Multi-source Multi-level Attention Networks for Visual Question Answering

DONGFEI YU, University of Science and Technology of China, China
JIANLONG FU, Microsoft Research, China
XINMEI TIAN, University of Science and Technology of China, China
TAO MEI, JD AI Research, China

In recent years, Visual Question Answering (VQA) has attracted increasing attention due to its requirement on cross-modal understanding and reasoning of vision and language. VQA is proposed to automatically answer natural language questions with reference to a given image. VQA is challenging, because the reasoning process on a visual domain needs a full understanding of the spatial relationship, semantic concepts, as well as the common sense for a real image. However, most existing approaches jointly embed the abstract low-level visual features and high-level question features to infer answers. These works have limited reasoning ability due to the lack of modeling of the rich spatial context of regions, high-level semantics of images, and knowledge across multiple sources. To solve the challenges, we propose multi-source multi-level attention networks for visual question answering that can benefit both spatial inferences by visual attention on context-aware region representation and reasoning by semantic attention on concepts as well as external knowledge. Indeed, we learn to reason on image representation by question-guided attention at different levels across multiple sources, including region and concept level representation from image source as well as sentence level representation from the external knowledge base. First, we encode region-based middle-level outputs from Convolutional Neural Networks (CNNs) into spatially embedded representation by a multi-directional two-dimensional recurrent neural network and, further, locate the answer-related regions by Multiple Layer Perceptron as visual attention. Second, we generate semantic concepts from high-level semantics in CNNs and select those question-related concepts as concept attention. Third, we query semantic knowledge from the general knowledge base by concepts and selected question-related knowledge as knowledge attention. Finally, we jointly optimize visual attention, concept attention, knowledge attention, and question embedding by a softmax classifier to infer the final answer. Extensive experiments show the proposed approach achieved significant improvement on two very challenging VQA datasets.

CCS Concepts: • **Computing methodologies → Computer vision representations**; *Information extraction*; *Computer vision tasks*; Knowledge representation and reasoning;

Additional Key Words and Phrases: Visual question answering, attention model, multi-modal representations, visual relationship

This work was supported in part by the National Key R&D Program of China under Contract No. 2017YFB1002203.
Authors' addresses: D. Yu, Building No 8, West Campus, University of Science and Technology of China, Hefei, Anhui Province, China; email: ydf2010@mail.ustc.edu.cn; J. Fu, Microsoft Research Asia, Microsoft Building No 2, Danling Street, Haidian District, Beijing, China; email: jianf@microsoft.com; X. Tian, Room 1203, Tech Building, University of Science and Technology of China, Hefei, Anhui Province, China; email: xinmei@ustc.edu.cn; T. Mei, JD AI Research 8F, Building A, North-Star Century Center No 8 Beichen West Street, Chaoyang District Beijing 100105, China; email: tmei@live.com.

1551-6857/2019/07-ART51 $15.00
https://doi.org/10.1145/3316767

ACM Reference format:
Dongfei Yu, Jianlong Fu, Xinmei Tian, and Tao Mei. 2019. Multi-source Multi-level Attention Networks for Visual Question Answering. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 2s, Article 51 (July 2019), 20 pages.
https://doi.org/10.1145/3316767

1 INTRODUCTION

Visual question answering (VQA) has been an emerging problem over the past few years and has drawn extensive attention from both the computer vision and natural language processing community. Given an image and a question in natural language, VQA requires grounding textural concepts to visual elements so as to infer the correct answer. The semantic complexity and diversity in questions determined the difficulty of VQA problems, which usually requires reasoning at different level representations of a visual image, such as attributes (e.g., color, shape, size, location), objects, action as well as prior knowledge. The capability of automatic VQA can enhance the connection between visual content and natural language. Toward practical applications, a VQA system can be further deployed in devices for the visually impaired, a personal assistant device, service robots, and so on, in the future.

The challenges of visual question answering are threefold: region-level visual understanding, concept-level semantic embedding, and cross-media knowledge reasoning. Most early works [12, 32, 38] follow similar work in image captioning tasks [7, 33, 43, 44, 52], which combines convolutional neural networks (CNNs) [26] and recurrent neural networks (RNNs) [16]. Specifically, these works utilize a pre-trained CNN model to extract global image representation and an RNN model to extract question representation. The major difference among these works lies in the combination methods of image features and question features. After the multi-modal features are jointly learned, they are further fed into either a decoder RNN to generate free-form answers or a softmax classifier to infer the best answer from a predefined answer set (e.g., 1K answer categories in VQA v1.0 dataset [3]). These early works achieved promising results on large VQA datasets. However, there are still some challenges that limit the further improvement of performance. First, visual question answering requires spatial inference at region level, because some answers can be only inferred from highly localized image regions for "what" and "where" questions. Second, semantic gaps between image and question domain prevent textual concepts to visual elements grounding. Natural language question conveys strong high-level semantics with explicit query intention, while real-world images with tens of thousands of pixels are relatively low level and abstract. Third, beyond pure vision problem, some questions such as "why" require prior knowledge from human experience to infer the answers. Effective knowledge discovering and reasoning remains an open problem in VQA task.

To deal with the challenges, the state-of-the-art approaches follow the research on VQA along two independent dimensions. First, some methods introduce a visual attention mechanism to localize the target regions queried by a question, while masking the irrelevant regions by a learned attention map [11, 21, 31, 34, 41, 50, 51]. However, these attention-based methods ignore the spatial dependency among regions without using the explicit spatial encoding for image regions. Second, others develop the high-level semantic representation for images by introducing semantic concepts, image captions, or even an external knowledge base into the typical CNN-RNN framework [27, 46–48]. However, research in this dimension still ignores using an attention mechanism to select the most discriminative concepts or knowledge for a natural language question, which achieved limited improvement although introducing more data from multiple sources.

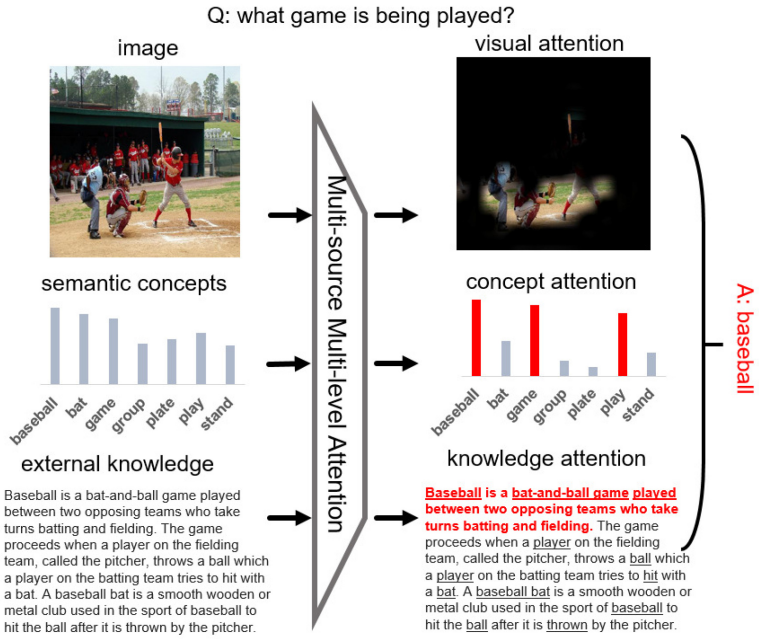

Fig. 1. Overview of the multi-source multi-level attention networks (MSMLAN). The proposed attention model highlights both question-related image regions, semantic concepts, and knowledge descriptions.

To simultaneously learn question-guided image representations across multiple sources and multiple levels, we unify the two dimensions into a holistic learning framework. Specifically, we propose a novel multi-source multi-level attention network (MSMLAN) for visual question answering by highlighting question-related local image regions, semantic concepts and external knowledge in end-to-end training. Figure 1 shows the advantages of the proposed MSMLAN by an intuitive example. The proposed MSMLAN consists of four major components: the first three for image representation across different sources and different levels and the last one for joint learning. First, context-aware spatial attention is proposed to infer the image regions that are queried by questions. Different from previous visual attention-based methods that calculate the attention weight for each region independently, we also incorporate contextual information of each region. Specifically, we first extracted local region representations from convolutional layers in CNN and further fed them into a multi-directional two-dimensional (2D) RNN [14] model by the 2D spatial relationship. Such a design enables spatial information of a region to be encoded from four neighboring regions as the surrounding context. An attention score for each region is further obtained by a multiple layer perceptron (MLP) with the input of both context-aware visual representation and question representation. Second, concept attention focuses on concept-level image representation by discovering the semantically close concepts to questions in the same vocabulary set and joint embedding space. These concepts correspond to highly frequent words in question–answer pairs and can represent high-level understanding for image content. Specifically, a CNN-based recognizer is trained for each concept, and the distribution over the semantic output layer in CNN constitutes the high-level representation of an image. Third, knowledge attention focuses on question-related knowledge discovering and reasoning by connecting concepts in an image to an external knowledge base (such as Wikipedia) and distilling those pieces of knowledge helpful to answer the question. Knowledge from external sources provides common sense well

known for humans, which is necessary to infer the answers for some questions, especially those starting with "why." We encode each sentence in the knowledge base as a fact and give each fact different weight according to its relevance with the image and the question. Fourth, joint learning incorporates attended regions, concepts, knowledge, and question features by element-wise multiplication, followed by a softmax layer to predict the most possible answer from the answer set.

The preliminary version of this work was published in Reference [54]. In this article, we improve the original model in two aspects. On one hand, we introduce an attention network to an external knowledge domain, which enables our model to reason across multiple sources and more levels. However, we propose to encode contextual information of image regions with 2D RNN instead of bidirectional RNN, which effectively captures the cross-region relations in two dimensions and four directions. Experiments show that the 2D RNN outperforms the original bidirectional RNN and improves significantly over visual attention without context encoding. Besides, we also include more experiments on the balanced VQA 2.0 dataset, and more ablation models are implemented and studied.

We summarize the main contributions as follows:

- We address the challenges of automatic visual question answering by jointly learning multi-source multi-level attention, which can simultaneously benefit region-level spatial inference, reduce the semantic gap from vision to language, and effectively incorporate the knowledge-based reasoning into the data-driven learning framework in the VQA task.
- We introduce a novel spatial encoding approach for visual attention, which extracts the context-aware visual features from image regions by a multi-directional 2D RNN model.
- We conduct experiments on two challenging VQA datasets [3, 13] and obtain top performance with a simple and effective framework.

The rest of the article is arranged as follows. In Section 2, we review the related work on visual question answering, visual attention, knowledge representation, and structured representation. In Section 3, we introduce our proposed multi-source multi-level attention networks and detail its major components. In Section 4, we describe the experiments we have made including the datasets we used, model training details, ablation study, results comparison and analysis in two datasets, and some visualization cases for qualitative study. In Section 5, we conclude our proposed methods and present future direction.

2 RELATED WORK

In this section, we first introduce the general CNN-RNN framework on both image captioning and visual question answering. Then, we summarize the most recent advances from two different dimensions. Besides, we will introduce a structured representation for contextual modeling in images.

CNN-RNN. In past decades, CNN as an automatic feature extraction method has achieved great success for various vision tasks, from image tagging [9, 10, 19] and retrieval [17, 18] to image captioning [36, 37, 52] and visual story telling [30] in the computer vision community. RNN have been widely explored for natural language understanding due to its simple and effective structure and end-to-end training methods. As a kind of vision-to-language task, CNN-RNN framework has been well studied in image captioning task and made breakthrough advances.

Inspired by the success of CNN-RNN framework in image captioning task, most early works tend to exploit variation of those models to visual question answering task [3, 12, 32, 38]. They extract visual features from images via pre-trained CNNs and encode questions by RNNs. Ren et al.

[38] took their inspiration from [43], where the image was treated as the first token and fed into RNNs together with descriptions to learn visual-semantic embedding. Instead of seeing an image once, Malinowski et al. [32] passed the image into RNNs at each time when encoded the question, which is similar to the framework in Reference [7] in the automatic image captioning task. Gao et al. [12] adapted m-RNN models [33] to deal with a VQA task in a multi-lingual setting. Agrawal et al. [3] released a large and human-annotated VQA dataset and evaluated several baseline models and human-level performance on this dataset, which accelerated advances in this task. Although these early approaches showed promising performance in the VQA task, it tended to fail on novel instances and highly relied on questions (did not change the answer across images) [1]. Reference [23] proposed a new dataset called the Task Driven Image Understanding Challenge (TDIUC), which has over 1.6 million questions organized into 12 different categories. They proposed new evaluation schemes that compensate for over-represented question -types and make it easier to study the strengths and weaknesses of algorithms.

Visual Attention. A visual attention mechanism is brought into VQA to address "where to look" problems. Question-guided visual attention uses the semantic representation of a question as the query to search for the regions in an image that are related to the answer [11, 21, 31, 41, 51]. Two types of soft attention mechanism are well explored in visual question answering tasks. The first type concatenates the question representation with each candidate region and then puts them into a multiple layer perceptron (MLP) to compute the soft attention weights, while the second type gets the attention score by the dot product of the two types of inputs [50]. Yang et al. [51] proposed a stacked attention model that queries the image multiple times to infer the answer progressively. Lu et al. [31] exploited a question–image co-attention strategy to handle not only related regions in images but also important words in questions. Recently, Nam et al. [34] proposed Dual Attention Networks, which refined the visual and textual attention via multiple reasoning steps. Our work is different from co-attention and dual attention in that we attend to high-level concepts extracted from the image rather than words from questions. The major advantage of using concepts over questions is that the concepts are the semantic representation of content in the image, not limited to words in the question. Another related work is the Focal Visual-Text Attention network proposed by Liang et al. [28], which aims to answer questions based on whole collections with sequences of photos or videos. They make use of a hierarchical process to dynamically determine what media and what time to focus on in the sequential data to answer the question. The main difference between their work and ours lies in attention strategies, where they apply temporal attention in the sequential data including image sequences and metadata around them, while our multi-source multi-level attention networks apply visual attention on 2D structured images and textual attention on the unordered concept or sentence collection.

High-level Concepts and Knowledge. There is another branch that shows a promising direction to address VQA problems. Instead of low-level visual features, they leverage high-level concepts [9, 10, 45], candidate answers [22], image captions or even a visual story [30], and an external knowledge base [46–48]. Each concept corresponds to a word mined from the training image descriptions and represents some kinds of attribute regaring the content of the image. These concepts act as semantic units between natural language and visual recognition and allow us to exchange information between the two modalities [39]. The external knowledge is mined from a general knowledge base, which is queried by the most relevant concepts found in the image. The discovered knowledge is encoded using Doc2Vec, which produces a fixed-length vector. However, the concepts generated from images and the documents queried from the knowledge base are usually noisy, and most information is irrelevant with the query intention of questions. Therefore, they only gain limited improvement, though concepts and external knowledge are incorporated into the

VQA framework. Besides, the spatial information is completely lost in the procedure of high-level concepts detection, which leads to inferior performance on a VQA task.

Structured Representation. In most previous attention-based models, the activation of the last convolutional layer of CNN was taken as the visual representation of an image. However, the spatial dependency among regions is neglected, which weakens the spatial reasoning of their models. Our previous work [54] and DMN [49] fed the convolutional feature map in a snakelike fashion to a bidirectional Gated Recurrent Units (GRU) layer to model the contextual information of individual regions, which may not be an optimal choice for 2D spatial structure in an image. More recent work [40] tried to build a fully connected network with the pairwise connection between regions for spatial reasoning in a VQA task. However, it is hard to scale due to the quadric complexity with respect to the number of candidate regions. Also, the pairwise relationship may be insufficient for reasoning over multiple objects. Recently, Hu et al. [20] proposed relation networks for object detection, which processes a set of objects simultaneously through interaction between their appearance feature and geometry, thus allowing the modeling of their relationships. It verifies the efficacy of modeling object relationships in CNN-based detection. The most similar work with ours is proposed in Reference [55], which tries to learn a multivariate distribution over a grid-structured Conditional Random Field on image regions as the structured representation. They implement the interactive approximate inference algorithms, as recurrent layers of neural networks, which iteratively refines the attention. Instead, we propose to feed the grid-structured feature map extracted from CNNs into 2D GRUs, which capture the long-term dependency and are easy to train.

3 MULTI-SOURCE MULTI-LEVEL ATTENTION NETWORKS

To simultaneously exploit spatial information, concept-level semantic information, and external knowledge, we propose a novel multi-source multi-level attention network. The overall framework is presented in Figure 2. Our framework consists of four major components. Component (A), which is defined as context-aware visual attention, aims at finding question-related regions and learning a visual representation of these regions. Component (B), which is defined as concept attention, aims at finding question-related concepts from the image. Component (C), which is defined as knowledge attention, aims at selecting question-related knowledge from the external knowledge base. Component (D) is designed to incorporate information from different-level representation of the image by joint attention learning. These four components are jointly optimized end to end, which learns spatial reasoning on image regions and knowledge-based reasoning across multiple sources and multiple levels.

3.1 Context-aware Visual Attention

Some questions in a VQA task focus on some specific regions rather than the whole image. For instance, the question "What color is the apple on the desk" is only about the color of the local region containing apples on the desk, while other regions should be masked as background. Visual attention has been widely used in recent VQA frameworks, due to its success on fine-grained visual representation and visualization for explainability. Compared with human attention, recent work [6] found that current VQA attention models do not seem to be "looking at" the consistent regions as humans do. One of the possible problems in current attention models is that they usually search for image regions independently by dividing the whole image into several isolated units. Although promising results have been achieved, further improvements are limited, because many concepts may interact with each other through the spatial and semantic relationship. For example, we should be aware of the spatial relationship of the cat and the toilet if we want to really understand and answer the question "What is the cat standing on." In this case, not only regions about "cat" but

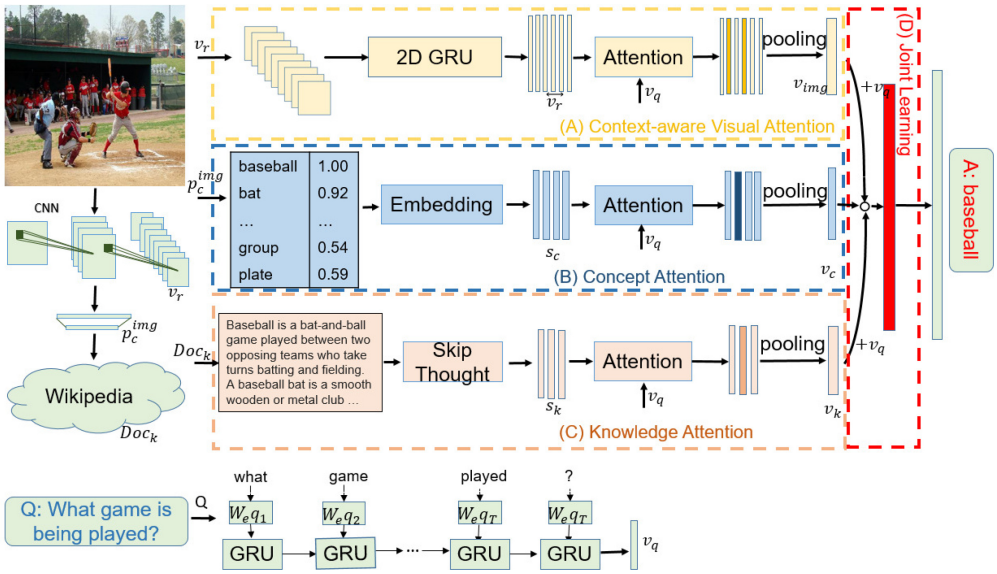

Fig. 2. Overall framework of Multi-source Multi-level Attention Networks. Our framework consists of four components: (A) context-aware visual attention, (B) concept attention, (C) knowledge attention, and (D) joint attention learning. Here, we denote by v_q the representation of the question Q and by v_{img}, v_c, and v_k the representation of image content on the region, concept, and knowledge level queried by the question, respectively. v_r and p_c^{img} is the activation of the last convolutional layer and the probability layer from the CNN.

those regions at bottom of the "cat" should be looked at and understood. To address this issue, we propose a context-aware visual attention mechanism in our VQA framework.

First, we incorporate the contextual information into the representation from each region by a multi-directional 2D GRU encoder, which is illustrated in Figure 3. Multi-Dimension RNN (MDRNN) was first proposed in Reference [14], which is implemented on the raw image pixels for image segmentation. Different from their application on small simple tasks, our proposed 2D GRU layer goes after the pre-trained CNN and takes the feature map as input. The 2D GRU module is a natural extension of GRU, where the main difference lies in the dimension of inputs and hidden states. In 2D GRU, the inputs and hidden states are a fixed-length vector arranged in a 2D array. For traditional GRU, there is one update gate and one reset gate. However, for 2D GRU, there are two update gates and two reset gates, which control information transmission in two orthogonal dimensions, respectively. Specifically, we use the pre-trained CNN model to extract visual features of local regions. We take the feature map of the last convolutional layer in the CNN model as our visual representation, which can preserve complete spatial information of each region. These visual representation on each region is denoted as $\{v_{i,j}, i = 1, 2, \ldots, M, j = 1, 2, \ldots, M\}$, where $v_{i,j}$ represents the feature vector of the region at (i, j). Then, we feed these feature vectors into the 2D GRU and concatenate the outputs from four different directions (left-top to right-bottom, left-bottom to right-top, right-top to left-bottom, and right-bottom to left-top) at each step to form a new feature vector for each region, which is given by:

$$\overleftrightarrow{v_{i,j}} = [2DGRU^{lt}(v_{i,j}), 2DGRU^{lb}(v_{i,j}), 2DGRU^{rt}(v_{i,j}), 2DGRU^{rb}(v_{i,j})], \quad (1)$$

where $\overleftrightarrow{v_{i,j}}$ is the context-aware visual representation of image region r. The new feature vectors contain not only the visual information of corresponding regions but also the contextual

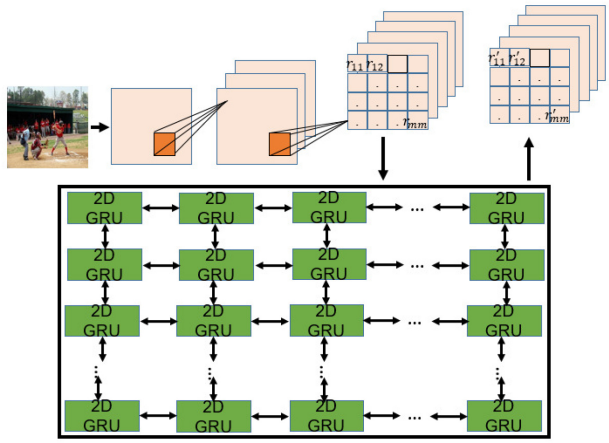

Fig. 3. An illustration of the context-aware visual representation for image regions by multi-directional 2D GRU. Regions in convolutional feature maps are encoded into GRU and outputs from four different directions are concatenated as the final outputs. Thus, each region is able to interact with its four-nearest-neighbor regions and incorporates contextual information into the visual representation. $v_{i,j}$ is the feature vector of location (i, j) in the original feature map.

information from surrounding regions. 2D GRU is the two-dimensional GRU, and the superscript symbols indicate the starting direction of propagation during the forward pass, i.e., lt, lb, rt, and rb represent left-top, left-bottom, right-top, and right-bottom, respectively. We formulate the 2D GRU in left-top direction as follows:

$$z_{i,j} = \sigma(W_z v_{i,j} + U_{z1} h_{i-1,j} + U_{z2} h_{i,j-1} + b_z), \tag{2}$$

$$z'_{i,j} = \sigma(W'_z v_{i,j} + U'_{z1} h_{i-1,j} + U'_{z2} h_{i,j-1} + b'_z), \tag{3}$$

$$r_{i,j} = \sigma(W_r v_{i,j} + U_{r1} h_{i-1,j} + U_{r2} h_{i,j-1} + b_r), \tag{4}$$

$$r'_{i,j} = \sigma(W'_r v_{i,j} + U'_{r1} h_{i-1,j} + U'_{r2} h_{i,j-1} + b'_r), \tag{5}$$

$$\tilde{h}_{i,j} = \tanh(W_h v_{i,j} + U_h(r_{i,j} \circ h_{i-1,j} + r'_{i,j} \circ h'_{i,j-1}) + b_h), \tag{6}$$

$$h_{i,j} = z_{i,j} \circ h_{i-1,j} + z'_{i,j} \circ h_{i,j-1} + (2 - z_{i,j} - z'_{i,j}) \circ \tilde{h}_{i,j}, \tag{7}$$

where $h_{i,j}$ is the output of current 2D GRUs with $v_{i,j}$ as input. The variables $z_{i,j}$, $z'_{i,j}$ are update gates for vertical dimension (from $i-1$ to i) and horizontal dimension (from $j-1$ to j), while $r_{i,j}$, $r'_{i,j}$ are reset gates correspondingly. We set the dimension of the hidden state in each GRU to the same with the question vector. For 2D GRU from left-top to right-bottom direction, the hidden state $h_{i,j}$ depends on current input $v_{i,j}$ and two hidden states $h_{i,j-1}$, $h_{i-1,j}$ from left and top region, respectively. 2D GRU in other directions (i.e., from left-bottom to right-top, from right-top to left-bottom, and from right-bottom to left-top) can be formulated in a similar way but with a different subscript index.

Second, we assign each region an attention score for modeling the relationship between the region and the question. We align the question and each region by element-wise multiplication of two vectors and then feed them into a multiple layer perceptron (MLP). Such a design enables the automatic learning of an attention function by parameter optimization in MLP. More specifically, we search for regions via multi-step reasoning as in Reference [51]. The main differences are twofold. First, we use a context-aware visual feature obtained in the last step to represent local regions rather than the independent representation from each region in convolutional neural

networks, which often lacks interactions between different regions. Second, we use element-wise multiplication instead of element-wise addition to align the question feature and visual feature for each region, which overcomes the scale inconsistency problem in multi-modal feature pooling. The comparison experiment in Section 4.4 demonstrates our assumption.

At first, we represent the question by a recurrent neural network. Specifically, given the question $Q = [q_1, q_2, \ldots, q_T]$, where q_t is the one hot vector representation of word at position t, we embed these words into a vector space through an embedding matrix W_e^q. At time step t, we feed the embedding vector x_t of word q_t to a GRU layer and pick the last hidden state h_T as the question representation, which is denoted as v_q. We use the following equation to formulate the question encoding model:

$$x_t = W_e^q q_t, \tag{8}$$

$$h_t = GRU(x_t, h_{t-1}), \tag{9}$$

$$v_q = h_T. \tag{10}$$

Next, we formulate our visual attention process as:

$$h_{i,j} = \tanh\Big((W_Q v_q + b_Q) \otimes (W_I \overleftrightarrow{v}_{i,j} + b_I)\Big), \tag{11}$$

$$M_{i,j} = \text{softmax}(W_p h_{i,j} + b), \tag{12}$$

where we denote $\otimes$ as the multiplication between a matrix and a vector, which is performed by element-wise multiplying each column of the matrix by the vector. W_Q and W_I are the corresponding embedding matrix. W_p is the parameter in multiple perceptron layers, and $M_{i,j}$ is the attention weights of image regions.

Finally, we pool these regions with a weighted sum to get the visual representation of image I queried by question Q, which is given by:

$$v_{img} = \sum_{i=1,j=1}^{m} M_{i,j} \overleftrightarrow{v}_{i,j}. \tag{13}$$

In practice, we repeat the above process once, as in Reference [51], using the addition of a question feature and an attended region feature as a guide. We ignore the details here for concision.

3.2 Concept Attention

Concept attention aims at finding important concept mining from the image to answer a question. For example, in Figure 1, although the concept detector has detected a set of objects and actions from the image (e.g., "group," "stand"), only those concepts that are semantically close to the question (i.e., "baseball," "game") should be highlighted by semantic attention. One of the core challenges in combining visual and linguistic modalities is that they have different levels of abstraction, where language usually refers to general categories, while hundreds of pixels in the image can point to one instance [39]. Previous works on image/video captioning [8, 36, 37, 52, 53] and visual question answering [46, 48] have shown that extracting explicit high-level concepts from images/videos can bring benefits to the interaction of visual content and language at the semantic level. Although an image can convey multiple semantics, not all of them are helpful to answer a particular question. Therefore, we propose to attend to concepts that should be not only relevant to images but also semantically close to questions. We achieve these goals through two steps.

In the first step, we train a concept detector by deep convolutional neural networks, which can produce the probability of semantic concepts for an image. Similarly to Reference [46], we first build a concept vocabulary, where each concept is defined as a single word. The top high-frequency words with the number of C from the question–answer training pairs are collected in

the concept vocabulary after the removal of stop words. Besides, a multi-label image dataset based on these concepts is constructed based on a COCO image captioning dataset [29], which is used to train the concept detector. As a result, a fixed-length vector p^c_{img} is created for each image I by taking the activation of f_c in the prediction layer of a CNN, which represents the probability of each concept occurring in the image. We denote the process of concept detection as:

$$p^c_{img} = f_c(I). \tag{14}$$

In the second step, we train an attention network to measure the semantic relevance between each concept in the vocabulary and the question. We share the same question features used in visual attention. Besides, we use the same vocabulary and embedding matrix for our concepts and questions, and, therefore, they can share the same semantic representation. Specifically, we represent the concept c with a semantic vector s_c by a two-layer stacked embedding layer. The first layer is designed to share the same word embedding layer as the question model, and the second layer is used to project the concept vector into the same dimension with the question representation, which is given by:

$$s^c = W^c_e(W^q_e c), \tag{15}$$

where c is the one hot vector representation of the concepts, W^q_e is the embedding weights shared with the question model, and W^c_e is the second embedding matrix, which embeds the concepts into the same dimension representation with the question. Next, we take the dot product of the projected concept vector s^c with the question vector v_q as an operation and pass the resultant value to a sigmoid activation layer to get the relevance score between the concept c with question Q. Further, we formulate the concept attention weights of the concept c as a multiplication of the concept-image relevance p^c_{img} and the concept–question relevance p^c_q, which is given by:

$$p^c_q = \text{sigmoid}(v_q \cdot s^c), \tag{16}$$

$$M^c = p^c_{img} p^c_q, \tag{17}$$

where the operator $\cdot$ represents the dot product of two vectors, p^c_q is the relevance score measuring the semantic similarity between the question Q and the concept c, and M^c is the concept attention weights over concepts. Finally, we represent the high-level semantic information of image I queried by question Q by a weighted sum over all concepts representation, which is given by:

$$v_c = \sum_{i=1}^{C} M^c_i s^c_i. \tag{18}$$

3.3 Knowledge Attention

Inspired by previous works [47, 48], we extend our multi-level attention networks to higher-level knowledge from an external source, which enables knowledge-based reasoning. We use Wikipedia as our external knowledge source. Though we only access the article summaries as the knowledge documents, they are usually general and noisy. We perform sentence-level attention on knowledge documents to extract the most useful descriptions to answer the question while filtering out the irrelevant information. In practice, we implement knowledge attention in three steps: knowledge extraction, knowledge encoding, and knowledge attention.

We first query the Wikipidia database by the top five concepts generated from the image with a popular API tool in the python library.[1] For each concept, we extract the summary part of the corresponding Wikipedia article as the knowledge documents. A set of sentences can be obtained in

[1] https://pypi.org/project/wikipedia/.

this step, which is denoted as S. Instead of Doc2Vec to model semantic meanings in the paragraph level, we turn to a pre-trained Skip Thought model [25] (denoted as Sent2Vec) on Wikipedia to encode knowledge in the sentence level. The knowledge encoding process can be simply formulated as:

$$s_i^k = Sent2Vec(i), \tag{19}$$

where s_i^k is the encoded vector of ith knowledge sentence in queried knowledge set S. Further, we use a similar formulation with concept attention on the knowledge vectors as knowledge attention,

$$M_i^k = \text{sigmoid}(v_q \cdot s_i^k), \tag{20}$$

$$v_k = \sum_{i=1}^{S} M_i^k s_i^k, \tag{21}$$

where M^k is the knowledge attention weight over the extracted knowledge sentence, and v_k is the attended knowledge vector, which can be considered as knowledge-level representation of image.

3.4 Joint Attention Learning

We use questions as the query to search for image information on different levels. In the low-level visual feature, we focus on question-related regions by visual attention, while in the high-level semantic feature, we focus on question-related concepts by concept attention and question-related knowledge by knowledge attention. The three-level attention is combined through the fusion of their attended representation. Particularly, we first add question vector into attended image features extracted from different layers, then we use an element-wise multiplication to combine the three types of attention together. Finally, we feed the joint feature into a softmax layer to predict the probability of predefined candidate answer set A. The candidate with the highest probability is determined as the final answer, which is given by:

$$u = (v_q + v_{img}) \circ (v_q + v_c) \circ (v_q + v_k), \tag{22}$$

$$p_a = \text{softmax}(Wu + b), \tag{23}$$

where we denote $\circ$ as the element-wise multiplication between two vectors; v_q, v_{img}, v_c, and v_k are the representation of question Q, the attended visual representation of image I, attended semantic representation of concept C, and attended semantic representation of knowledge S, respectively. u is the joint representation from question and image across multiple levels and multiple sources. W and b are the parameters of the last full connected layer, and p_a is the output of the softmax layer, i.e., the distribution of probability of answer candidates. The candidate with the maximum probability is picked out as the predicted answer.

4 EXPERIMENT

4.1 Dataset

We evaluate our model on two large-scale VQA datasets, i.e., the VQA v1.0 and VQA v2.0 datasets, due to the large amount of training instances and the diversity of question types.

VQA v1.0 is a large-scale visual question answering dataset that contains 204,721 images from the COCO dataset and a newly created abstract scene dataset that contains 50,000 scene images. We evaluate our model on this dataset for only real images. For each image in the VQA dataset, three questions are annotated, and each question has 10 answers from 10 different annotators. We only report our results in the open-ended task, which is more challenging. In the open-ended task, we select the answer with the highest activation from all possible outputs. We collect the most frequent 3,000 answers in the training data as the candidate answer set. We evaluate the

proposed approach not only on validation dataset but also on a test server, which is provided for blind evaluation in the test set for fair comparison [3].

VQA v2.0 [13] is the second version of VQA dataset, which is more balanced in answer distribution than the first version. To solve the language bias problem in the VQA task, each question is associated with a pair of similar images that results in two different answers to the question. In this way, it is harder to infer the answer for VQA models without observing the image. The images in VQA v2.0 are totally the same with VQA v1.0, but the number of questions is about twice of VQA v1.0. We use same setting to preprocess the data and train models on the VQA v2.0 dataset.

4.2 Evaluation Metrics

Visual QA is formulated as a multi-class classification problem on both datasets. We follow the evaluation metrics as the baseline approaches on the two datasets. For the VQA dataset, Reference [3] set an evaluation server publicly for blind evaluation on the test set. The test set is divided into four splits: test-dev, test-standard, test-challenge, and test-reserve, each of which contains about 20K images. We evaluate our ablation model for experiment analysis on the test-dev set of VQA v1.0 and evaluate our best model on the test-dev and test-standard set of both the VQA v1.0 and VQA v2.0 datasets. For the open-ended task, Reference [3] used a voting mechanism to score the accuracy of a predicted answer:

$$acc(ans) = min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\},$$

where *ans* is the answer predicted by visual QA models. We use the same evaluation metrics in both the VQA v1.0 and VQA v2.0 datasets.

4.3 Experiment Setting

We show our experimental settings, hyper-parameters, and training process here.

For the question model, we use the natural language toolkit NLTK[2] to tokenize questions, cast all words into lowercase, and only keep those words appearing at least twice in the train-val set. We do not make any additional preprocessing to those words, e.g., removing stop words and stemming. A two-layer GRU is used to encode the question, which has 620-dimensional word vectors and 2,400-dimensional hidden states. We initialize all weights in the GRU layer from scratch. We take the last hidden state of the GRU layer as the question representation so that the dimension of the question feature vector is 2,400.

For the image model, we extract visual features from the last convolutional layer (i.e., "res5c") of the pre-trained ResNet-152 model by feeding resized images with 448×448 resolution. Each feature vector has a dimension of 2,048 and corresponds to a 32×32 pixels region of the input image. Thus, we get a feature map of $14 \times 14 \times 2{,}048$ dimensions for each input image. The dimension of hidden state in 2D GRU is set to the same with the question feature, i.e., 2,400. We get a $14 \times 14 \times 2{,}400$ dimensional contextual aware feature map after passing the original feature map into the 2D GRU module. Then we project both the question feature and context-aware visual representation of the image into the same 512-dimensional space and perform attention on visual representation.

For the concept model, we select the most frequent 256 words appearing in question–answer training pairs as our concept vocabulary after removing stop words. A multi-label dataset is derived from the MSCOCO dataset based on the concept vocabulary and captions associated with the image. Further, we train a multi-label classification model with ResNet-152 [15] on the multi-label

[2] http://www.nltk.org/.

dataset. We detect concepts from images by taking the activation of the last layer of the multi-label classification network. There are two major differences in our concept detector from Reference [46]. We use a more powerful classification model, i.e., ResNet with 152 layers pre-trained on ImageNet, instead of VGGNet with 19 layers [42]. Besides, we use the popular loss function "SigmoidCrossEntropyLoss" in the multi-label classification task to fine-tune the network. For each concept, we get the same embedding vector with the same question word, i.e., 2,400 dimensions. Similarly to the image model, we project the question vector and concept vector to the 512-dimensional space and then perform attention on concepts.

For the knowledge model, the top 5 concepts generated in the concept model are queried in the knowledge base (i.e., Wikipedia). For each concept, we crawl the top 5 sentences from the summary of the corresponding Wikipedia articles. Therefore, we get a total of 25 sentences for each image for knowledge representation. As mentioned in Section 3.3, each sentence is encoded into a 2,400-dimensional vector with the pretrained Skip-Thought model. Further, we the project question vector and knowledge vector to the 512-dimensional space and perform attention on knowledge.

We sum pool the visual feature, concept feature, and knowledge feature respectively weighted with their attention scores as attended representation of the image at different levels. In this way, we get three 2,400-dimensional vectors for multi-level image representation as well as the same dimensional question vector. These feature are joint embedded in the form of Equation (22) and fed into a softmax layer to predict the answer.

In our experiments, we use stochastic gradient descent with momentum 0.9 as the solver. The batch size is fixed to 100. We set the base learning rate to 0.05. After 15 epochs, we drop the learning rate to 1 of 10 of the previous one every 5 epochs. In addition, gradient clipping technology and dropout are exploited in training. To further promote the performance of our model, we also pretrain our model on the Visual Genome [5] dataset and then fine-tune the model on the VQA dataset. We keep the same question and answer vocabulary in Visual Genome dataset with the one in VQA dataset.

4.4 Ablation Model

To analyze the contribution of each component in our model and demonstrate how the multi-source multi-level attention works better than single-level attention, we ablate the full model and demonstrate the effectiveness of each component.

- Q + I [3]: a baseline model with no attention.
- Q + Visual Att.: our visual attention model without context-aware visual representation
- Q + Context-aware Visual Att. (BiGRU): context-aware visual representation with Bidirectional GRUs for context modeling, used in our previous work [54].
- Q + Context-aware Visual Att. (2DGRU): component (A) of our model
- Att-CNN + LSTM [46]: the attribute representation as the first input of LSTM, then following the question
- Q + Concept: a simple version of concept attention, taking the output of concept detector as the attention weights, independent on the question
- Q + Semantic Att.: component (B) of our model, taking the relation of concepts with both image and question into the attention weights
- Q + Knowledge: remove attention parts of component (C)
- Q + Knowledge Att.: component (C) of our model
- Q + Concept Att. + Knowledge Att.: component (B) + component (C) of our model

Table 1. Ablation Model on the Test-dev Split of VQA v1.0 Dataset

Ablation Model	Accuracy
Q + I [3]	57.75
Q + Visual Att.	62.29
Q + Context-aware Visual Att. (BiGRU)	62.50
Q + Context-aware Visual Att. (2DGRU)	62.84
Att-CNN + LSTM [46]	55.57
Q + Concept	56.62
Q + Concept Att.	59.28
Q + Knowledge	56.31
Q + Knowledge Att.	57.46
Q + Concept Att. + Knowledge Att.	60.78
MLAN [54]	63.69
MSMLAN w/o Knowledge Att.	64.05
MSMLAN	**64.43**

The top four models only utilize visual features from convolution layer, the next three models only utilize the concepts instead of visual features and then followed by three knowledge-based models. MLAN is our previous work [54] that applies attention to the multi-level representation of images. MSMLAN is our extended work that applies attention to the multi-source multi-level representation of images. MLAN w/o Knowledge Att. removes attention part from knowledge attention module (C) in the full model.

- MLAN [54]: component (A) + component (B) of our model, the final model proposed in our previous work [54]
- MSMLAN w/o Knowledge Att.: MSMLAN but remove attention part in component (C)
- MSMLAN: the final model in this work, the fusion of Visual Att, Concept Att, and Knowledge Att.

We report the performance of our ablation models on the test-dev set of the VQA v1.0 dataset in Table 1. These models are trained on the training dataset and half of the validation set, as in Reference [51]. Further analysis will be given in the next section.

4.5 Results and Analysis

We will explain how each component works in our model by the ablation experiment shown in Table 1. It is observed that our multi-source multi-level attention model outperforms all single-level attention model significantly, i.e., attention on the region-based visual feature, attention on high-level concepts, as well as attention on external knowledge.

The first four rows in Table 1 compare our context-aware visual attention model with those models without attention mechanism. The results prove our two contributions to visual attention mechanism. We use element-wise multiplication to replace addition in the SAN [51] model and get better performance, which supports our assumption that element-wise multiplication is a better multimodal fusion approach than addition in visual question answering task. The second contribution is that we incorporate contextual information from surrounding regions into target regions, which benefits the spatial inference in images. In our previous work [54], the context of the individual region is modeled by a bidirectional GRU, which achieves marginal improvement compared to the traditional visual attention method. In this work, we improve this part by using multi-directional 2D GRUs instead of the bidirectional GRUs and get 0.55% absolute promotion.

The advantage of multi-directional 2D GRUs over bidirectional GRUs lies in twofold: On one hand, multi-directional GRU can model the region's dependency in two dimensions and four directions, avoiding reordering problems from 2D spatial configuration to the 1D sequence in traditional GRUs. On the other hand, the number of regions in each dimension that 2D GRUs need to memorize is much smaller than that in 1D sequence (14 vs. 196 in our experiments), alleviating the long-term dependency problems. Besides, the 2D GRUs model is more efficient than bidirectional GRUs model in context-aware visual encoding. The 2D GRUs has two independent dimensions to control information flow, and each dimension will have two possible directions, resulting in four GRU layers in four different directions. Therefore, we will use visual feature $v_{i,j}$ as inputs four times in total. Bidirectional GRU uses each visual feature twice; however, they need model quadratic length dependency than 2D GRUs, since they flatten the 2D structured visual features into linear sequences. In our experiments, the visual feature map has 14×14 spatial arrangement, so we have $14 \times 6 \times 4$ times operations in 2D GRUs while $196 \times 4 \times 2$ times operations. So, 2D GRUs have more efficient computation than bi-GRU.

The middle three rows in Table 1 incorporate high-level concepts into the VQA framework. The first model uses the output of the final probability layer of CNN as the high-level representation of the image, while the second model tries to ground the semantic concepts with the question by sharing their word embedding parameters. We get 2.7% performance gain when we attend to concepts related to both images and questions. This demonstrates attention on high-level concepts is effective and could find more important semantic information from the image and remove noisy information irrelevant to the question.

The third block in Table 1 show the effectiveness of the knowledge discovery and knowledge-based reasoning in VQA task. "Q + Knowledge Att." model introduces attention mechanism into the knowledge-level description, which leads to 1.15% increase over "Q + Knowledge" model. Besides, "Q + Concept Att. + Knowledge Att." model gets 1.5% increase over "Q + Concept Att." model, which should give the credit to the knowledge-based reasoning module.

The last block in Table 1 joins different-level attention into one unified framework and achieve significant improvement compared with any single-level attention model. MLAN [54] is our previous work, which combined context-aware visual attention and concept attention together. When we extend this work to the knowledge domain, we get another 0.74% performance improvement. This demonstrates the attention mechanism not only on different level image representation but also on different sources are complementary and can benefit each other. Besides, we also remove the attention part from the knowledge component to demonstrate the contribution of attention mechanism in the knowledge representation.

We compare our model with the state-of-the-art methods on two large datasets. The results are shown in Table 2 on the VQA v1.0 dataset and Table 3 on the VQA v2.0 dataset, respectively. For a fair comparison, we report the results using the single model with several settings. In the VQA v1.0 dataset shown in Table 2, our proposed model achieves the best performance when compared with the state-of-the-art models including bilinear pooling-based methods [11, 24], though we use a more simple and effective framework. References [31] and [34] use two methods that exploit both visual attention and textual attention; the difference is that they perform textual attention on questions rather than high-level concepts in our model. We achieve better results than both of them, because we exploit more concepts from the image as well as knowledge rather than the question itself. In the VQA v2.0 dataset shown in Table 3, our proposed models achieve comparable model with more recent proposed structured attention [55] and bottom-up attention [2]. Reference [55] proposed a structured representation of image with a grid-structured CRF model, which is similar to our 2D GRU module for context-aware image encoding. Reference [2] proposed a more effective image feature encoding method that replaced the convolutional features with object-level

Table 2. Comparison Results on VQA v1.0 Dataset

	Approach	test-dev				test-standard			
		Yes/No	Number	Other	All	Yes/No	Number	Other	All
1	LSTM Q + I [3]	78.9	35.2	36.4	53.7	79.0	35.6	36.8	54.1
	deeper + norm [3]	80.5	36.8	43.1	57.8	80.6	36.5	43.7	58.2
	DPPnet [35]	80.7	37.2	41.7	57.2	80.3	36.9	42.2	57.4
2	SAN [51]	79.3	36.6	46.1	58.7	-	-	-	58.9
	FDA [21]	81.1	36.2	45.8	59.2	-	-	-	59.5
	DMN+[49]	80.5	36.8	48.3	60.3	-	-	-	60.4
	MCB + Att. [11]	82.2	37.7	54.8	64.2	-	-	-	-
	MCB + Att. + Glove [11]	82.5	37.6	55.6	64.7	-	-	-	-
	MCB + Att. + GloVe + VG [11]	82.3	37.2	57.4	65.4	-	-	-	-
	MLB [24]	**84.1**	38.2	54.9	64.1	-	-	-	-
	MLB + Att. [24]	83.9	37.9	56.8	65.8	-	-	-	-
3	AC [48]	79.8	36.8	43.1	57.5	79.7	36.0	43.4	57.6
	ACK [48]	81.0	38.4	45.2	59.2	81.1	37.1	45.8	59.4
	ACSK [47]	-	-	-	-	81.1	37.2	45.9	59.5
4	HieCoAtt [31]	79.7	38.7	51.7	61.8	-	-	-	62.1
	DAN [34]	83.0	39.1	53.9	64.3	82.8	39.1	54.0	64.2
5	MLAN (ResNet, train)[54]	82.9	39.2	52.8	63.7	-	-	-	-
	MLAN (ResNet, train+val +VG)[54]	81.8	**41.2**	**56.7**	65.3	81.3	**41.9**	56.5	65.2
	MSMLAN (ResNet, train)	83.6	39.9	53.6	64.4	-	-	-	-
	MSMLAN (ResNet, train+val +VG)	83.8	**41.2**	**56.7**	**66.1**	**83.7**	41.3	**56.6**	**66.2**

We divide compared approaches into five categories based on different attention mechanisms. Category 1 does not use any attention. Category 2 uses only visual attention. Category 3 extracts high-level concepts for image representation. Category 4 applies attention to both images and questions. Category 5 includes different variations of our approach.

Table 3. Results on the VQA v2.0 Dataset

	Approach	test-dev				test-standard			
		Yes/No	Number	Other	All	Yes/No	Number	Other	All
1	prior	61.0	0.3	1.1	25.7	61.2	0.4	1.2	26.0
	language_only	67.2	31.4	27.4	44.2	67.0	31.6	27.4	42.3
	deeper_lstm_Q_norm_I [3]	73.1	35.4	41.9	54.0	73.5	35.18	41.8	54.2
2	SAN [51]	81.0	44.1	52.5	63.3	-	-	-	-
	MF-SIG + VG [55]	81.3	43.0	55.6	64.7	-	-	-	-
	Bottom-up Att. [2]	81.8	44.2	56.1	65.3	82.2	43.9	56.3	65.67
3	mcb benchmark [11]	78.4	38.8	53.2	62.0	78.8	38.3	53.4	62.3
	MUTAN [4]	79.1	39.0	53.5	62.4	-	-	-	-
4	MLAN [54]	80.1	45.0	55.3	64.3	80.4	44.7	54.9	64.4
	MSMLAN	81.2	44.8	55.3	64.8	81.5	45.1	55.3	65.0
5	MSMLAN+MUTAN	82.1	45.2	56.0	65.6	-	-	-	-
	MSMLAN+Bottom-up Att.	**82.6**	**45.5**	**56.8**	**66.3**	-	-	-	-

We compare our results with state-of-the-art models. We divide compared approaches into five categories. Category 1 does not use any attention and builds baseline on this dataset. Category 2 includes three popular visual attention models. Category 3 shows two bilinear pooling-based methods. Category 4 show our proposed multi-source multi-level attention models. Category 5 combines our methods with bilinear operation and uses more powerful image features to achieve better performance.

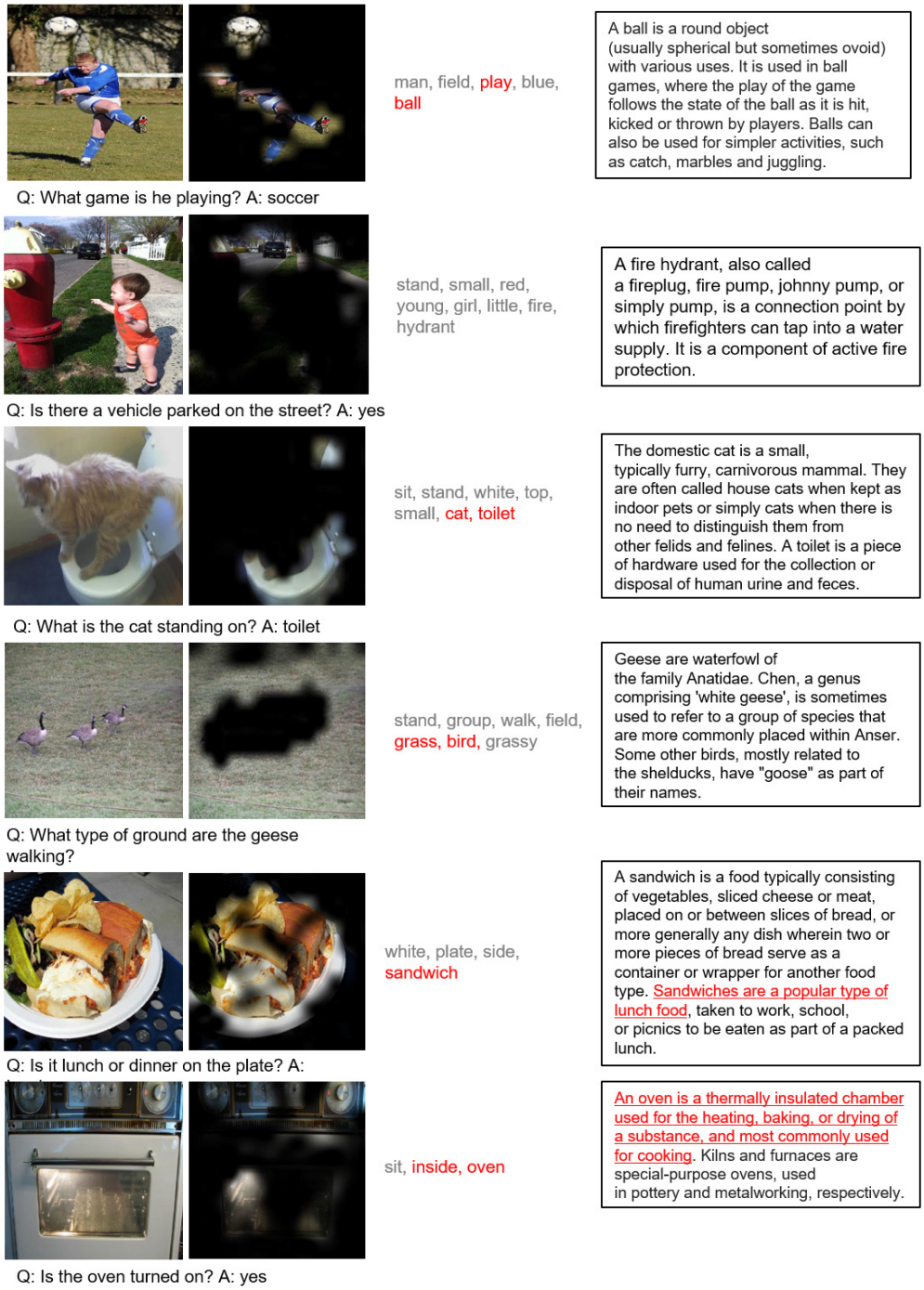

Fig. 4. Qualitative results from visual question answering with attention visualization. Image regions related to the question are marked by the gray mask, while important concepts and knowledge sentences are highlighted in red. Better view in color.

features derived from an object detection model. When we use the same image feature with them, we achieve the best performance in VQA v2.0, which demonstrates the comparibilty of our model with these state-of-the-art methods. Besides, we also show some qualitative results in Figure 4 to intutively visualize the attention in multi-source multi-level representation of images. Examples in the first two rows show that correct attended image regions lead to the true answer. The middle two rows show those cases where answers can be found directly from attended concepts. The bottom two rows show those cases where external knowledge can help reasoning.

5 CONCLUSION

We propose Multi-source Multi-level Attention Networks to incorporate visual attention, concept attention, as well as knowledge attention into an end-end framework to address automatic visual question answering. Visual attention enables region-level visual understanding queried by questions while concept attention narrows the domain gap between questions and images. The incorporation of knowledge-level attention empowers our VQA model with the capability of knowledge reasoning. Our model makes use of the complementarity of an attention mechanism on different level representations and across multiple sources. Extensive experiments on two large datasets demonstrate we not only outperform any single-level attention model but also achieve top results via a simple but effective framework. Future work includes further exploration on learning visual relationship for spatial reasoning and extracting useful knowledge from the external visual-textual dataset.

ACKNOWLEDGMENTS

Dongfei Yu was involved in this work during his internship in Microsoft Research, Beijing.

REFERENCES

[1] A. Agrawal, D. Batra, and D. Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1955–1960.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[4] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. 2017. MUTAN: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2612–2620.

[5] A. Das, H. Agrawal, et al. 2016. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* 123, 1 (2016), 32–73.

[6] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Comput. Vis. Image Understand.* 163, C (2017), 90–100.

[7] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2625–2634.

[8] H. Fang, S. Gupta, F. Landola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1473–1482.

[9] J. Fu, J. Wang, Y. Rui, X. Wang, T. Mei, and H. Lu. 2015. Image tag refinement with view-dependent concept representations. *IEEE Trans. Circ. Syst. Vid. Technol.* 25, 28 (2015), 1409–1422.

[10] J. Fu, Y. Wu, T. Mei, J. Wang, H. Lu, and Y. Rui. 2015. Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging. In *Proceedings of the IEEE International Conference on Computer Vision*. 1985–1993.

[11] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 457–468.

[12] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. 2015. Are you talking to a machine? Dataset and methods for multilingual image question answering. In *Advances in Neural Information Processing Systems*. 2296–2304.
[13] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913.
[14] A. Graves, S. Fernández, and J. Schmidhuber. 2007. Multi-dimensional recurrent neural networks. In *Proceedings of the International Conference on Artificial Neural Networks*. 549–558.
[15] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[16] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neur. Comput.* 9, 8 (1997), 1735–1780.
[17] R. Hong, Z. Hu, R. Wang, M. Wang, and D. Tao. 2016. Multi-view object retrieval via multi-scale topic models. *IEEE Trans. Image Process.* 25, 12 (2016), 5814–5827.
[18] R. Hong, L. Li, J. Cai, D. Tao, M. Wang, and Q. Tian. 2017. Coherent semantic-visual indexing for large-scale image retrieval in the cloud. *IEEE Trans. Image Process.* 26, 9 (2017), 4128–4138.
[19] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, and X. Wu. 2014. Image annotation by multiple-instance learning with discriminative feature mapping and selection. *IEEE Trans. Cybernet.* 44, 5 (2014), 669–680.
[20] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. 2018. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3588–3597.
[21] I. Ilievski, S. Yan, and J. Feng. 2016. A focused dynamic attention model for visual question answering. In *arXiv preprint arXiv:1604.01485*.
[22] A. Jabri, A. Joulin, and L. v. d. Maaten. 2016. Revisiting visual question answering baselines. In *Proceedings of the European Conference on Computer Vision*. 727–739.
[23] K. Kafle and C. Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*. 1965–1973.
[24] J. Kim, K. On, W. Lim, J. Kim, J. Ha, and B. Zhang. 2017. Hadamard product for low-rank bilinear pooling. In *Proceedings of the 5th International Conference on Learning Representations*.
[25] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*. 3294–3302.
[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[27] Q. Li, J. Fu, D. Yu, T. Mei, and J. Luo. 2018. Tell-and-answer: Towards explainable visual question answering using attributes and captions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1338–1346.
[28] L. Liang, L. Jiang, L. Gao, L. Li, and A. Hauptmann. 2018. Facal visual-text attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6135–6143.
[29] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. 2014. Microsoft COCO: Common objects in ConText. In *Proceedings of the European Conference on Computer Vision*. 740–755.
[30] Y. Liu, J. Fu, T. Mei, and C. Chen. 2017. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. 1445–1452.
[31] J. Lu, J. Yang, D. Batra, and D. Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*. 289–297.
[32] M. Malinowski, M. Rohrbach, and M. Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1–9.
[33] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). In *Proceedings of the 3rd International Conference on Learning Representations*.
[34] H. Nam, J. Ha, and J. Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 299–307.
[35] H. Noh, P. H. Seo, and B. Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 30–38.
[36] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4594–4602.
[37] Y. Pan, T. Yao, H. Li, and T. Mei. 2017. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6504–6512.
[38] M. Ren, R. Kiros, and R. S. Zemel. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*. 2953–2961.

[39] M. Rohrbach. 2017. Attributes as semantic units between natural language and visual recognition. In *Visual Attributes*. 301–330.
[40] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*. 4967–4976.
[41] K. J. Shih, S. Singh, and D. Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4613–4621.
[42] K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*.
[43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.
[44] C. Wang, H. Yang, and C. Meinel. 2018. Image captioning with deep bidirectional LSTMs and multi-task learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 40 (2018), 1–20. Issue 2s.
[45] J. Wang, J. Fu, T. Mei, and Y. Xu. 2016. Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*. 3484–3490.
[46] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 203–212.
[47] Q. Wu, C. Shen, P. Wang, A. Dick, and A. v. d. Hengel. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2017), 1367–1381.
[48] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4622–4630.
[49] C. Xiong, S. Merity, and R. Socher. 2016. Dynamic memory networks for visual and texual question answering. In *Proceedings of the International Conference on Machine Learning*. 2397–2406.
[50] H. Xu and K. Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Proceedings of the European Conference on Computer Vision*. 451–466.
[51] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 21–29.
[52] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*. 4894–4902.
[53] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4651–4659.
[54] D. Yu, J. Fu, T. Mei, and Y. Rui. 2017. Multi-level attention network for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4709–4717.
[55] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma. 2017. Structured attentions for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 1291–1300.

Received June 2018; revised February 2019; accepted February 2019