R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering

Pan Lu*
Dept. of Computer Science
Tsinghua University
lupantech@gmail.com

Lei Ji[1,2]
Microsoft Research Asia[1], Institute of Computing Technology, CAS[2]
leiji@microsoft.com

Wei Zhang†
School of CSSE
East China Normal University
zhangwei.thu2011@gmail.com

Nan Duan
Microsoft Research Asia
Microsoft Corporation
nanduan@microsoft.com

Ming Zhou
Microsoft Research Asia
Microsoft Corporation
mingzhou@microsoft.com

Jianyong Wang
Dept. of Computer Science
Tsinghua University
jianyong@mail.tsinghua.edu.cn

ABSTRACT

Recently, Visual Question Answering (VQA) has emerged as one of the most significant tasks in multimodal learning as it requires understanding both visual and textual modalities. Existing methods mainly rely on extracting image and question features to learn their joint feature embedding via multimodal fusion or attention mechanism. Some recent studies utilize external VQA-independent models to detect candidate entities or attributes in images, which serve as semantic knowledge complementary to the VQA task. However, these candidate entities or attributes might be unrelated to the VQA task and have limited semantic capacities. To better utilize semantic knowledge in images, we propose a novel framework to learn visual relation facts for VQA. Specifically, we build up a Relation-VQA (R-VQA) dataset based on the Visual Genome dataset via a semantic similarity module, in which each data consists of an image, a corresponding question, a correct answer and a supporting relation fact. A well-defined relation detector is then adopted to predict visual question-related relation facts. We further propose a multi-step attention model composed of visual attention and semantic attention sequentially to extract related visual knowledge and semantic knowledge. We conduct comprehensive experiments on the two benchmark datasets, demonstrating that our model achieves state-of-the-art performance and verifying the benefit of considering visual relation facts.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; • **Information systems** → *Question answering*;

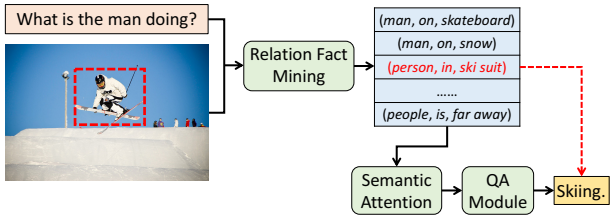

Figure 1: Our proposed model, which learns to mine relation facts with semantic attention for visual question answering.

KEYWORDS

visual question answering; relation fact mining; semantic knowledge; attention network; question answering

ACM Reference Format:
Pan Lu, Lei Ji[1,2], Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. 2018. R-VQA: Learning Visual Relation Facts with Semantic Attention for Visual Question Answering. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220036

1 INTRODUCTION

With the great development of natural language processing, computer vision, knowledge embedding and reasoning, and multimodal representation learning, Visual Question Answering has become a popular research topic in recent years. The VQA task is required to provide the correct answer to a question with a corresponding image, which has been regarded as an important Turing test to evaluate the intelligence of a machine. The VQA problem can be easily expanded to other tasks and play a significant role in various applications, including human-machine interaction and medical assistance. However, it is difficult to address the problem, as the AI system needs to understand both language and vision content, to extract and encode necessary common sense and semantic knowledge, and then to make reasoning to obtain the final answer. Thanks to multimodal embedding methods and attention mechanisms, researchers have made remarkable progress in VQA development.

*This work was mainly performed when Pan Lu was visiting Microsoft Research.
†Wei Zhang is the corresponding author.

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08.
https://doi.org/10.1145/3219819.3220036

The predominant methods first extract language feature embedding by an RNN model and image feature embedding by a pre-trained model, then learning their joint embedding by multimodal fusion like element-wise addition or multiplication, and finally feeding it to a sequential network to generate free-form answers or to a multi-class classifier to predict most related answers. Inspired by image captioning, some VQA approaches [15, 31, 32] introduce semantic concepts such as entities and attributes from off-the-shelf CV methods, which provide various semantic information for the models. Compared with entities and attributes, relation facts have larger semantic capacities as they consist of three elements: subject entity, relation, and object entity, leading to a large number of combinations. For example, in Figure 1, given the question "what is the man doing" and the image, relation facts like *(man, standing on, skateboard)*, *(man, on, ice)*, *(person, in, ski suit)* enable providing important semantic information for question answering.

The main challenge for VQA lies in the semantic gap from language to image. To deal with the semantic gap, existing attempts come in two forms. To be specific, some methods extract high-level semantic information [15, 31, 32], such as entities, attributes, or even retrieval results in knowledge base [15], such as DBpedia [2] and Freebase [5]. Other methods introduce visual attention [13, 35, 38] to select related image regions corresponding to salient visual information. Unfortunately, these progressions of introducing semantic knowledge are still limited in two aspects. On one hand, they use entities or attributes as high-level semantic concepts, which are individual and only cover restricted knowledge information. On the other hand, as they extract the concept based on off-the-shelf CV methods in other tasks or datasets, the candidate concepts might be irrelevant to the VQA task.

To make full use of semantic knowledge in images, we propose a novel semantic attention model for VQA. We build a large-scale Relation-VQA (R-VQA) dataset including over 335k data samples based on the Visual Genome dataset. Each data instance is composed of an image, a relevant question, and a relation fact semantically similar to the image-question pair. We then adopt a relation detector to predict the most related visual relation facts given an image and a question. We further propose a novel multi-step attention model to incorporate visual attention and semantic attention into a sequential attention framework. Our model is composed of three major components (see Figure 4). The visual attention module (Subsection 5.1) is designed to extract image feature representation. The output of the visual attention module is then fed into semantic attention (Subsection 5.2), which learns to select important relation facts generated by the relation detector (Section 4). Finally, joint knowledge learning (Subsection 5.3) is applied to simultaneously learn visual knowledge and semantic knowledge based on visual and semantic feature embeddings.

The main contributions of our work are four-fold.

- We propose a novel VQA framework which enables learning visual relation facts as semantic knowledge to help answer the questions.
- We develop a multi-step semantic attention network (MSAN) which combines visual attention and semantic attention sequentially to simultaneously learn visual and semantic knowledge representations.
- To achieve that, we build up a large-scale VQA dataset accompanied by relation facts and design a fine-grained relation detector model.
- We evaluate our model on two benchmark datasets and achieve state-of-the-art performance. We also conduct substantial experiments to illustrate the ability of our model.

2 RELATED WORK

2.1 Visual Question Answering

As the intersection of natural language processing, knowledge representation and reasoning, and computer vision, the task of Visual Question Answering has attracted increasing interest recently in multiple research fields. A series of large-scale datasets have been constructed, including VQA [1], COCO-QA [26], and Visual Genome [14] datasets. A commonly used framework is to first encode each question as a semantic vector using a long short-term memory network (LSTM) and to extract image features via a pre-trained convolution neural network (CNN), then to fuse these two feature embeddings to predict the answer. In contrast to work in [11, 25, 34] which use simple feature fusion like element-wise operation or concatenation, effective bilinear pooling methods are well studied in [4, 8, 13].

2.2 Attention Methods

Attention networks have recently shown remarkable success in many applications of knowledge mining and natural language processing, such as neural machine translation [3], recommendation systems [29], advertising [42], document classification [39], sentiment analysis [21], question answering [16], and others. Bahdanau et al. [3] introduced an attention mechanism to automatically select parts of words in a source sentence relevant to predicting a target word, which improves the performance of basic encoder-decoder architecture. Long et al. [21] propose a cognition based attention (CBA) layer for neural sentiment analysis to help capture the attention of words in source sentences. Different from above works focusing on word-level attention, sentence-level [39] and document-level attention [29] pay more holistic attention to the whole textual content. An attention mechanism has also been successfully applied to computer vision tasks like image captioning [36, 40], image retrieval [20], image classification [33], image popularity prediction [43], et al.

Inspired by the great success achieved by attention mechanisms on natural language processing and computer vision, lots of VQA approaches perform attention mechanism to improve model capacity. Current attention methods [35, 38] for VQA mainly perform visual attention to learn image regions relevant to the question. Some recent works [8, 12, 13] integrate effective multimodal feature embedding with visual attention to further improve VQA performance. More recently, Lu et al. [23] design a novel dual attention network which introduces two types of visual features and enables learning question-releted free-form and detection-based image regions. Different from these studies, we propose a novel sequential attention mechanism to seamlessly combine visual and semantic clues for VQA.

2.3 Semantic Facts

Relation facts, standing for relationships between two entities, play an important role in representation and reasoning in knowledge graph. The encoding and applications of relation facts have been widely studied in multiple tasks of knowledge representation [6, 28]. Visual relationship detection is an emerging task aiming to generate relation facts[18, 22], e.g. *(man, riding, bicycle)* and *(man, pushing, bicycle)*, which capture various interactions between pairs of entities in images.

Existing relevant VQA methods involve using knowledge information to either obtain retrieval results of entities and attributes [15, 31, 32], or detect high-level concepts in the image according to the question query [30]. However, it is not effective enough to exploit the complicated semantic relations between the question and image by simply treating semantic knowledge in images as entities and attributes. To the best of our knowledge, it is still rare to incorporate relation facts in VQA to provide rich semantic knowledge. In order to utilize relation facts in the VQA task, we propose effectively learning relation facts and selecting related facts via semantic attention.

3 PRELIMINARY

In this section, we first formulate the VQA problem addressed in this paper, and then clarify the predominant framework for the problem.

3.1 Problem Formulation

Given a question Q and a related image I, the VQA algorithm is designed to predict the most possible answer $\hat{a}$ based on both the language and image content. The predominant approaches formalize VQA as a multi-class classification problem in the space of candidate answer phrases from most frequent answers in training data. This can be formulated as

$$\hat{a} = \arg\max_{a \in \Omega} p(a|Q, I; \Theta), \tag{1}$$

where Θ denotes the parameters of the model and Ω is the set of candidate answers.

3.2 Common Framework

The common frameworks for VQA are composed of three major parts: the image embedding model, the question embedding model, and the joint feature learning model. CNN models like [9, 27] are used in the image model to extract image feature representation. For example, a typical deep residual network ResNet-152 [9] can extract the image feature map v from the last convolution layer before the pooling layer, which is given by:

$$v = \text{CNN}(I). \tag{2}$$

Before being fed into the CNN model, the input image is resized to be 448×448 pixels from the raw image. The convolution feature map extracted from the CNN model has a size of $2048 \times 14 \times 14$, where 14×14 is its spatial size corresponding to different image regions, and 2048 represents the number of feature embedding dimension of each region.

For the question model, recurrent neural networks like Long Short-Term Memory (LSTM) [10] and Gated Recurrent Unit (GRU) [7] are utilized to obtain the question semantic representation, which is given by:

$$q = \text{RNN}(Q). \tag{3}$$

To be specific, given a question with T words, the embedding of each question word is sequentially fed into the RNN model. The final hidden state h_T of the RNN model is taken as the question embedding.

The question and image representations are then jointly embedded into the same space through multimodal pooling, including element-wise product or sum, as well as the concatenation of these representations

$$h = \Phi(v, q), \tag{4}$$

where Φ is the multimodal pooling module. The joint representation h is then fed to a classifier which predicts the final answer.

A large quantity of recent works incorporate visual attention mechanisms for more effective visual feature embedding. In general, a semantic similarity layer is introduced to calculate the relevance between the question and image regions defined as:

$$m_i = \text{sigmoid}(\psi(q, v_i)), \tag{5}$$

where ψ is the module of semantic similarity, sigmoid is a sigmoid-type of function, such as softmax, to map the semantic results to the value interval [0,1], and m_i is the semantic weight of one image region. Finally, the visual representation of the image is updated by the weighted sum over all image regions as:

$$\tilde{v} = \sum_{i=1}^{14\times14} m_i v_i, \tag{6}$$

which is able to highlight the representations of image regions most related to the input question.

4 RELATION FACT DETECTOR

In this section, we describe the process of collecting our Relation-VQA (R-VQA) dataset, as well as the data analysis in Subsection 4.1. We then design a relation fact detector in Subsection 4.2 based on R-VQA to predict visual relation facts related to given questions and images, which is further incorporated into our VQA model in Section 5.

4.1 Data Collection for Relation-VQA

Survey of Existing Datasets Existing VQA datasets like VQA [1] and COCO-QA [26] are made up of images, questions and labeled answers, not involving supporting semantic relation facts. Although the Visual Genome Dataset [14] provides semantic knowledge information such as objects, attributes, and visual relationships about parts of images, which is not aligned with their corresponding question-answer pairs. Therefore, we expand Visual Genome based on semantic similarity and build up the Relation-VQA dataset, which is composed of questions, images, answers, and aligned semantic knowledge. The dataset will be released at https://github.com/lupantech/rvqa .

Data Collection We first define relation facts used in our paper as shown in Table 1. The relation facts are categorized as one of three types: *entity concept*, *entity attribute*, and *entities relation*,

Semantic Knowledge	Fact Templates	Examples
Entity concept	(there, is, object)	(*there, is, train station*)
Entity attribute	(subject, is, attribute)	(*plate, is, white*)
Entities' relation	(subject, relation, object)	(*computer, under, desk*)

Table 1: Types of relation facts.

based on the semantic data of concepts, attributes, and relationships in Visual Genome, respectively. For simplicity, the *attribute* in an entity attribute can be an adjective, noun, or preposition phrase. For Visual Genome, most images are provided with related question-answer pairs, and parts of images are annotated with semantic knowledge. Thus, we keep images containing both question-answer pairs and semantic knowledge, and treat these semantic knowledge as candidate facts with the form of the above templates.

Response Ranking, a semantic similarity ranking method proposed in [37], is then adopted to compute the relevance between each QA pair and its candidate facts. It should be noted that as the ranking algorithm is not our work's main focus and various ranking algorithms are compatible in our framework, we simply adopt one of the state-of-the-art ranking methods, such as Response Ranking. We leave the choice or design of a better ranking algorithm in future work. The relevance matching score obtained from the Response Ranking module ranges from 0 to 1, and value 0 means the candidate fact is completely unrelated to the given QA data, while value 1 means perfect correlation. In the end, after removing candidate facts below a certain threshold matching score, the fact with the largest score is chosen as the ground truth. We randomly partition the generated data into a training set (60%), a development set (20%) and a testing set (20%). Table 2 shows the data sizes of R-VQA with different matching score thresholds.

Score THR	# Train	# Dev	# Test	# Total	# Unique Img.	Match
0.20	286,972	95657	95658	478,287	78,863	75.6%
0.25	207,589	69,196	69,198	345,983	72,993	86.1%
0.30	119,333	39,777	39,779	198,889	60,473	90.8%
0.35	28,345	9,448	9,449	47,242	25,884	91.7%
0.40	24,668	8,222	8,224	41,114	23,480	93.1%
0.45	756	252	253	1,261	1,096	95.0%

Table 2: Basic statistics of the R-VQA dataset.

Human Evaluation To ensure the quality of matched facts, we employ crowdsourcing workers to label whether the generated facts are closely related to the given QA pair. For each generated dataset with a certain threshold score, we randomly sample 1,000 examples for human labeling and ask three workers to label them. The final accuracy for each dataset is the average accuracy obtained by the three workers. Additionally, the workers are encouraged to label every question-answer-fact tuple in more than three seconds. As we can see in Table 2, with the increase of relevance score threshold, the R-VQA dataset has higher matched accuracy, together with a smaller data size. Figure 2 shows two examples in the R-VQA dataset with a score threshold value of 0.30.

Data Analysis To balance the quality of matched facts and quantity of data sample, we compromise by choosing a matched

Q: What sport is the person playing?
A: tennis
R: (*A man, playing, tennis*)

Q: How many animals are there?
A: two
R: (*two horses, stand on, the grass*)

Figure 2: Examples on the R-VQA dataset. For each image-question-answer pair, the dataset provides its aligned relation fact.

Top subjects		Top relations		Top objects		Top Facts	
man	7.80 %	*is*	42.68 %	*white*	6.69 %	*sky, is, blue*	2.22 %
woman	2.86 %	*on*	21.74 %	*blue*	4.08 %	*grass, is, green*	1.75 %
sky	2.81 %	*in*	8.23 %	*green*	4.02 %	*cloud, in, sk*	0.76 %
there	2.35 %	*wearing*	3.18 %	*black*	2.79 %	*plate, is, white*	0.72 %
grass	2.24 %	*holding*	2.71 %	*red*	2.43 %	*train, on, track*	0.59 %
cat	2.12 %	*near*	1.64 %	*brown*	2.27 %	*snow, is, white*	0.47 %
dog	1.85 %	*behind*	1.60 %	*plate*	2.11 %	*tree, is, green*	0.39 %
train	1.50 %	*above*	1.56 %	*table*	2.06 %	*toilet, is, white*	0.37 %
tree	1.45 %	*sitting on*	1.47 %	*wall*	1.61 %	*man, wearing, shirt*	0.35 %
plate	1.41 %	*has*	1.12 %	*water*	1.58 %	*snow, on, ground*	0.34 %

Table 3: Top relation facts in the R-VQA dataset.

score threshold value of 0.30, leading to a dataset of 198,889 samples with an average matched accuracy of 90.8% for all question-answer-fact tuples. There are 5,833, 2,608, and 6,914 unique subjects, relations, and objects, respectively, covering a wide range of semantic topics. In Table 3, we can see the distribution of the most frequent subjects, relations, objects, and facts on the generated R-VQA dataset.

4.2 Relation Fact Detector

The Relation-VQA Dataset provides 198,889 image-question-answer-fact with a matching score of 0.30. That is to say, for each image in the dataset, a question and a correct answer corresponding to the image content are provided, as well as a relation fact well supporting the question-answer data. As stated before, a relation fact describes semantic knowledge information, which benefits a VQA model a lot with better image understanding. For these reasons, we develop a relation fact detector to obtain a relation fact related to both the question and image semantic content. The fact detector will be further expanded in our relation fact-based VQA model, as illustrated in Section 5.

Detector Modeling Given the input image and question, we formulate the fact prediction as a multi-task classification following [18, 19]. For the image embedding layer, we feed the resized image to a pre-trained ResNet-152[9], and take the output of the last convolution layer as a spatial representation of the input image content. Then we add a spatial average pooling layer to extract a dense image representation $v \in \mathcal{R}^{2048}$ as

$$v = \text{Meanpooling}(\text{CNN}(I)). \tag{7}$$

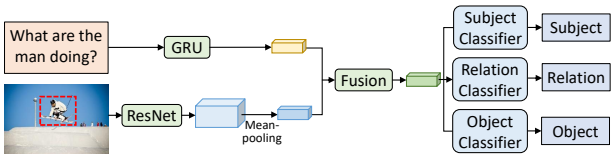

Figure 3: Relation Fact Detector.

The Gated Recurrent Unit (GRU) network is adopted to encode the input question semantic feature as $q \in \mathcal{R}^{2400}$

$$q = \text{GRU}(Q). \tag{8}$$

To encode the image and question in a shared semantic space, the feature representations v and q are fed into a linear transformation layer followed by a non-linear activate function, respectively, as the following equations,

$$f_v = \tanh(W_v v + b_v), \ f_q = \tanh(W_q q + b_q), \tag{9}$$

where W_v, W_b, b_v, b_q are the learnable parameters for linear transformation, and tanh is a hyperbolic tangent function.

A joint semantic feature embedding is learned by combing the image and question embeddings in the common space,

$$h = \tanh(W_{vh} f_v + W_{qh} f_q + b_h). \tag{10}$$

where element-wise addition is employed for the fusion strategy of two modalities. After fusing the image and question representations, a group of linear classifiers are learned for predicting the *subject*, *relation* and *object* in a relation fact,

$$p_{sub} = \text{softmax}(W_{hs} h + b_s), \tag{11}$$

$$p_{rel} = \text{softmax}(W_{hr} h + b_r), \tag{12}$$

$$p_{obj} = \text{softmax}(W_{ho} h + b_o), \tag{13}$$

where $p_{sub}, p_{rel}, p_{obj}$ denote the classification probabilities for subject, relation and object over pre-specific candidates, respectively. Our loss function combines the group classifiers as

$$L_t = \lambda_s L(s, \hat{s}) + \lambda_r L(r, \hat{r}) + \lambda_o L(o, \hat{o}) + \lambda_w \|W\|_2, \tag{14}$$

where s, r, o are target subjects, relations, and objects, and $\hat{s}, \hat{r}, \hat{o}$ are the predicted results. $\lambda_s = 1.0, \lambda_r = 0.8, \lambda_o = 1.2$ are hyper-parameters obtained though grid search on the development set. L denotes the cross entropy criterion function used for multi-class classification. An L2 regularization term is added to prevent over-fitting, and the regularization weight λ_w is set to 1×10^{-7} in our experiment.

Experiments Given an input image and question, the goal of the proposed relation detector is to generate a set of relation facts *subject,relation,object* related to semantic contents of both image and question. The possibility of a predicted fact is the sum of probabilities of the subject, relation, and object in Eqs 11-13. We conduct experiments on the training and development sets for learning, and the testing set for evaluation.

Before carrying out the experiments, some essential operations of data preprocessing are performed. It is observed that there exist some similar and synonymous elements in facts on R-VQA, which may confuse the training of fact detection. For example, "*on*" vs. "*on the top of*" vs. "*is on*", "*tree*" vs. "*trees*", etc. Therefore, we merge these ambiguous elements to their simplest forms based on alias concept dictionaries labeled by [14], e.g., "*on the top of*" and "*is on*" are simplified to "*on*". The merging results are shown in Table 4. We take the most frequent subjects, relations, and objects from all unique candidates in training data, which leads to 2,000 subjects, 256 relations and 2,000 objects, respectively, with more details shown in Table 4.

Operation	Top subjects (2k)		Top relations (256)		Top objects (2k)	
	No.	Perc.	No.	Perc.	No.	Perc.
Before merging	114,797	96.20	115,159	96.50	113,398	95.10
After merging	115,581	96.86	116,008	97.21	113,962	95.50

Table 4: Merging results of the R-VQA dataset for relation detector. After merging similar elements, the top element candidate in relation facts can cover more training data.

Models	Element (Accuracy)			Fact (Recall)		
	Sub.	Rel.	Obj.	R@1	R@5	R@10
V only	3.25	39.19	2.11	0.14	0.43	0.72
Q only	56.66	77.34	40.76	23.14	37.82	43.16
ours - no merge	65.98	74.79	43.61	25.23	44.25	51.26
ours - final	**66.47**	**78.80**	**45.13**	**27.39**	**46.72**	**54.10**

Table 5: Results for the relation detector.

The evaluation metrics we report are **recall@1**, **recall@5**, and **recall@10**, similar to [22]. **recall@k** is defined as the fraction of numbers the correct relation fact is predicted in the top **k** ranked predicted facts. The RMSProp learning method is adopted to train the detector, with an initial learning rate of 3×10^{-4}, a momentum of 0.98 and a weight decay of 0.01. The batch size is set to 100, and dropout strategy is applied before every linear transformation layer.

Results Table 5 shows the experiment results on the R-VQA test set. The first part of Table 5 reports two baseline models, which fully supports that both image and question semantic information is beneficial to relation fact prediction. On the one hand, the model without question content (denoted as **V only**) shows a sharp drop in the accuracy of predicted facts. This phenomenon is intuitive since semantic facts and questions both come from textual modality, while images come from visual modality. In order to improve the semantic space of relation facts, we formulate fact prediction as a multi-objective classification problem, and candidate facts are combinations of three elements, namely a subject, a relation, and an object. Therefore, it is important to provide the question semantic information to reduce the space of candidate facts. On the other hand, the model without image content (denoted as **Q only**) suffers from limited prediction performance, indicating images also contain some useful semantic knowledge.

The second part of Table 5 illustrates that the model based on the merged R-VQA data (denoted as **Ours - final**) works much better than the model based on initial R-VQA data (denoted as **Ours - no merge**). Although existing methods have made good progress in visual detection achieving **Rec@100** accuracy of 10-15% on Visual

Genome for visual facts, these approaches are not suitable to predict question-related visual facts. In contrast with these works, our model incorporates the question feature for fact prediction, and achieves a much higher accuracy with a smaller candidate number of **k**, as well as a much simpler framework. In future work, it will be still meaningful to design a fine-grained model to obtain better prediction performance.

5 VISUAL QUESTION ANSWERING WITH FACTS

The overall framework of our proposed multi-step attention network for VQA is demonstrated in Figure 4, which takes a semantic question and an image as inputs, and learns visual and semantic knowledge sequentially to infer the correct answer. Our proposed network consists of three major components: (A) Context-aware Visual Attention, (B) Fact-aware Semantic Attention, and (C) Joint Knowledge Embedding Learning. Context-aware visual attention is designed to select image regions associated with the input question and to obtain visual semantic representation of these regions. Fact-aware semantic attention aims to weigh detected relevant relation facts by the learned visual semantic representation, and to learn semantic knowledge. Finally, a joint knowledge embedding learning model is able to jointly encode visual and semantic knowledge and infer the most possible answer.

5.1 Context-aware Visual Attention

Similar to many previous VQA approaches [8, 35, 41], we adopt a question-aware visual attention mechanism to choose related image regions.

Image Encoding We apply a ResNet-152 network [9] to extract image feature embedding for an input image. The $2048 \times 14 \times 14$ feature map from the last convolution layer is taken as the image visual feature v, which corresponds to 14×14 image regions with 2048 feature channels.

Question Encoding A gate recurrent unit (GRU) [7] is used to encode the question embedding, which is widely adopted in NLP and multimodal tasks [16, 21, 41]. To be specific, given a question with T words $Q = [q_1, q_2, ..., q_T]$, where q_t is the one hot vector of the question word at position t, we first embed them into a dense representation via a linear transformation $x_t = W_e q_t$. At each time t, we feed the word embedding x_t into the GRU layer sequentially, and the GRU recursively updates the current hidden state $h_t = \text{GRU}(h_{t-1}, x_t)$ with the input x_t and previous hidden state h_{t-1}. Finally, we take the last hidden state h_T as the question representation.

Visual Attention A visual attention mechanism is adopted to highlight image regions related to question semantic information, and to learn more effective multimodal features between textual and visual semantic information. First, we apply the multimodal low-rank bilinear pooling (MLB) method [13] to merge two modalities of the question and image as

$$c = \text{MLB}(q, v). \tag{15}$$

where context vector c contains both question and image semantic content. We map the context vector to attention weights via a linear transformation layer followed by a softmax layer,

$$m = \text{softmax}(W_c c + b_c), \tag{16}$$

where weights m has a size of 14×14, and the value of each dimension represents the semantic relevance between corresponding image region and the input question. The context-aware visual feature is calculated as weighted sum of representations over all image regions, which is given by:

$$\tilde{v} = \sum_{i=1}^{14\times14} m(i) v(i). \tag{17}$$

We further combine the context-aware visual feature with the question feature to obtain the final visual representation as

$$f_v = \tilde{v} \circ \tanh(W_q q + b_q), \tag{18}$$

where $\circ$ denotes element-wise multiplication.

5.2 Fact-aware Semantic Attention

Visual attention enables the mining of visual context-aware knowledge, such as object and spatial information, which is beneficial to questions mainly focusing on object detection. However, models only with visual attention may suffer from limited performance when more relation reasoning is required. Therefore, we incorporate a list of relation facts as semantic clues and propose a semantic attention model to weigh different relation facts for better answer prediction. Some existing studies mine semantic concepts or attributes as semantic knowledge to assist VQA models. Our proposed model differs from these works in two ways. On one hand, existing methods only mine concepts or attributes, while our model extracts relation facts containing concepts and attributes, obviously increasing the semantic capacity of the semantic knowledge used. On the other hand, concepts or attributes in previous works may be irrelevant to VQA, because they are extracted only considering image content and based on data or pre-trained CNN models from other tasks like caption and object recognition [41]. In contrast, we build up the Relation-VQA dataset to train the relation fact detector directly focusing on both the input image and question.

Fact Detection First, we incorporate the fact detector introduced previously in Section 4 into our VQA model. Given the input image and question, the fact detector is used to generate the most possible K relation facts as a candidate set $T = [t_1; t_2; ...; t_K]$. For a fact $t_i = (s_i, r_i, o_i)$, we embed each element of the fact into a common semantic space $\mathcal{R}^n$, and concatenate these three embeddings as the fact embedding as follows:

$$f_{t_i} = [W_{sh} s_i, W_{rh} r_i, W_{oh} o_i] \in \mathcal{R}^{3n}. \tag{19}$$

Then we can obtain the representation of K fact candidate, denoted as $f_T = [f_{t_1}; f_{t_2}; ...; f_{t_K}] \in \mathcal{R}^{K\times3n}$.

Semantic Attention Second, we develop a semantic attention to find out important facts considering the input image and question. Concretely, we use the context-aware visual representation as a query to select significant facts in a candidate set. Similar to context-aware visual attention, given the context-aware visual embedding f_v and fact embedding f_T, we first obtain joint context

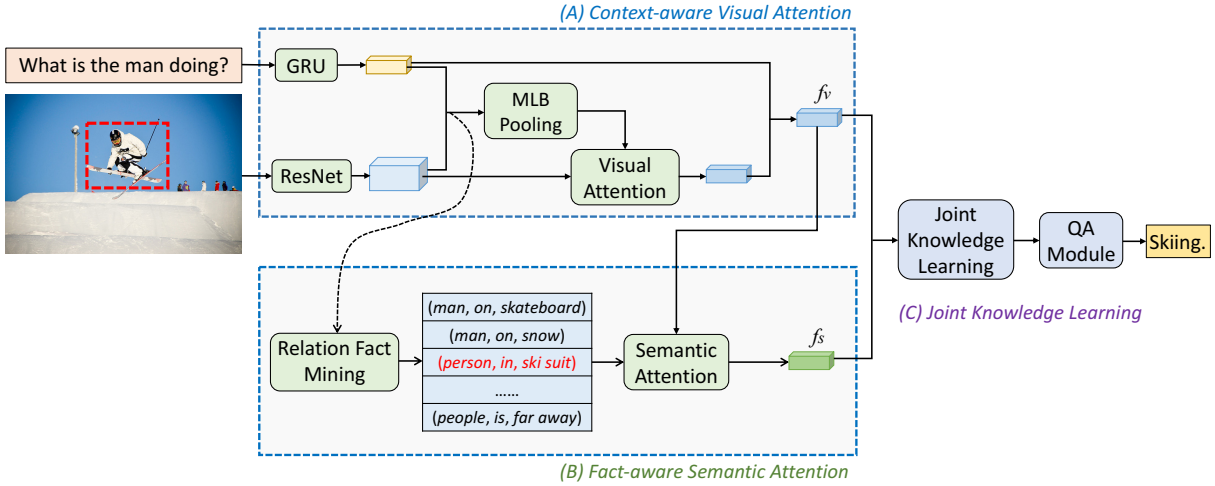

Figure 4: Our proposed multi-step attention network for VQA.

representation c_t and then calculate attention weight vector m_t as follows:

$$c_t = \text{MLB}(f_v, f_T), \tag{20}$$

$$m_t = \text{softmax}(W_{c_t} c_t + b_{c_t}). \tag{21}$$

The final attended fact representation over candidate facts is calculated as

$$f_s = \sum_{i=1}^{K} m_t(i) f_T(i), \tag{22}$$

which serves as semantic knowledge information for answering visual questions.

5.3 Joint Knowledge Embedding Learning

Our proposed multi-step attention model consists of two attention components. One is visual attention which aims to select related image regions and output context-ware visual knowledge representation f_v. Another is semantic attention which focuses on choosing related relation facts and output fact-ware semantic knowledge representation f_s. We merge these two representations via element-wise addition with linear transformation and a non-linear activation function to jointly learn visual and semantic knowledge,

$$h = \tanh(W_{vh} f_v + b_v) + \tanh(W_{sh} f_s + b_s). \tag{23}$$

As we formulate VQA as a multi-class classification task, a linear classifier is trained to infer the final answer,

$$p_{ans} = \text{softmax}(W_a h + b_a). \tag{24}$$

6 EXPERIMENTS

6.1 Datasets and Evaluation Metrics

We evaluate our proposed model on two popular benchmark datasets, th VQA dataset [1] and the COCO-QA dataset [26], due to large data sizes and various question types.

The VQA dataset is annotated by Amazon Mechanical Turk (AMT), and contains 248,349 training instances, 121,512 validation instances and 244,302 testing instances based on a number of 123,287 unique natural images. The dataset is made up of three question categories including *yes/no*, *number* and *other*. For each question, ten answers are provided by different annotators. We take the top 2,000 most frequent answers following previous work [13] as candidate answer outputs, which cover 90.45% of answers in training and validation sets. For testing, we train our model on the train+val set and report the testing result on the test-dev set from a VQA evaluation server maintained by [1]. There are two different tasks, an open-ended task and a multi-choice task. For the open-ended task, we select the most possible answer from our candidate answer set, while for the multi-choice task, we choose the answer with the highest activation score among the given choices.

The COCO-QA dataset is another benchmark dataset, including 78,736 training questions and 38,948 testing questions. There are four question types, *object*, *number*, *color*, and *location*, which cover 70%, 7%, 17% and 6% of total question-answer pairs, respectively. All of the answers in the dataset are single words. As the COCO-QA dataset is smaller, we select all the unique answers as possible answers, which leads to a candidate set with a size of 430.

Evaluation Metric For the VQA dataset, we report the results following the evaluation metric provided by the authors of the dataset, where a predicted answer is considered correct only if more than three annotators vote for that answer, that is to say,

$$Acc(ans) = \min(1, \frac{\#\text{humans vote for } ans}{3}). \tag{25}$$

For the COCO-QA dataset, a predicted answer is regarded as correct if it is the same as the labeled answer in the dataset.

6.2 Implementation Details

For encoding question, the embedding size for each word is set to 620. For encoding facts in the VQA model, the top ten facts are generated and the size of element embedding size m is set as 900. All other visual and textual representations are vectors of size 2400.

Method	Open-Ended				Multi-Choice			
	All	Y/N	Num.	Other	All	Y/N	Num.	Other
LSTM Q+I [1]	53.74	78.94	35.24	36.42	57.17	78.85	35.80	43.41
DPPnet [25]	57.22	80.71	37.24	41.69	62.48	80.79	38.94	52.16
FDA [11]	59.24	81.14	36.16	45.77	64.01	81.50	39.00	54.72
DMN+ [34]	60.30	80.50	36.80	48.30	-	-	-	-
SMem [35]	57.99	80.87	37.32	43.12	-	-	-	-
SAN [38]	58.70	79.30	36.60	46.10	-	-	-	-
QRU [17]	60.72	82.29	37.02	47.67	65.43	82.24	38.69	57.12
MRN [12]	61.68	82.28	38.82	49.25	66.15	82.30	**40.45**	58.16
MCB [8]	64.20	82.20	37.70	54.80	68.60	-	-	-
MLB [13]	64.53	83.41	37.82	54.43	-	-	-	-
V2L [30]	57.46	78.90	36.11	40.07	-	-	-	-
AMA [30]	59.17	81.01	38.42	45.23	-	-	-	-
MLAN [41]	64.60	83.80	**40.20**	53.70	64.80	-	-	-
RelAtt (ours)	**65.69**	**83.55**	36.92	**56.94**	**69.60**	**83.58**	38.56	**64.65**

Table 6: Evaluation results for our proposed model and compared methods on the VQA dataset.

Method	All	Obj.	Num.	Color	Loc.
2VIS+BLSTM [26]	55.09	58.17	44.79	49.53	47.34
IMG-CNN [24]	58.40	-	-	-	-
DDPnet [25]	61.16	-	-	-	-
SAN [38]	61.60	65.40	48.60	57.90	54.00
AMA [32]	61.38	63.92	51.83	57.29	54.84
QRU [17]	62.50	65.06	46.90	60.50	56.99
RelAtt (ours)	**65.15**	**67.50**	**48.81**	**62.64**	**58.37**

Table 7: Evaluation results for our proposed model and compared methods on the COCO QA dataset.

We implement our model with the Torch computing framework, one of the most popular recent deep learning libraries. In our experiments, we utilize the RMSProp method for the training process with mini-batches of 200, an initial learning rate of 3×10^{-4}, a momentum of 0.99, and a weight-decay of 10^{-8}. The validation is performed every 10,000 iterations and early stopping is applied if the validation accuracy does not improve at the last five validations. We use a drop strategy with a probability of 0.5 at every linear transformation layer to reduce overfitting.

6.3 Comparison with State-of-the-art

Table 6 demonstrates our proposed model for both open-ended and multi-choice tasks with state-of-the-arts on the VQA test set. Note that all listed approaches apply only one type of visual feature and the report results of a single model.

The first part in the table shows models using simple multimodal joint learning without an attention mechanism. Models in the second part are based on visual attention, while models in the third part apply semantic attention to learn semantic knowledge like concepts and attributes. It's shown that our proposed multi-step semantic attention network (denoted as **RelAtt**) improves the state-of-the-art **MLAN** [41] model from 64.60% to 65.69% on the open-ended task, and from 64.80% to 69.60% on the multi-choice task. To be specific, our model obtains the improvement of 2.51% in the question types *Other*. As the state-of-the-art model, apart form visual attention, **MLAN** uses semantic attention to mine important concepts based on image content. In contrast, our model **RelAtt** introduces relation facts instead of concepts as semantic knowledge, which obviously increase semantic capacity. Moreover, we train a relation detector to learn facts based on both visual and textual content, instead of only using the image [41]. As our proposed R-VQA dataset extended from Visual Genome dataset shares similar image semantic space with current datasetes like VQA and COCO-QA, semantic knowledge learned from the fact detector can be easily transferred to the VQA task. These are the main reasons that **RelAtt** beats **MLAN** significantly.

Table 7 compares our approach with state-of-the-arts on the COCO-QA dataset. Different from the VQA dataset, COCO-QA doesn't contain the multi-choice task, and fewer results are reported on it. Our model improves the state-of-the art **QRU** [17] from 62.50% to 65.15% with a growth of 2.65%. In particular, our model significantly outperforms the state-of-the-art semantic attention model **AMA** [32] by 3.77%, indicating the benefits of modeling semantic relation facts and learning semantic knowledge from R-VQA dataset.

6.4 Ablation Study

In this section, we conduct five ablation experiments to study the role of individual components designed in our model. Table 8 reports the ablation results of compared baseline models, which are trained on the training set, and evaluated on the validation set. Specifically, the ablation experiments are as follows:

- *Q+I*, where we only take the image and question to infer the answer, and image-question joint representation is learned by a simple fusion method of element wise addition.
- *Q+R*, where only the question and relation facts generated by the detector are considered to predict the answer.
- *Q+I+Att*, where we apply visual attention to learn the joint representation of the image and question.
- *RelAtt-Average*, where the semantic attention mechanism denoted in Eqs 20 - 22 is removed from our best model **RelAtt**. Instead, the fact representation is calculated by averaging different fact embeddings.
- *RelAtt-MUL*, where element-wise addition is replaced by multiplication in Eq 23 to learn the joint knowledge embedding.

The results of first three ablated models indicate that visual attention provides limited visual information for question answering and relation facts can play an important role as they contain semantic information. A drop of 0.79% in accuracy for *RelAtt-Average* illustrates that semantic attention is essential to encode relation facts. Moreover, it is shown that the fusion method of element-wise addition might work better than multiplication when encoding joint visual-textual knowledge representation.

6.5 Case Study

To illustrate the capability of our model in learning relation facts as semantic knowledge, we show some examples on the VQA testing set with the image, question and predicted answer. We also list relation facts generated by the fact detector and their attention weights in the semantic attention component. For saving space,

Q: Does this building have a clock ?
A: yes
R:

(*clock, on, building*)	0.81
(*window*, *on*, *building*)	0.07
(*building*, *has*, *clock*)	0.03
(*clock*, *near*, *tower*)	0.01
(*clock tower*, *near*, *building*)	0.01

(a)

Q: What are the animals standing on ?
A: grass
R:

(*animal, standing on, grass*)	0.27
(*animal, standing on, field*)	0.20
(*zebra, standing on, grass*)	0.19
(*zebra, standing on, field*)	0.14
(*animal, standing on, gazing*)	0.13

(b)

Q: What color is the ball ?
A: yellow
R:

(*ball*, *is*, *yellow*)	0.43
(*tennis ball*, *is*, *yellow*)	0.36
(*ball*, is, *green*)	0.12
(*there*, is, *tennis ball*)	0.05
(*ball*, *in*, *air*)	0.02

(c)

Q: What kind of court is this ?
A: tennis court
R:

(*there, is, tennis court*)	0.52
(*white lines*, *in*, *tennis court*)	0.21
(*line*, *in*, *tennis court*)	0.09
(*grass*, *in*, *tennis court*)	0.08
(*court*, *is*, *orange*)	0.03

(d)

Figure 5: Testing samples on the VQA test set.

Method	Accuracy
Q+I	53.22
Q+R	51.34
Q+I+Att	57.40
RelAtt-Average	57.84
RelAtt-MUL	58.12
RelAtt (final)	**58.63**

Table 8: Ablation study on the VQA dataset.

only five in ten relation facts are shown in Figure 5. In Figures 5 (a) and (b), the fact detector mines semantic fact candidates related to both the image and the question, and semantic attention highlights the most possible facts for question answering. In Figures 5 (c) and (d), although given the same image, the fact detector can depend on the different questions to generate corresponding semantic facts.

7 CONCLUSION

In this paper, we aim to learn visual relation facts from images and questions for semantic reasoning of visual question answering. We propose a novel framework by first learning a relation factor detector based on the built Relation-VQA (R-VQA) dataset. Then a multi-step attention model is developed to incorporate the detected relation facts with sequential visual and semantic attentions, enabling the effective fusion of visual and semantic knowledge for answering. Our comprehensive experiments show our method outperforms state-of-the-art approaches and demonstrate the effectiveness of considering visual semantic knowledge.

ACKNOWLEDGMENTS

We would like to thank our anonymous reviewers for their constructive feedback and suggestions. This work was supported in part by the National Natural Science Foundation of China under Grant No. 61532010 and No. 61702190, in part by the National Basic Research Program of China (973 Program) under Grant No. 2014C B340505, in part by the Shanghai Sailing Program under Grant No. 17YF1404500, in part by the Shanghai Chenguang Program under Grant No. 16CG24, and in part by Microsoft Corporation.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV '15)*. 2425–2433.
[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.
[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR '14)*.
[4] Hedi Ben-Younes, Rémi Cadène, Nicolas Thome, and Matthieu Cord. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *International Conference on Computer Vision (ICCV '17)*.
[5] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD '08)*. ACM, 1247–1250.
[6] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems (NIPS '13)*. 2787–2795.
[7] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Empirical Methods in Natural Language Processing (EMNLP '14)*. 1724–1734.
[8] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Empirical Methods in Natural Language Processing (EMNLP '16)*. 457–468.
[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. 770–778.
[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[11] Ilija Ilievski, Shuicheng Yan, and Jiashi Feng. 2016. A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485* (2016).
[12] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual qa. In *Advances In Neural Information Processing Systems (NIPS '16)*.
[13] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations (ICLR '17)*.
[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123 (2017), 32–73.
[15] Guohao Li, Hang Su, and Wenwu Zhu. 2017. Incorporating External Knowledge to Answer Open-Domain Visual Questions with Dynamic Memory Networks. *arXiv preprint arXiv:1712.00733* (2017).
[16] Huayu Li, Martin Renqiang Min, Yong Ge, and Asim Kadav. 2017. A Context-aware Attention Network for Interactive Question Answering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD '17)*. ACM, 927–935.
[17] Ruiyu Li and Jiaya Jia. 2016. Visual question answering with question representation update (qru). In *Advances In Neural Information Processing Systems (NIPS '16)*.

[18] Yikang Li, Wanli Ouyang, Xiaogang Wang, et al. 2017. Vip-cnn: Visual phrase guided convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*. 7244–7253.

[19] Xiaodan Liang, Lisa Lee, and Eric P Xing. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *IEEE Conference on Computer Vision and Pattern Recognitio (CVPR '17)*. IEEE, 4408–4417.

[20] Guang-Hai Liu, Jing-Yu Yang, and ZuoYong Li. 2015. Content-based image retrieval using computational visual attention model. *Pattern Recognition* 48, 8 (2015), 2554–2566.

[21] Yunfei Long, Lu Qin, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. A Cognition Based Attention Model for Sentiment Analysis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP '17)*. 462–471.

[22] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual Relationship Detection with Language Priors. In *European Conference on Computer Vision (ECCV '16)*.

[23] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. 2018. Co-attending Free-form Regions and Detections with Multi-modal Multiplicative Feature Embedding for Visual Question Answering.. In *The AAAI Conference on Artificial Intelligence (AAAI'18)*. 7218–7225.

[24] Lin Ma, Zhengdong Lu, and Hang Li. 2016. Learning to Answer Questions from Image Using Convolutional Neural Network.. In *The AAAI Conference on Artificial Intelligence (AAAI '16)*.

[25] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*.

[26] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances In Neural Information Processing Systems (NIPS '16)*.

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

[28] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems (NIPS '13)*. 926–934.

[29] Xuejian Wang, Lantao Yu, Kan Ren, Guanyu Tao, Weinan Zhang, Yong Yu, and Jun Wang. 2017. Dynamic attention deep model for article recommendation by learning human editors' demonstration. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD '17)*. ACM, 2051–2059.

[30] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. 2016. What value do explicit high level concepts have in vision to language problems?. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. 203–212.

[31] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2017. Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[32] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. 4622–4630.

[33] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*. 842–850.

[34] Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning (ICML '16)*.

[35] Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision (ECCV '16)*. 451–466.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML '15)*. 2048–2057.

[37] Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, Zhoujun Li, and Jianshe Zhou. 2016. Docchat: An information retrieval approach for chatbot engines using unstructured documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16)*, Vol. 1. 516–525.

[38] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. 21–29.

[39] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL '16)*. 1480–1489.

[40] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. 4651–4659.

[41] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017. Multi-level attention networks for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*. 4709–4717.

[42] Shuangfei Zhai, Keng-hao Chang, Ruofei Zhang, and Zhongfei Mark Zhang. 2016. Deepintent: Learning attentions for online advertising with recurrent neural networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD '16)*. ACM, 1295–1304.

[43] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided Hierarchical Attention Network for Multi-modal Social Image Popularity Prediction. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. 1277–1286.