Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering

Huijuan Xu and Kate Saenko(✉)

Computer Science, Boston University, Boston, USA
{hxu,saenko}@bu.edu

Abstract. We address the problem of Visual Question Answering (VQA), which requires joint image and language understanding to answer a question about a given photograph. Recent approaches have applied deep image captioning methods based on convolutional-recurrent networks to this problem, but have failed to model spatial inference. To remedy this, we propose a model we call the Spatial Memory Network and apply it to the VQA task. Memory networks are recurrent neural networks with an explicit attention mechanism that selects certain parts of the information stored in memory. Our Spatial Memory Network stores neuron activations from different spatial regions of the image in its memory, and uses attention to choose regions relevant for computing the answer. We propose a novel question-guided spatial attention architecture that looks for regions relevant to either individual words or the entire question, repeating the process over multiple recurrent steps, or "hops". To better understand the inference process learned by the network, we design synthetic questions that specifically require spatial inference and visualize the network's attention. We evaluate our model on two available visual question answering datasets and obtain improved results.

Keywords: Visual question answering · Spatial attention · Memory network · Deep learning

1 Introduction

Visual Question Answering (VQA) is an emerging interdisciplinary research problem at the intersection of computer vision, natural language processing and artificial intelligence. It has many real-life applications, such as automatic querying of surveillance video [1] or assisting the visually impaired [2]. Compared to the recently popular image captioning task [3–6], VQA requires a deeper understanding of the image, but is considerably easier to evaluate. It also puts more focus on artificial intelligence, namely the inference process needed to produce the answer to the visual question.

In one of the early works [8], VQA is seen as a Turing test proxy. The authors propose an approach based on handcrafted features, combining a semantic parse

B. Leibe et al. (Eds.): ECCV 2016, Part VII, LNCS 9911, pp. 451–466, 2016.
DOI: 10.1007/978-3-319-46478-7_28

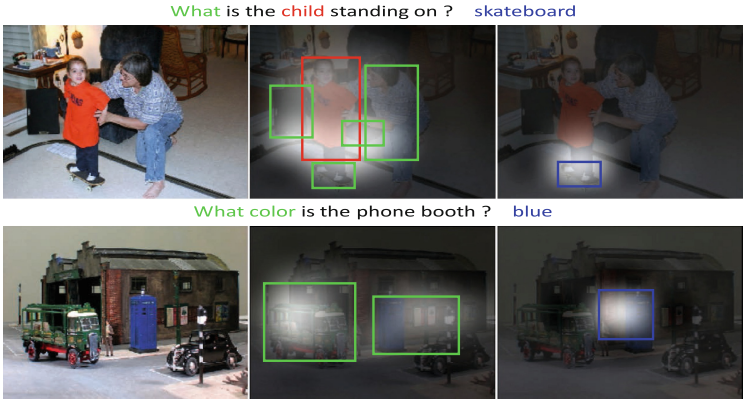

Fig. 1. We propose a Spatial Memory Network for VQA (SMem-VQA) that answers questions about images using spatial inference. The figure shows the inference process of our two-hop model on examples from the VQA dataset [7]. In the first hop (middle), the attention process captures the correspondence between individual words in the question and image regions. High attention regions (bright areas) are marked with bounding boxes and the corresponding words are highlighted using the same color. In the second hop (right), the fine-grained evidence gathered in the first hop, as well as an embedding of the entire question, are used to collect more exact evidence to predict the answer. (Best viewed in color.) (Color figure online)

of the question with visual scene analysis in a latent-world Bayesian framework. More recently, several end-to-end deep neural networks that learn features directly from data have been applied to this problem [9,10], featuring networks adapted directly from captioning models [3–5]. These methods utilize a recurrent LSTM network to encode the question words and Convolutional Neural Net (CNN) image features into a hidden state, then predict the answer. Despite a great improvement compared to the handcrafted feature method [8], the LSTM-based methods have their own drawbacks. First, conditioning on both the image and question encodings does not provide a clear improvement over conditioning just on the question encoding alone [9,10]. Second, the rather complicated LSTM models obtain similar or worse accuracy compared to a baseline model which concatenates CNN features and a bag-of-words question embedding[1] to predict the answer, such as the IMG+BOW model in [10] and the iBOWIMG model in [11].

A major limitation of the existing models is that they rely on whole-image features with no explicit notion of object position, and do not support the computation of intermediate results based on spatial attention. Our intuition is that answering visual questions often involves paying attention to individual spatial regions and comparing their contents and/or locations. For example, to answer the questions in Fig. 1, we must first find the regions corresponding to certain

[1] Weighted average of the word vectors.

words in the question ("child", "phone booth"), and then analyse them or their nearby regions.

Inspired by this intuition, we propose a new deep learning approach to VQA that incorporates explicit spatial attention, which we call the Spatial Memory Network VQA (SMem-VQA). Our approach is based on memory networks, which have recently been proposed for text Question Answering (QA) [12,13]. Memory networks combine learned text embeddings with an attention mechanism and multi-step inference. The text QA memory network stores textual knowledge in its "memory" in the form of sentences, and selects relevant sentences to infer the answer. However, in VQA, the knowledge is in the form of an image, thus the memory and the question come from different modalities. We adapt the end-to-end memory network [13] to solve visual question answering by storing the convolutional network outputs obtained from different receptive fields into the memory, which explicitly allows spatial attention over the image. We also propose to repeat the process of gathering evidence from attended regions, enabling the model to update the answer based on several attention steps, or "hops". The entire model is trained end-to-end and the evidence for the computed answer can be visualized using the attention weights.

To summarize our contributions, in this paper we:

- propose a novel multi-hop memory network with spatial attention for the VQA task which allows one to visualize the spatial inference process used by the deep network (a CAFFE [14] implementation is available at https://github.com/VisionLearningGroup/Ask_Attend_and_Answer);
- design a word-guided attention architecture which captures fine-grained alignment between the words and regions in the first hop;
- create a series of synthetic questions that explicitly require spatial inference to analyze the working principles of the network, and demonstrate that it is able to learn logical inference rules through visualizations; and
- provide an extensive evaluation and comparison with several existing models on the same publicly available datasets.

Section 2 reviews relevant work on memory networks and attention models. Section 3 describes our design of the multi-hop memory network architecture for visual question answering (SMem-VQA). Section 4 visualizes the inference rules learned by the network for synthetic spatial questions and shows the experimental results on DAQUAR [8] and VQA [7] datasets. Section 5 concludes the paper.

2 Related Work

Before visual question answering (VQA) became popular, text question answering (QA) had already been established as a mature research problem in the area of natural language processing. Previous QA methods include: searching for the key words of the question using a search engine [15]; parsing the question as a

knowledge base (KB) query [16]; or embedding the question and using a similarity measurement to find evidence for the answer [17]. Recently, memory networks were proposed for solving the QA problem. [12] first introduces the memory network as a general model that consists of a memory and four components: input feature map, generalization, output feature map and response. The model is investigated in the context of text QA, where the long-term memory acts as a dynamic knowledge base and the output is a textual response. [13] proposes the "end-to-end" memory network which uses less supervision and implements a recurrent attention model over a large external memory. The related Neural Turing Machine (NTM) [18] couples a neural network to external memory and interacts with it by attentional processes to infer simple algorithms such as copying, sorting, and associative recall from input and output examples. In this paper, we propose a multimodal memory network architecture based on [13] that is the first to address visual question answering (a related model was recently independently proposed in [19]).

The neural attention mechanism has been widely used in different areas, for example, in image captioning [20], video description generation [21], machine translation [22,23] and machine reading systems [24]. Most methods use the soft attention mechanism [22], which adds a layer to the network that predicts soft weights and uses them to compute a weighted combination of the items in memory. The two main types of soft attention mechanisms differ in the function that combines the input feature vector and the candidate feature vectors in order to compute the soft attention weights. The first type uses an alignment function based on "concatenation" of the input and each candidate (we use the term "concatenation" as described in [23]). This function adds an input vector (e.g. hidden state vector of the LSTM) to each candidate feature vector, embeds the resulting vectors into scalar values, and then applies the softmax function to generate the attention weight for each candidate. [20–22,24] use the "concatenation" alignment function in their soft attention models and [25] gives a literature review of such models applied to different tasks. The second type uses an alignment function based on the dot product of the input and each candidate. It first projects both inputs to a common vector embedding space, then takes the dot product of the two input vectors, and applies a softmax function to produce the attention weight for each candidate. Motivated by the use of dot product alignment in the end-to-end memory network [13] and a study that found it superior to concatenation alignment [23], we also use this form of alignment in our Spatial Memory Network.

Several early VQA papers directly adapted image captioning models to solve the problem [9,10] by generating the answer using a recurrent LSTM network conditioned on the CNN output, but lacked spatial attention. [26] uses a spatial attention model similar to that in image captioning [20], but does not provide results on the more common VQA benchmark [7], and our own implementation of this model is less accurate on [7] than other baseline models. [11] summarizes several recent results on the VQA dataset [7] on arxiv.org and proposes a simple but strong baseline model (iBOWIMG). This baseline concatenates the image

features with the bag-of-words question representation and feeds them into a softmax classifier to predict the answer. The iBOWIMG model beats most VQA models considered in the paper. Here, we compare our proposed model to the VQA models (namely, the ACK model [27] and the DPPnet model [28]) which have comparable or better results than the iBOWIMG model. The ACK model in [27] is essentially the same as the LSTM model in [10], except that it uses image attribute features, the generated image caption and relevant external knowledge from a knowledge base as the input to the LSTM's first time step. The DPPnet model in [28] tackles VQA by learning a dynamic parameter prediction network that uses a Gate Recurrent Unit (GRU) to generate a question representation, and, based on this, predicts the CNN weights via hashing. Neither of these models [27,28] contain a spatial attention mechanism, and they both use external data in addition to the VQA dataset [7], e.g. the knowledge base in [27] and the large-scale text corpus used to pre-train the GRU question representation [28]. In this paper, we explore a complementary approach of spatial attention to both improve performance and visualize the network's inference process, and obtain improved results without using external data compared to the iBOWIMG model [11] as well as the ACK model [27] and the DPPnet model [28] which use external data.

3 Spatial Memory Network for VQA

We start with an overview of the first time step (hop) of our proposed SMem-VQA network, illustrated in Fig. 2(a). The input is a question comprised of a variable-length sequence of words and an image of fixed size. Each word is first represented as a one-hot vector in the size of the vocabulary, and then embedded into a real-valued word vector, $V = \{v_j \mid v_j \in \mathbb{R}^N; j = 1, \cdots, T\}$, where T is the maximum number of words and N is the dimensionality of the embedding space. Sentences with length less than T are padded with all-zero word vectors.

The question words are used to compute attention over the visual memory, which contains extracted image features. We use $S = \{s_i \mid s_i \in \mathbb{R}^M; i = 1, \cdots, L\}$ to represent spatial CNN features at each of the L grid locations (in this work, the last convolutional layer of GoogLeNet (*inception_5b/output*) [29].) The image features are embedded into the same number of dimensions as the word vectors using two different embeddings: the "attention" embedding W_A and the "evidence" embedding W_E. The attention embedding generates the attention weights, while the evidence embedding maps the features to semantic concepts such as objects. The embedded features are multiplied with the attention weights and summed over all locations to generate a visual evidence vector S_{att}. Finally, S_{att} is combined with a representation of the question to predict the answer. We describe this one-hop model and its attention mechanism in more detail in the next section, then discuss adding more hops in Sect. 3.2.

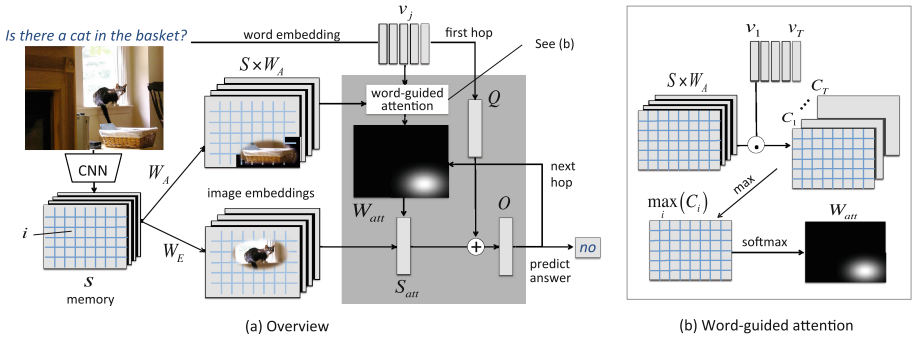

Fig. 2. (a) Overview of our proposed Spatial Memory Network for Visual Question Answering (SMem-VQA). Unlike previous models that disregard object location, ours uses a spatial attention mechanism to attend to relevant regions and gather visual evidence for predicting the answer (see Sect. 3 for details). (b) The word-guided spatial attention model used in the first hop of the network (see Sect. 3.1 for details.) (Color figure online)

3.1 Word Guided Spatial Attention in the First Hop

Rather than using the entire question representation, such as a bag-of-words, to guide attention, the architecture in the first hop (Fig. 2(b)) uses each word vector separately to extract correlated visual features in memory. The intuition is that the BOW representation may be too coarse, and letting each word select a region may provide more fine-grained attention. The correlation matrix $C \in \mathbb{R}^{T \times L}$ between word vectors V and visual features S is computed as

$$C = V \cdot (S \cdot W_A + b_A)^T \tag{1}$$

where $W_A \in \mathbb{R}^{M \times N}$ contains the attention embedding weights of visual features S, and $b_A \in \mathbb{R}^{L \times N}$ is the bias term. This correlation matrix is the result of the dot product of each word embedding and each spatial location's embedding, thus each value in C measures the similarity between a word and a region.

The spatial attention weights W_{att} are calculated by taking the maximum of C over the word dimension T, thus selecting the highest correlation value for each spatial location, and then applying the softmax function

$$W_{att} = \text{softmax}(\max_{i=1,\cdots,T}(C_i)),\ C_i \in \mathbb{R}^L \tag{2}$$

The resulting attention weights $W_{att} \in \mathbb{R}^L$ are high for selected locations and low for other locations, with the sum of weights equal to 1. For instance, the example question "Is there a cat in the basket?" in Fig. 2 might produce high attention weights for the location of the basket because of high correlation of the word vector for *basket* with the embedded features at that location. Note that W_A controls which image features have high correlation with which words.

The evidence embedding W_E projects visual features S to produce high activations for certain semantic concepts. E.g., in Fig. 2, it may have high activations in the regions containing objects such as *cat*. The results of this evidence embedding are then multiplied by the generated attention weights W_{att}, and summed to produce the selected visual "evidence" vector $S_{att} \in \mathbb{R}^N$,

$$S_{att} = W_{att} \cdot (S \cdot W_E + b_E) \tag{3}$$

where $W_E \in \mathbb{R}^{M \times N}$ are the evidence embedding weights of the visual features S, and $b_E \in \mathbb{R}^{L \times N}$ is the bias term. In our running example, this step would accumulate evidence of objects such as *cat* at the *basket* location.

Finally, the sum of this evidence vector S_{att} and an embedding of the question Q is used to predict the answer for the given image and question. While many question representations, such as an LSTM, can be used for Q, we use the BOW as it has fewer parameters yet has shown good performance compared to LSTM [30]. Specifically, we compute

$$Q = W_Q \cdot V + b_Q \tag{4}$$

where $W_Q \in \mathbb{R}^T$ represents the BOW weights for word vectors V, and $b_Q \in \mathbb{R}^N$ is the bias term. The final prediction P is computed as

$$P = \text{softmax}(W_P \cdot f(S_{att} + Q) + b_P) \tag{5}$$

where $W_P \in \mathbb{R}^{K \times N}$, bias term $b_P \in \mathbb{R}^K$, and K is the number of possible answers. f is the activation function, and we use ReLU here. In our running example, this step would add the evidence gathered for objects near the basket location to the question, and, since *cat* was not detected there, predict the answer "no". The attention and evidence computation steps can be optionally repeated in another hop before predicting the final answer, as detailed in the next section.

3.2 Spatial Attention in the Second Hop

We can add hops to promote deeper inference, gathering additional evidence at each hop. Recall that the visual evidence vector S_{att} is added to the question representation Q in the first hop to produce an updated question vector,

$$O_{hop1} = S_{att} + Q \tag{6}$$

On the next hop, this vector $O_{hop1} \in \mathbb{R}^N$ is used in place of the individual word vectors V to extract additional visual evidence from spatial memory based on the updated question.

While the correlation matrix C in the first hop provides fine-grained local evidence from each word vectors V in the question, the correlation vector C_{hop2} in the next hop considers the global evidence from the updated question O_{hop1}. The correlation vector $C_{hop2} \in \mathbb{R}^L$ is calculated by

$$C_{hop2} = (S \cdot W_{A_2} + b_{A_2}) \cdot O_{hop1} \tag{7}$$

where $W_{A_2} \in \mathbb{R}^{M \times N}$ is the attention embedding of visual features S in the second hop and $b_{A_2} \in \mathbb{R}^{L \times N}$ is the bias term. Based on experimental results, we share the attention embedding in the second hop and the evidence embedding in the first hop, such that $W_{A_2} = W_E$ and $b_{A_2} = b_E$.

The attention weights in the second hop W_{att2} are obtained by applying the softmax function to the correlation vector C_{hop2},

$$W_{att2} = \text{softmax}(C_{hop2}) \tag{8}$$

Then, the attended visual information in the second hop $S_{att2} \in \mathbb{R}^N$ is extracted using attention weights W_{att2}.

$$S_{att2} = W_{att2} \cdot (S \cdot W_{E_2} + b_{E_2}) \tag{9}$$

where $W_{E_2} \in \mathbb{R}^{M \times N}$ is the evidence embedding of visual features S in the second hop, and $b_{E_2} \in \mathbb{R}^{L \times N}$ is the bias term.

The final answer P is predicted by combining the whole question representation Q, the local visual evidence S_{att} from each word vector in the first hop and the global visual evidence S_{att2} from the whole question in the second hop,

$$P = \text{softmax}(W_P \cdot f(O_{hop1} + S_{att2}) + b_P) \tag{10}$$

where $W_P \in \mathbb{R}^{K \times N}$, bias term $b_P \in \mathbb{R}^K$, and K is the number of possible answers. More hops can be added in this manner.

The entire network is differentiable and is trained using stochastic gradient descent via standard backpropagation, allowing image feature extraction, image embedding, word embedding and answer prediction to be jointly optimized on the training image/question/answer triples.

4 Experiments

In this section, we conduct a series of experiments to evaluate our model. To explore whether the model learns to perform the spatial inference necessary for answering visual questions that explicitly require spatial reasoning, we design a set of experiments using synthetic visual question/answer data in Sect. 4.1. The experimental results of our model in standard datasets (DAQUAR [8] and VQA [7] datasets) are reported in Sect. 4.2.

4.1 Exploring Attention on Synthetic Data

The questions in the public VQA datasets are quite varied and difficult and often require common sense knowledge to answer (e.g., "Does this man have 20/20 vision?" about a person wearing glasses). Furthermore, past work [9,10] showed that the question text alone (no image) is a very strong predictor of the answer. Therefore, before evaluating on standard datasets, we would first like to understand how the proposed model uses spatial attention to answer simple visual questions where the answer cannot be predicted from question alone. Our visualization demonstrates that the attention mechanism does learn to attend to objects and gather evidence via certain inference rules.

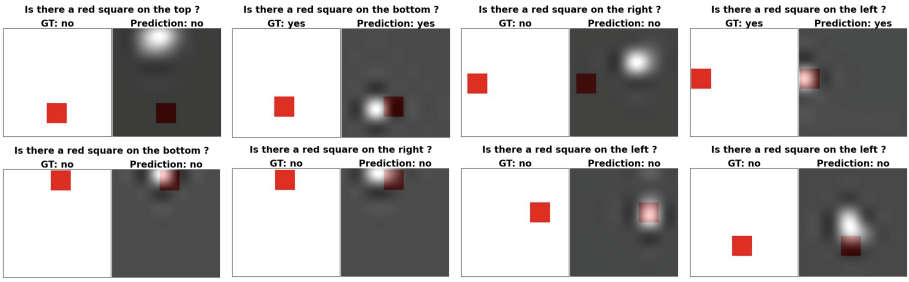

Fig. 3. Absolute position experiment: for each image and question pair, we show the original image (left) and the attention weights W_{att} (right). The attention follows one of two learned rules. The first rule (top row) looks at the position specified in the question (top|bottom|right|left), and answers "yes" if it contains a square and "no" otherwise. The second rule (bottom row) looks at the region containing the square, and answers "yes" if the question refers to that position and "no" otherwise. (Color figure online)

Absolute Position Recognition. We investigate whether the model has the ability to recognize the rough absolute location of the object in the image. We design a simple task where an object (a red square) appears in some region of a white-background image, and the question is "Is there a red square on the [top|bottom|left|right]?" For each image, the square is randomly placed in one of the four regions, and the four questions are generated together with three "no" and one "yes" answer. The generated data is split into training and testing sets.

Due to the simplicity of this synthetic dataset, the SMem-VQA one-hop model achieves 100 % test accuracy. However, the baseline model (iBOWIMG) [11] cannot infer the answer and only obtains accuracy of around 75 %, which is the prior probability of the answer "no" in the training set. The SMem-VQA one-hop model is equivalent to the iBOWIMG model if the attention weights in our one-hop model are set equally for each location, since the iBOWIMG model uses mean pooling of the same convolutional features (*inception_5b/output* in GoogLeNet). We visualize the attention weights (Fig. 3) and find that the relationship between the high-attention regions and the answer can be expressed by one of two logical expressions: (1) Look at the position specified in the question (top|bottom|right|left), if it contains a square, then answer "yes", otherwise, answer "no"; (2) Look at the region containing the square, then answer "yes" if the question is about that position and "no" otherwise.

In the iBOWIMG model, the mean-pooled GoogLeNet visual features lose spatial information and thus cannot distinguish images with a square in different positions. On the contrary, our SMem-VQA model can select different regions according to the question, and generate an answer based on the selected region, using some learned inference rules. This experiment demonstrates that the attention mechanism in our model is able to make absolute spatial location inference based on the spatial attention.

Fig. 4. Relative position experiment: for each image and question pair, we show the original image (left), the evidence embedding W_E of the convolutional layer (middle) and the attention weights W_{att} (right). The evidence embedding W_E has high activations on both cat and red square. The attention weights follow similar inference rules as in Fig. 3, with the difference that the attention position is relative to the cat. (Color figure online)

Relative Position Recognition. To check whether the model has the ability to infer the position of one object *relative* to another object, we collect all the cat images from the MS COCO Detection dataset [31], and add a red square on the [top|bottom|left|right] of the bounding box containing the cat. For each generated image, we create four questions, "Is there a red square on the [top|bottom|left|right] of the cat?" together with three "no" answers and one "yes" answer. We select 2639 training cat images and 1395 testing cat images from MS COCO Detection dataset.

Our SMem-VQA one-hop model achieves 96 % test accuracy on this synthetic task, while the baseline model (iBOWIMG) accuracy is around 75 %. We also check that another simple baseline that predicts the answer based on the absolute position of the square in the image gets around 70 % accuracy. We visualize the image features after the evidence embedding W_E (max pooled over channel dimension) and the attention weights W_{att} of several typical examples in Fig. 4. The evidence embedding W_E has high activations on the cat and the red square, while the attention weights are high at certain locations relative to the cat. We can analyze the attention in the correctly predicted examples using the same rules as in the absolute position recognition experiment. These rules still work, but the position is now relative to the cat object: (1) Check the specified position relative to the cat, if it has the square, then answer "yes", otherwise "no"; (2) Find the square, then answer "yes" if it is in the specified relative position, and "no" otherwise. We also check the images where our model makes mistakes, and find that they mainly occur in images with more than one cat. The red square appears near only one of the cats, but our model might focus on the other cats. We conclude that our SMem-VQA model can infer the relative spatial position based on the spatial attention around the specified object, which can also be represented by logical inference rules.

Table 1. Accuracy results on the DAQUAR dataset (in percentage).

	DAQUAR
Multi-World [8]	12.73
Neural-Image-QA [9]	29.27
Question LSTM [9]	32.32
VIS+LSTM [10]	34.41
Question BOW [10]	32.67
IMG+BOW [10]	34.17
SMem-VQA One-Hop	36.03
SMem-VQA Two-Hop	**40.07**

4.2 Experiments on Standard Datasets

Results on DAQUAR. The DAQUAR dataset [8] is a relatively small dataset which builds on the NYU Depth Dataset V2 [32]. We use the reduced DAQUAR dataset. The evaluation metric for this dataset is 0-1 accuracy. The embedding dimension is 512 for our models running on the DAQUAR dataset. We use several reported models on DAQUAR as baselines, which are listed below:

- **Multi-World** [8]**:** an approach based on handcrafted features using a semantic parse of the question and scene analysis of the image combined in a latent-world Bayesian framework.
- **Neural-Image-QA** [9]**:** uses an LSTM to encode the question and then decode the hidden information into the answer. The image CNN feature vector is shown at each time step of the encoding phase.
- **Question LSTM** [9]**:** only shows the question to the LSTM to predict the answer without any image information.
- **VIS+LSTM** [10]**:** similar to Neural-Image-QA, but only shows the image features to the LSTM at the first time step, and the question in the remaining time steps to predict the answer.
- **Question BOW** [10]**:** only uses the BOW question representation and a single hidden layer neural network to predict the answer, without any image features.
- **IMG+BOW** [10]**:** concatenates the BOW question representation with image features, and then uses a single hidden layer neural network to predict the answer. This model is similar to the iBOWIMG baseline model in [11].

Results of our SMem-VQA model on the DAQUAR dataset and the baseline model results reported in previous work are shown in Table 1. We see that models based on deep features significantly outperform the Multi-World approach based on hand-crafted features. Modeling the question only with either the LSTM model or Question BOW model does equally well in comparison, indicating the question text contains important prior information for predicting the answer. Also, on this dataset, the VIS+LSTM model achieves better accuracy than Neural-Image-QA model; the former shows the image only at the first

Fig. 5. Visualization of the spatial attention weights in the SMem-VQA One-Hop and Two-Hop models on VQA (top row) and DAQUAR (bottom row) datasets. For each image and question pair, we show the original image, the attention weights W_{att} of the One-Hop model, and the two attention weights W_{att} and W_{att2} of the Two-Hop model in order. (Color figure online)

timestep of the LSTM, while the latter does so at each timestep. In comparison, both our One-Hop model and Two-Hop spatial attention models outperform the IMG+BOW, as well as the other baseline models. A major advantage of our model is the ability to visualize the inference process in the deep network. To illustrate this, two attention weights visualization examples in SMem-VQA One-Hop and Two-Hop models on DAQUAR dataset are shown in Fig. 5 (bottom row).

Results on VQA. The VQA dataset [7] is a recent large dataset based on MS COCO [31]. We use the full release (V1.0) open-ended dataset, which contains a train set and a val set. Following standard practice, we choose the top 1000 answers in train and val sets as possible prediction answers, and only keep the examples whose answers belong to these 1000 answers as training data. The question vocabulary size is 7477 with the word frequency of at least three. Because of the larger training size, the embedding dimension is 1000 on the VQA dataset. We report the test-dev and test-standard results from the VQA evaluation server. The server evaluation uses the evaluation metric introduced by [7], which gives partial credit to certain synonym answers: $Acc(ans) = \min\{(\#\text{ humans that said}ans)/3, 1\}$.

For the attention models, we do not mirror the input image when using the CNN to extract convolutional features, since this might cause confusion about the spatial locations of objects in the input image. The optimization algorithm used is stochastic gradient descent (SGD) with a minibatch of size 50 and momentum of 0.9.

For the VQA dataset, we use the simple iBOWIMG model in [11] as one baseline model, which beats most existing VQA models currently on arxiv.org. We also compare to two models in [27,28] which have comparable or better results to the iBOWIMG model. These three baseline models as well the best model in the VQA dataset paper [7] are listed in the following:

Table 2. Test-dev and test-standard results on the Open-Ended VQA dataset (in percentage). Models with * use external training data in addition to the VQA dataset.

	Test-dev				Test-standard			
	Overall	Yes/No	Number	Others	**Overall**	Yes/No	Number	Others
LSTM Q+I [7]	53.74	78.94	35.24	36.42	54.06	-	-	-
ACK* [27]	55.72	79.23	36.13	40.08	55.98	79.05	36.10	40.61
DPPnet* [28]	57.22	80.71	37.24	41.69	57.36	80.28	36.92	42.24
iBOWIMG [11]	55.72	76.55	35.03	42.62	55.89	76.76	34.98	42.62
SMem-VQA One-Hop	56.56	78.98	35.93	42.09	-	-	-	-
SMem-VQA Two-Hop	**57.99**	**80.87**	**37.32**	**43.12**	**58.24**	**80.8**	**37.53**	**43.48**

- **LSTM Q+I** [7]**:** uses the element-wise multiplication of the LSTM encoding of the question and the image feature vector to predict the answer. This is the best model in the VQA dataset paper.
- **ACK** [27]**:** shows the image attribute features, the generated image caption and relevant external knowledge from knowledge base to the LSTM at the first time step, and the question in the remaining time steps to predict the answer.
- **DPPnet** [28]**:** uses the Gated Recurrent Unit (GRU) representation of question to predict certain parameters for a CNN classification network. They pre-train the GRU for question representation on a large-scale text corpus to improve the GRU generalization performance.
- **iBOWIMG** [11]**:** concatenates the BOW question representation with image features (GoogLeNet), and uses softmax classification to predict the answer.

The overall accuracy and per-answer category accuracy for our SMem-VQA models and the baseline models on VQA dataset are shown in Table 2. From the table, we can see that the SMem-VQA One-Hop model obtains slightly better results compared to the iBOWIMG model. However, our SMem-VQA Two-Hop model achieves an improvement of 2.27 % on test-dev and 2.35 % on test-standard compared to the iBOWIMG model, demonstrating the value of spatial attention. If we set uniform attention weights over the image regions, we get test-dev accuracy 55.97 % for our One-Hop model and 55.83 % for our Two-Hop model. Our One-Hop model with uniform attention weights is equivalent to the baseline model iBOWIMG, except that the question and image features are summed rather than concatenated. The SMem-VQA Two-Hop model also shows best performance in the per-answer category accuracy.

The DPPnet model uses a large-scale text corpus to pre-train the Gated Recurrent Unit (GRU) network for question representation. DPPnet without pre-training (RAND-GRU) gets the test-dev result 55.46 % compared to 57.99 % of our Two-Hop model. Similar pre-training work on extra data to improve model accuracy has been done in [33]. Considering the fact that our model does not use extra data to pre-train the word embeddings, its results are very competitive. We tried the layer-wise weight sharing strategy in [13] and got the overall test-dev result of 55.76 % for the Two-Hop model, which is lower than the adjacent

Fig. 6. Visualization of the original image (left), the spatial attention weights W_{att} in the first hop (middle) and one correlation vector from the correlation matrix C for the location with highest attention weight in the SMem-VQA Two-Hop model on the VQA dataset. Higher values in the correlation vector indicate stronger correlation of that word with the chosen location's image features. (Color figure online)

weight sharing strategy that we take. We also experimented with adding a third hop into our model on the VQA dataset, but the result did not improve further.

The attention weights visualization examples for the SMem-VQA One-Hop and Two-Hop models on the VQA dataset are shown in Fig. 5 (top row). From the visualization, we can see that the two-hop model collects supplementary evidence for inferring the answer, which may be necessary to achieve an improvement on these complicated real-world datasets. We also visualize the fine-grained alignment in the first hop of our SMem-VQA Two-Hop model in Fig. 6. The correlation vector values (blue bars) measure the correlation between image regions and each word vector in the question. Higher values indicate stronger correlation of that particular word with the specific location's image features. We observe that the fine-grained visual evidence collected using each local word vector, together with the global visual evidence from the whole question, complement each other to infer the correct answer for the given image and question, as shown in Fig. 1.

5 Conclusion

In this paper, we proposed a memory network architecture with a spatial attention mechanism adapted to the visual question answering task. We designed a set of synthetic spatial questions and demonstrated that our model learns inference rules based on spatial attention through attention weight visualization. Evaluation on the challenging DAQUAR and VQA datasets showed improved results over previously published models and no-attention baselines. Our model can be used to visualize the inference steps learned by the deep network, giving some insight into its processing.

Acknowledgments. This work was supported by NSF Award IIS-1212928 and a Google Faculty Research Award. The authors would like to thank Trevor Darrell, Raymond Mooney, Marcus Rohrbach and Subhashini Venugopalan for valuable discussions.

References

1. Tu, K., Meng, M., Lee, M.W., Choe, T.E., Zhu, S.C.: Joint video and text parsing for understanding events and answering queries. IEEE MultiMedia **21**(2), 42–70 (2014)
2. Lasecki, W.S., Zhong, Y., Bigham, J.P.: Increasing the bandwidth of crowdsourced visual question answering to better support blind users. In: Proceedings of the 16th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 263–264. ACM (2014)
3. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. arXiv preprint arXiv:1411.4389 (2014)
4. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. arXiv preprint arXiv:1411.4555 (2014)
5. Karpathy, A., Joulin, A., Li, F.F.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in Neural Information Processing Systems, pp. 1889–1897 (2014)
6. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., et al.: From captions to visual concepts and back. arXiv preprint arXiv:1411.4952 (2014)
7. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. CoRR abs/1505.00468 (2015)
8. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. CoRR abs/1410.0210 (2014)
9. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: a neural-based approach to answering questions about images. arXiv preprint arXiv:1505.01121 (2015)
10. Ren, M., Kiros, R., Zemel, R.S.: Exploring models and data for image question answering. CoRR abs/1505.02074 (2015)
11. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)
12. Weston, J., Chopra, S., Bordes, A.: Memory networks. CoRR abs/1410.3916 (2014)
13. Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: End-to-end memory networks. arXiv preprint arXiv:1503.08895 (2015)
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint arXiv:1408.5093 (2014)
15. Yahya, M., Berberich, K., Elbassuoni, S., Ramanath, M., Tresp, V., Weikum, G.: Natural language questions for the web of data. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, pp. 379–390 (2012)
16. Berant, J., Liang, P.: Semantic parsing via paraphrasing. In: Proceedings of ACL, vol. 7, p. 92 (2014)
17. Bordes, A., Chopra, S., Weston, J.: Question answering with subgraph embeddings. arXiv preprint arXiv:1406.3676 (2014)

18. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401 (2014)
19. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. CoRR abs/1603.01417 (2016)
20. Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044 (2015)
21. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4507–4515 (2015)
22. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
23. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
24. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems, pp. 1684–1692 (2015)
25. Cho, K., Courville, A., Bengio, Y.: Describing multimedia content using attention-based encoder-decoder networks (2015)
26. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7W: grounded question answering in images. arXiv preprint arXiv:1511.03416 (2015)
27. Wu, Q., Wang, P., Shen, C., van den Hengel, A., Dick, A.: Ask me anything: free-form visual question answering based on knowledge from external sources. arXiv preprint arXiv:1511.06973 (2015)
28. Noh, H., Seo, P.H., Han, B.: Image question answering using convolutional neural network with dynamic parameter prediction. arXiv preprint arXiv:1511.05756 (2015)
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR 2015 (2015)
30. Shih, K.J., Singh, S., Hoiem, D.: Where to look: focus regions for visual question answering. arXiv preprint arXiv:1511.07394 (2015)
31. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_48
32. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33715-4_54
33. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014)