision. *CoRR*, abs/1607.01205, 2016.
- [27] Y. Pang, M. Sun, X. Jiang, and X. Li. Convolution in convolution for network in network. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1587–1597, May 2018.
- [28] Juan Pavez, Héctor Allende, and Héctor Allende-Cid. Working memory networks: Augmenting memory networks with a relational reasoning module. In *Proceedings of the 56th*

- Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018.
- [29] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.
- [30] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3208–3218, 2018.
- [31] Sandro Pezzelle and Raquel Fernández. Big generalizations with small data: Exploring the role of training samples in learning adjectives of size. In *Proceedings of the Beyond Vision and LANGUAGE: inTEgrating Real-world kNOWLEDge (LANTERN)*, pages 18–23, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [32] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.
- [33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252, Dec. 2015.
- [34] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy P. Lillicrap. A simple neural network module for relational reasoning. *CoRR*, abs/1706.01427, 2017.
- [35] T. Shen, G. Lin, C. Shen, and I. Reid. Bootstrapping the performance of webly supervised semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1363–1371, June 2018.
- [36] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [37] Joseph Suarez, Justin Johnson, and Fei-Fei Li. DDRprog: A CLEVR differentiable dynamic reasoning programmer. 2018.
- [38] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [39] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS’15*, pages 2440–2448, Cambridge, MA, USA, 2015. MIT Press.
- [40] Hao Tan and Mohit Bansal. Object ordering with bidirectional matchings for visual reasoning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 444–451, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [41] D. Tsai, Yushi Jing, Yi Liu, H. A. Rowley, S. Ioffe, and J. M. Rehg. Large-scale image annotation using visual synset. In *2011 International Conference on Computer Vision*, pages 611–618, Nov 2011.
- [42] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [43] Guangyu Robert Yang, Igor Ganiev, Xiao-Jing Wang, Jonathon Shlens, and David Sussillo. A dataset and architecture for visual reasoning with a working memory. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 729–745, Cham, 2018. Springer International Publishing.
- [44] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [45] Xuchen Yao and Benjamin Van Durme. Information extraction over structured data: Question answering with Freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 956–966, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- [46] Yiquan Yao, Jiaming Xu, Feng Wang, and Bo Xu. Cascaded mutual modulation for visual reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 975–980, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.
- [47] Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331, Beijing, China, July 2015. Association for Computational Linguistics.
- [48] Y. Yuan, G. Xun, F. Ma, Y. Wang, N. Du, K. Jia, L. Su, and A. Zhang. Muvan: A multi-view attention network for multivariate temporal data. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 717–726, Nov 2018.
- [49] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 408–424, Cham, 2014. Springer International Publishing.
- [50] Yuke Zhu, Joseph J. Lim, and Li Fei-Fei. Knowledge acquisition for visual question answering via iterative querying. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [51] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base for visual question answering. *CoRR*, abs/1507.05670, 2015.