Multi-level Attention Networks for Visual Question Answering*

Dongfei Yu, Jianlong Fu, Tao Mei, Yong Rui
University of Science and Technology of China, Microsoft Research, Beijing, China
ydf2010@mail.ustc.edu.cn, {jianf, tmei}@microsoft.com, yongrui@outlook.com

Abstract

Inspired by the recent success of text-based question answering, visual question answering (VQA) is proposed to automatically answer natural language questions with the reference to a given image. Compared with text-based QA, VQA is more challenging because the reasoning process on visual domain needs both effective semantic embedding and fine-grained visual understanding. Existing approaches predominantly infer answers from the abstract low-level visual features, while neglecting the modeling of high-level image semantics and the rich spatial context of regions. To solve the challenges, we propose a multi-level attention network for visual question answering that can simultaneously reduce the semantic gap by semantic attention and benefit fine-grained spatial inference by visual attention. First, we generate semantic concepts from high-level semantics in convolutional neural networks (CNN) and select those question-related concepts as semantic attention. Second, we encode region-based middle-level outputs from CNN into spatially-embedded representation by a bidirectional recurrent neural network, and further pinpoint the answer-related regions by multiple layer perceptron as visual attention. Third, we jointly optimize semantic attention, visual attention and question embedding by a softmax classifier to infer the final answer. Extensive experiments show the proposed approach outperforms the-state-of-arts on two challenging VQA datasets.

1. Introduction

Visual question answering (VQA) has attracted extensive attention recently, since VQA is considered approaching towards the milestone of "AI-complete" that enables a machine to reason across language and vision as humans [38]. Compared with text-based QA system in natural language processing (NLP), VQA takes one step further, which is able to answer a natural language question by considering the correspondence between a question and a reference image. The capability of automatic VQA can significantly promote the mutual understanding between language and vision, and further benefit a variety of applications, such as visually-impaired assistant devices, early education, service robots, and so on.

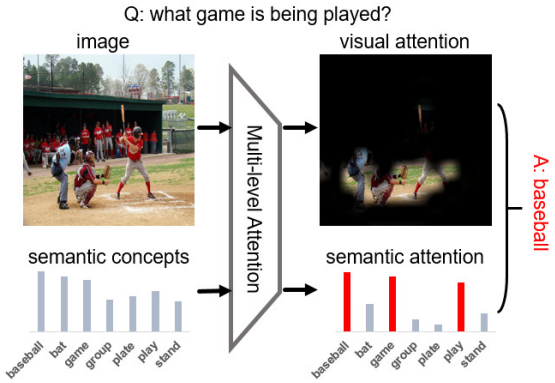

Figure 1. Overview of the multi-level attention network (MLAN). The proposed attention model highlights both question-related semantic concepts (*i.e.*, "baseball," "game," "play") and image regions (*i.e.*, those regions of "bat" and "glove").

The challenges of visual question answering are twofold: effective semantic embedding and fine-grained visual understanding. Most early works transfer image captioning framework [5, 19, 28, 35] to VQA tasks [10, 18, 24] by the combination of convolutional neural networks (CNN) [14] and recurrent neural network (RNN) [12]. Specifically, these works extract global image representation from a pre-trained CNN model and extract question representation from a RNN model. They further feed the joint embedding features across language and vision 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 dataset [2]). Although promising results have been reported, further improvement suffers from the following limitations. First, human 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,

*This work was performed when Dongfei Yu was visiting Microsoft Research as a research intern.

which pose grand challenges for deep image understanding due to the well-known semantic gap. Second, visual question answering requires fine-grained spatial inference because some answers can be only inferred from highly-localized image regions for "what" and "where" questions.

To deal with the challenges, the state-of-the-art approaches proceed the research on VQA along two independent dimensions. First, some methods develop the high-level semantic representation for images by introducing semantic concepts, image captions or even external knowledge base into the typical CNN-RNN framework [30, 31]. Second, others focus on using region-based features to discover the most important regions to answer a question [9, 13, 17, 20, 26, 33, 34]. However, previous research still ignores using semantic attention to select the most discriminative concepts for a natural language question and using the explicit spatial encoding for image regions.

To simultaneously learn semantic and spatial representation from images, we unify the two dimensions into a holistic learning framework. Specifically, we propose a novel multi-level attention network (MLAN) for visual question answering by highlighting both question-related semantic concepts and local image regions in end-to-end training. Figure 1 shows the advantages of the proposed MLAN by an intuitive example. The proposed MLAN consists of three major components. First, semantic attention attends on high-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. Second, spatial attention is proposed to infer the image regions which can be attended by questions. Local region representations are first extracted from convolutional layers in CNN and further fed into a bidirectional RNN model by a pre-defined order. Such a design enables spatial information of a region to be encoded from surrounding context. Attention scores for each region are further obtained by a multiple layer perceptron (MLP) with the input of both context-aware visual representation and question representation. Third, joint learning incorporates attended regions, attended concepts and question features by element-wise multiplication, followed by a softmax layer to predict the most possible answer from an answer set. We summarize the main contributions as follows:

- We address the challenges of automatic visual question answering by jointly learning multi-level attention, which can simultaneously reduce the semantic gap from vision to language and benefit fine-grained inference in VQA tasks.
- We introduce a novel spatial encoding approach for visual attention, which extracts the context-aware visual features from ordered image regions by a bidirectional RNN model.
- We conduct experiments on two widely-used VQA datasets [2, 37], and obtain significant performance gains over both visual-only and semantic-only attention models.

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.

CNN-RNN. 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 [2, 10, 18, 24]. They extract visual features from images via pre-trained convolutional neural networks (CNNs) and encode questions by recurrent neural networks (RNNs). Ren *et al*. [24] took their inspiration from [28], where the image was treated as the first token and fed into RNNs together with descriptions to learn visual-semantic embedding. Instead of seeing image once, Malinowski *et al*. [18] passed the image into RNNs at each time when encoded the question, which is similar to the framework of [5] in automatic image captioning task. Gao *et al*. [10] adapted m-RNN models [19] to deal with VQA task in the multi-lingual setting. Agrawal *et al*. [2] released a large and human-annotated VQA dataset and evaluate several baseline models and human-level performance on this dataset, which accelerated advances in this task. Despite these early approaches show promising performance in VQA task, it tends to fail on novel instances and highly rely on questions (do not change the answer across images) [1].

Visual Attention. Visual attention mechanism is brought into VQA to address "where to look" problems. Question-guided visual attention uses semantic representation of a question as query to search for the regions in an image that are related to the answer [9, 13, 17, 26, 34]. Two types of soft attention mechanism are well explored in visual question answering task. The first type concatenates the question representation with each candidate region and then put 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 ways of inputs [33]. Yang *et al*. [34] propose a stacked attention model which queries the image multiple times to infer the answer progressively. Lu *et al*. [17] exploit a question-image co-attention strategy to attend not only related regions in images but also important words in questions. Recently, Nam *et al*. [20] proposed Dual Attention Networks, which refined the visual and textual attention via multiple reasoning steps.

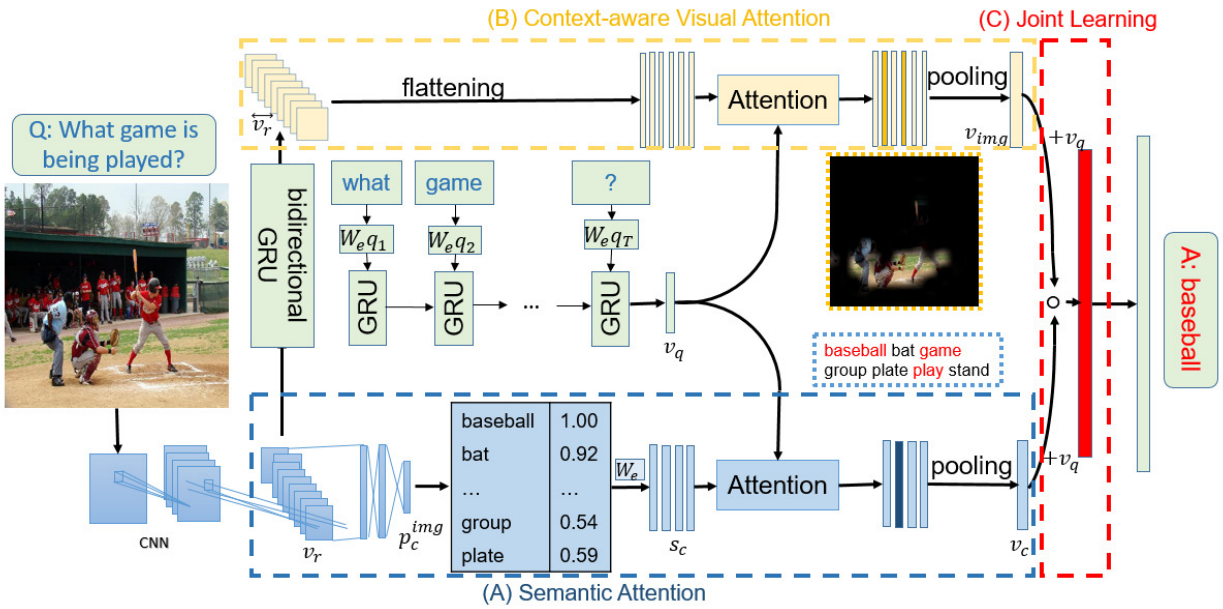

Figure 2. Overall framework of multi-level attention networks. Our framework consists of three components: (A) semantic attention, (B) context-aware visual attention and (C) joint attention learning. Here, we denote by v_q the representation of the question Q, by v_{img}, v_c the representation of image content on the visual and semantic 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.

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. In [9], Fukui *et al*. incorporate a powerful feature fusion method into visual attention, and achieve impressive results in VQA task. However, they have to keep a much higher dimension after fusion at cost of more computation and storage.

High-level Concepts. There is another branch also showing a promising direction to address VQA problems. Instead of low-level or middle-level visual features, they leverage high-level concepts[7, 8, 29], image captions or even visual story[16], external knowledge base [30, 31]. Each concept corresponds to a word mined from the training image descriptions and represents some kinds of attributes about the content of the image. These concepts act as semantic units between natural language and visual recognition, allow us to exchange information between the two modalities [25]. However, the spatial information is completely lost in the procedure of high-level concepts detection, which leads to inferior performance on VQA task.

3. Multi-level Attention Networks

To simultaneously exploit higher-level semantic information and spatial information, we propose a novel multi-level attention network. The overall framework is presented in Figure 2. Our framework consists of three major components. Component (A), which is defined as semantic attention, aims at finding question-related concepts from the image. Component (B), which is defined as context-aware visual attention, aims at finding question related regions and learning visual representation of these regions. Component (C) is designed to incorporate information from different-level layers in the CNN by joint attention learning. These three components are joint optimized end-to-end, which bridges the semantic gap between language and vision, and learns fine-grained representation from image regions.

3.1. Semantic Attention

Semantic attention aims at finding important concepts 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 which 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 modality 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 [25]. Previous works on image/video captioning [6, 22, 23, 35, 36] and visual question answering [30, 31] 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 on concepts, which should be not only relevant to images, but semantically close to questions. We achieve these goals by 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. Similar to [30], we first build a concept vocabulary, where each concept is defined as a single word. The top highly-frequent words with the number of C from the question-answer training pairs are collected in the concept vocabulary after stop words removal. Besides, a multi-label image dataset based on these concepts is constructed based on COCO image captioning dataset [15], 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:

$$p_c^{img} = f_c(I). \tag{1}$$

In the second step, we train an attention network to measure the semantic relevance between each concept in the vocabulary and the question. At first, we represent the question by a recurrent neural network. Specifically, given the question $Q = [q_1, q_2, ..., q_T]$, 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. For each time step t, we feed the embedding vector x_t of word q_t to a Gate Recurrent Unit (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:

$$x_t = W_e^q q_t, \tag{2}$$

$$h_t = GRU(x_t, h_{t-1}), \tag{3}$$

$$v_q = h_T. \tag{4}$$

Besides, we use the same vocabulary and embedding matrix for our concepts and questions, 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 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:

$$s_c = W_e^c(W_e^q c), \tag{5}$$

where c is the one hot vector representation of the concepts, W_e^q is the embedding weights shared with the question model, W_e^c 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 semantic attention weights of the concept c as the multiplication of the concept-image relevance p_c^{img} and the concept-question relevance p_c^q, which is given by:

$$p_c^q = \text{sigmoid}(v_q \cdot s_c), \tag{6}$$

$$M_c = p_c^{img} p_c^q, \tag{7}$$

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, M_c is the semantic 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:

$$v_c = \sum_{i=1}^{C} M_c(i) s_c(i). \tag{8}$$

3.2. Context-aware Visual Attention

Although semantic attention bridges the semantic gap between the questions and images, it ignores the spatial information in images, which is important to represent the spatial context for image regions, and thus is crucial in the visual question answering task. Hence visual attention has been widely used in recent VQA frameworks, due to its success on fine-grained visual representation and visualization. Compared with human attention, recent work [4] finds current VQA attention models do not seem to be "looking at" the consistent regions as human do. One of the possible problems in current attention model is that they usually search for image regions one after another, 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 action and position relations. 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 those regions at bottom of the "cat" should be looked at and understood. In order to address this issue, we propose a context-aware visual attention mechanism into our VQA framework.

Specifically, we first incorporate the context information into the representation from each region by a bidirectional GRU encoder, which is illustrated in figure 3. We use the fine-tuned CNN model for concept detection from

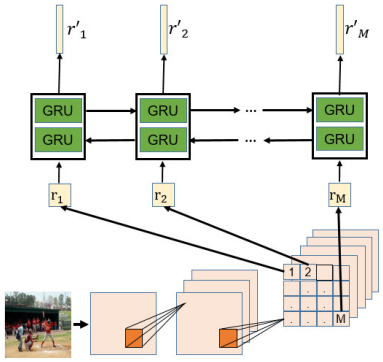

Figure 3. An illustration of the context-aware visual representation for image regions by bidirectional GRU. Regions in convolutional feature maps are encoded into GRU with the order of left-to-right and top-to-bottom.

the precious step to extract visual features for 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. We denote these visual representation on each region as $\{v_r, r = 1,2,...,R\}$, where v_r represents the feature vector of the r^{th} ordered region. We feed these feature vectors into the bidirectional GRU and combine the output from the forward and backward direction at each step to form a new feature vector for each region, which is given by:

$$\overleftrightarrow{v_r} = GRU^f(v_r) + GRU^b(v_r) \tag{9}$$

where $\overleftrightarrow{v_r}$ 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 information from surrounding regions. We set the dimension of the hidden state in each GRU to the same with the question vector.

Second, we assign each region an attention score for modeling the relation between the region and the question. Different from semantic attention, which measures the semantic similarity between the question and the concept word by the dot-product of two vectors, we align the question and each region by element-wise multiplication of two vectors, and then fed them into a multiple layer perceptron (MLP). Such a design enables the automatic learning of attention function by parameter optimization in MLP. More specifically, we search for regions via multi-step reasoning as [34]. The main differences come from two-fold. 1) We use 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. 2) 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 the section 4.4 demonstrates our assumption. Specifically, we formulate our visual attention process as:

$$h = \tanh\big((W_Q v_q + b_Q) \otimes (W_I \overleftrightarrow{v_r} + b_I)\big), \tag{10}$$

$$M_r = \text{softmax}(W_p h + b), \tag{11}$$

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, M_r is the attention weights of image regions.

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

$$v_{img} = \sum_{i=1}^{R} M_r(i) \overleftrightarrow{v_r}(i). \tag{12}$$

In practice, we repeat the above process once as in [34], using the addition of question feature and attended region feature as guide. We ignore the details here for concision.

3.3. Joint Attention Learning

We use questions as 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 semantic attention. The two-level attention is combined by fusion of their attended representation. Particularly, we first add question vector into attended image features extracted from different layers, then we use a element-wise multiplication to combine the two types of attentions 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:

$$u = (v_q + v_{img}) \circ (v_q + v_c), \tag{13}$$

$$p_a = \text{softmax}(Wu + b), \tag{14}$$

where we denote $\circ$ as the element-wise multiplication between two vectors. v_q, v_{img},v_c are the representation of question Q, the attended visual representation of Image I, and attended semantic representation of concept C, respectively. u is the joint representation from question, image and concepts, which are extracted from the image. W and b is the parameter of the last full connected layer, 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*., VQA and Visual7W dataset, due to large amount of training instances and the diversity of question types.

VQA is a large-scale visual question answering dataset which contains 204,721 images from the COCO dataset and a newly created abstract scene dataset which contains 50,000 scene images. We evaluate our model on this dataset for only real images. For each image in VQA dataset, three questions are annotated, and each question has 10 answers from 10 different annotators. We report our results on two different tasks, which are open-ended and multiple-choice tasks. In open-ended task, we select the answer with the highest activation from all possible outputs, and in multiple-choice task, we pick the answer that has the highest activation from the given choices. We collect the most frequent 3000 answers in training data as candidate answer set. We evaluate the proposed the approach not only on validation dataset, but on a test server, which is provided for blind evaluation in the test set for fair comparison [2].

Visual7W is a more recent VQA dataset built by [37], which is a subset of Visual Genome [3] (the largest visual QA dataset to date with 1.7 million QA pairs). Visual7W contains 327,939 question-answer pairs on 47,300 COCO images. Each question-answer pair is associated with 4 human-generated multiple-choices, and only one of them is the correct answer. There are two major highlights on Visual7W. First, Visual7W provides dense annotations on object-level groundings for establishing an explicit link between QA pairs and image regions. Second, Visual7W allows pointing questions with visual answers, where the correct answer is one of four image regions. We evaluate our model only in multiple-choices setting on this dataset.

4.2. Evaluation Metrics

Visual QA is formulated as multi-class classification problem on both datasets. We follow the evaluation metrics as the baseline approaches on the two datasets. For VQA dataset, [2] 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, and evaluate our best model on both the test-dev and test-standard set. For open-ended task, [2] use a voting mechanism to score the accuracy of a predicted answer:

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

where *ans* is the answer predicted by visual QA models. For Visual7W dataset, we use the evaluation code released by [37], supposing the model is correct on a question if it selects the correct answer candidate. Accuracy is used to measure the performance.

Table 1. Ablation model on test-dev set. The first three models only utilize semantic attention, while the middle three models only perform visual attention. MLAN denotes our full model which applies attention on multi-level representation of images.

Ablation Model	Accuracy
Att-CNN + LSTM [30]	55.57
Q + Concept	56.62
Q + Semantic Attention	59.28
SAN [34]	58.68
Q + Visual Attention	62.29
Q + context-aware Visual Attention	62.50
MLAN (Ours)	**63.69**

4.3. Experiment Setting

We show our experimental settings, hyper-parameters and training process here. For question model, we use the natural language toolkit NLTK[1] to tokenize questions, cast all words into lowercase, and only keep those words appearing at least twice in the train-val set. We don't make any additional preprocessing to those words, *e.g*. removing stop words, stemming. Finally, we get a question vocabulary with 9853 words in VQA dataset. As mentioned in section 3.1, a single layer GRU is used to encode the question, which has 620-dimension word vectors and 2400-dimension hidden states. We take the last hidden state of the GRU layer as the question representation, so that the dimension of question feature vector is 2400.

For concepts model, we select the most frequent 256 words appearing in question-answer training pairs as our concept vocabulary after removing stop words. We detect concepts from images by taking the activation of the last layer of ResNet model [11] fine-tuned on our multi-label dataset derived from MSCOCO dataset. There are two major differences in our concept detector from [30]. We use a more powerful classification model, *i.e*. ResNet with 152 layers pre-trained on ImageNet, instead of VGGNet with 19 layers [27]. Besides, we use the most common loss function "SigmoidCrossEntropyLoss" in 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*. 2400 dimensions. We project question vector and concept vector to the 512-dimension space, and then perform attention on concepts.

For image model, we extract visual features from the last convolutional layer (*i.e*. "res5c") from the same ResNet-152 model with the concept detection. Each feature vector has a dimension of 2048 and corresponds to a 32×32 pixels

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

Table 2. Comparison results on VQA dataset. We divide compared approaches into five categories based on different attention mechanisms. Category I does not use any attention. Category II uses only visual attention. Category III extracts high-level concepts for image representation. Category IV applies attention on both images and questions. Category V includes different variations of our approach.

	Approach	test-dev					test-standard				
		Open Ended				MC	Open Ended				MC
		Yes/No	Number	Other	All	All	Yes/No	Number	Other	All	All
I	LSTM Q + I [2]	78.9	35.2	36.4	53.7	57.2	79.0	35.6	36.8	54.1	57.8
	deeper + norm [2]	80.5	36.8	43.1	57.8	62.7	80.6	36.5	43.7	58.2	63.1
	DPPnet [21]	80.7	37.2	41.7	57.2	-	80.3	36.9	42.2	57.4	-
II	SAN [34]	79.3	36.6	46.1	58.7	-	-	-	-	58.9	-
	FDA [13]	81.1	36.2	45.8	59.2	-	-	-	-	59.5	-
	DMN+[32]	80.5	36.8	48.3	60.3	-	-	-	-	60.4	-
	MCB+Att. [9]	82.2	37.7	54.8	64.2	68.6	-	-	-	-	-
	MCB + Att. + GloVe [9]	82.5	37.6	55.6	64.7	69.1	-	-	-	-	-
	MCB + Att. + GloVe + VG [9]	82.3	37.2	**57.4**	**65.4**	**69.9**	-	-	-	-	-
III	AC [31]	79.8	36.8	43.1	57.5	-	79.7	36.0	43.4	57.6	-
	ACK [31]	81.0	38.4	45.2	59.2	-	81.1	37.1	45.8	59.4	-
IV	HieCoAtt [17]	79.7	38.7	51.7	61.8	65.8	-	-	-	62.1	66.1
	DAN [20]	83.0	39.1	53.9	64.3	69.1	82.8	39.1	54.0	64.2	69.0
V	MLAN (ResNet)	82.9	39.2	52.8	63.7	68.9	-	-	-	-	-
	MLAN (ResNet, train+val)	**83.8**	40.2	53.7	64.6	69.8	**83.7**	40.9	53.7	64.8	69.9
	MLAN (ResNet, train+val +VG)	81.8	**41.2**	56.7	**65.3**	**70.0**	81.3	**41.9**	**56.5**	**65.2**	**70.0**

region of the input image. As with attention on semantic level, we embed the 2048-dimension feature vector to 2400-dimension by bidirectional GRU, project image and this context-aware representation into the same 512-dimension space, and then perform attention on visual representation.

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 one of ten of the previous one every 5 epochs. In addition, gradient clipping technology and dropout are exploited in training. For visual7W dataset, we use the exactly same parameter setting and training options with the VQA dataset. We evaluate our model only in multiple-choices setting, and split the dataset into train, validation and test following [37].

4.4. Ablation model

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

- Att-CNN + LSTM [30]: the attribute representation as the first input of LSTM, then following the question
- Q + Concept: a simple version of semantic attention, taking the output of concept detector as the attention weights, independent on the question
- Q + Semantic Attention: the first component of our model, taking the relation of concepts with both image and question into the attention weights
- SAN [34]: a visual attention model similar with our second components
- Q + Visual Attention: our visual attention model without context-aware visual representation
- Q + context-aware Visual Attention: the second components of our model, removing semantic attention from the full model
- Q + Multi-level Attention: our full model, fusing attention on different level image representation

We report the performance of our ablation models on test-dev set of VQA dataset in Table 1. These models are trained on the training dataset and half of validation set, as in [34]. Further analysis will be given in next section.

4.5. Result and Analysis

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

The first three rows in Table 1 compare our semantic attention model with those models using high-level concepts but without attention mechanism. 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 image and remove noisy information irrelevant with the question.

The middle three rows in Table 1 proves our two contributions on visual attention mechanism. We use element-wise multiplication to replace addition in SAN[34] model and get better performance, which supports our assump-

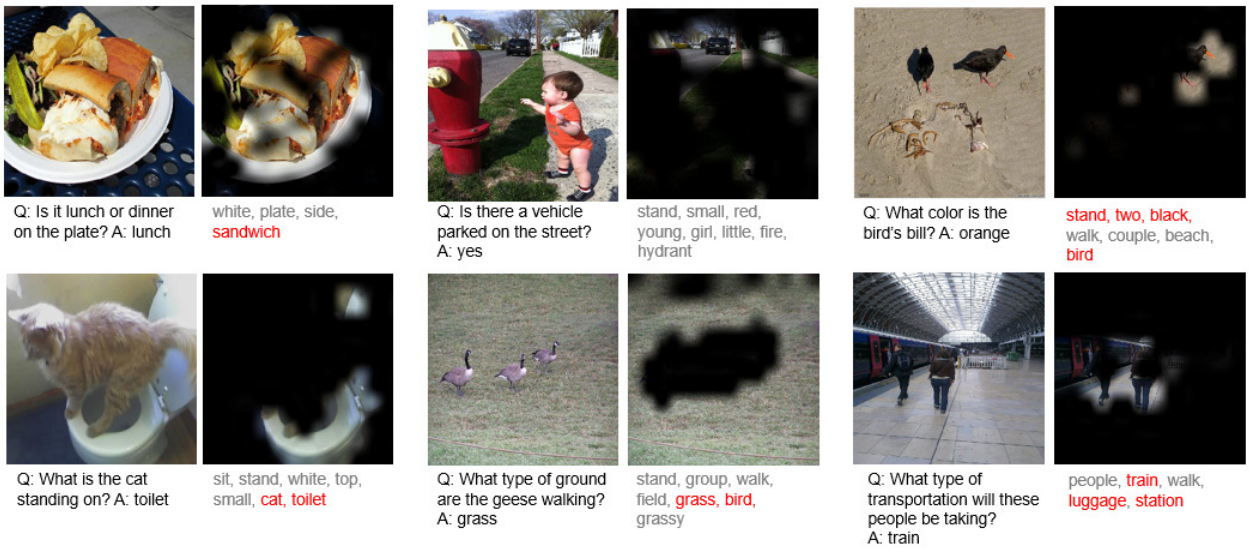

Figure 4. qualitative results from visual question answering with attention visualization. Both image regions related to the question and high-level concepts are highlighted. Examples in the first row shows correct attended image regions lead to the true answer, while the second row shows those cases where answer can be found directly from attended concepts.

tion 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. The promotion is marginal than we think. We conjecture that there might be two reasons. First, our current context encoding scheme suffers from long-term dependency problems by bidirectional GRU and is not symmetric for surrounding regions in horizontal and vertical direction because bidirectional GRU can only model a sequence rather than a 2D spatial map. Second, most images from COCO only contain a few objects, therefore, the interaction among objects is not so common as the natural scenario. We will verify this in our future work.

The last row in Table 1 joins different-level attention into one unifying framework and achieves significant improvement compared with any single-level attention model. This demonstrates attention mechanism at different level image features are complementary and could benefit each other.

We compare our model with the state of art methods on two large datasets. The results are showed in Table 2 on VQA dataset and Table 3 on Visual7W dataset respectively. For a fair comparison, we report the results using the single model with several setting. [9] achieve a comparable performance with ours when they add glove tricks and additional training data. However, their method uses a much higher dimension fusion method (16,000 dim v.s. 2400 dim), and drop almost over 1% if they use comparable dimensional features. Their model has to make a trade-off between effectiveness and efficiency. [17] and [20] are two methods also exploiting 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 than the question itself.

Table 3. Results on Visual7W dataset. We report the independent and average accuracy on six question types, including "what, where, when, who, why and how."

Method	Wht.	Whr.	Whn.	Who	Why	How	Avg
LSTM-Att [37]	51.5	57.0	75.0	59.5	55.5	49.8	54.3
MCB+Att. [9]	60.3	70.4	79.5	69.2	58.2	51.1	62.2
MLAN (Ours)	60.5	71.2	79.6	69.4	58.0	50.8	62.4

5. Conclusion

We propose a novel Multi-level Attention Network to join visual attention and semantic attention into an end-end framework to address automatic visual question answering. Visual attention enables fine-grained visual understanding queried by questions while semantic attention narrows the domain gap between questions and images. Our model makes use of the complementarity of attention mechanism on different level representation. Extensive experiments on two large dataset demonstrate we not only outperforms any single-level attention model, but also achieves top results via a simple but effective framework. Future work includes further exploring on spatial encoding with context information, attention on sentence-level representation and better fusion methods to join different level attention.

References

[1] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, 2016.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. In *ICCV*, 2015.

[3] A. Das, H. Agrawal, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. In *IJCV*, 2016.

[4] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *EMNLP*, 2016.

[5] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[6] 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. From captions to visual concepts and back. In *CVPR*, 2015.

[7] J. Fu, J. Wang, Y. Rui, X.-J. Wang, T. Mei, and H. Lu. Image tag refinement with view-dependent concept representations. *IEEE T-CSVT*, 25(28):1409–1422, 2015.

[8] J. Fu, Y. Wu, T. Mei, J. Wang, H. Lu, and Y. Rui. Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging. In *ICCV*, 2015.

[9] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.

[10] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[13] I. Ilievski, S. Yan, and J. Feng. A focused dynamic attention model for visual question answering. In *ECCV*, 2016.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceddings of the IEEE*, 86(11):2278–2324, 1998.

[15] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[16] Y. Liu, J. Fu, T. Mei, and C. W. Chen. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *AAAI*, pages 1445–1452, 2017.

[17] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.

[18] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015.

[19] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). In *ICLR*, 2015.

[20] H. Nam, J. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *arXiv:1611.00471*, 2016.

[21] H. Noh, P. H. Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, 2016.

[22] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[23] Y. Pan, T. Yao, H. Li, and T. Mei. Video captioning with transferred semantic attributes. In *arXiv preprint arXiv:1611.07675*, 2016.

[24] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015.

[25] M. Rohrbach. Attributes as semantic units between natural language and visual recognition. In *arXiv:1604.03249*, 2016.

[26] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[28] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[29] J. Wang, J. Fu, T. Mei, and Y. Xu. Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. In *IJCAI*, 2016.

[30] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR*, 2016.

[31] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016.

[32] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and texual question answering. In *ICML*, 2016.

[33] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016.

[34] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[35] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *arXiv preprint arXiv:1611.01646*, 2016.

[36] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *CVPR*, 2016.

[37] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016.

[38] C. L. Zitnick, A. Agrawal, S. Antol, M. Mitchell, D. Batra, and D. Parikh. Measuring machine intelligence through visual question answering. *AI Magazine*, 37(1):63–72, 2016.