Jiang AW, Liu B, Wang MW. Deep multimodal reinforcement network with contextually guided recurrent attention for image question answering. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 32(4): 738–748 July 2017. DOI 10.1007/s11390-017-1755-6

Deep Multimodal Reinforcement Network with Contextually Guided Recurrent Attention for Image Question Answering

Ai-Wen Jiang [1], *Member, CCF*, Bo Liu [2], and Ming-Wen Wang [1,*], *Senior Member, CCF*

[1] *College of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China*
[2] *College of Computer Science and Software Engineering, Auburn University, Auburn, AL36849, U.S.A.*

E-mail: jiangaiwen@jxnu.edu.cn; boliu.umass@gmail.com; mwwang@jxnu.edu.cn

Received December 19, 2016; revised May 26, 2017.

Abstract Image question answering (IQA) has emerged as a promising interdisciplinary topic in computer vision and natural language processing fields. In this paper, we propose a contextually guided recurrent attention model for solving the IQA issues. It is a deep reinforcement learning based multimodal recurrent neural network. Based on compositional contextual information, it recurrently decides where to look using reinforcement learning strategy. Different from traditional "static" soft attention, it is deemed as a kind of "dynamic" attention whose objective is designed based on reinforcement rewards purposefully towards IQA. The finally learned compositional information incorporates both global context and local informative details, which is demonstrated to benefit for generating answers. The proposed method is compared with several state-of-the-art methods on two public IQA datasets, including COCO-QA and VQA from dataset MS COCO. The experimental results demonstrate that our proposed model outperforms those methods and achieves better performance.

Keywords image question answering, recurrent attention, deep reinforcement learning, multimodal recurrent neural network, multimodal fusion

1 Introduction

Question answering (QA) is a well-defined topic in natural language processing area. It is traditionally a pure NLP (natural language processing)-related problem as both the question and facts that answers lie in are in the form of language. However, as the research goes deeper, people are not satisfied with the textual-only QA, and inclined to extend it to vision area. When textual facts are alternated by vision facts, there consequently comes a newly defined research topic — image question answering[1-4].

Image question answering takes an image and an image content related question as inputs, and directly infers a reasonable answer as an output automatically. In IQA (image question answering), grounding information on answering questions lies in the vision facts from images. Therefore, in order to answer visual questions correctly, the IQA system needs to understand both images and questions. This kind of multimodal working mode is closely related to the cognition behavior of human brain. It is a very challenging task, and has attracted great interests from computer vision and natural language processing areas.

A large number of architectures ranging from symbolic to neural based framework have been proposed to solve the IQA problem recently. Depending on the visual features used, they are categorized as: 1) explicit visual representations on bounding box surrounding object of interests, 2) holistic image feature, or 3) soft attention to combine regional information. Bounding boxes or soft attention based methods give much concern on local visual information, neglecting global context for understanding holistic semantics of an image. Global representation based methods extract full image content in a high level, inevitably losing important local details for answer grounding.

Regular Paper
Special Issue on Deep Learning
This work was supported by the National Natural Science Foundation of China under Grant Nos. 61365002 and 61462045, and the Science and Technology Project of the Education Department of Jiangxi Province of China under Grant No. GJJ150350.
*Corresponding Author

We believe that both global contextual information and local conceptual details are helpful to correctly answer visual questions. When doing the IQA task, we generally perform holistic image understanding first based on global context. Then, based on the learned semantic information from specific questions, we pay attention to some relevant details. At last, we composite the global contextual understanding and local details as reliable vision evidence for question answering. We identify this sequential way as "where to look under contextual guidance".

Under the presumption of sequential and hierarchical semantic understanding, we propose a contextual-guided recurrent attention model for IQA in this paper. Our proposed system is a deep multimodal recurrent neural network. At each step, based on current compositional contextual information, it decides where to look with reinforcement learning strategy, extracting multi-resolution crops on predicted locations on feature mapping. The extracted regional information is memorized into its internal representation which is recurrently combined with the question and global visual context to locate next informative region. The process continues until reaching a predefined maximum iteration for regional observation.

Compared with general "static" soft attention methods, our proposed mechanism is dynamic. Under the rewards of reinforcement strategy, our attentions can be more purposeful towards IQA's objectives. Furthermore, compared with purely global representation, our learned compositional information includes both global context and informative local details, which bring many benefits to solve the IQA problem.

We train an end-to-end system for answer prediction and grounding, and compare it with several state-of-the-art methods on two public IQA datasets, COCO-QA[1] and VQA from dataset MS COCO[3] (MSCOCO-VQA for short). The experimental results demonstrate that our method can achieve better performance.

The main contribution of this work is that we have proposed a deep reinforcement learning based dynamic recurrent attention mechanism for the IQA task. The contextually guided attention mechanism can learn effective sampled "hard" attention trajectories through reinforcement learning rewards. The local attention trajectory possesses complementary detailed information to global context, which can purposefully improve the accuracy of IQA.

The rest of this paper is organized as follows. In Section 2, related work is introduced. In Section 3, details of our proposed network are described. In Section 4, comprehensive experimental comparisons are presented. Conclusive remarks and discussion are given in Section 5.

2 Related Work

2.1 Local Spatial Attention Based Methods

Visual attention based models have been extensively explored for IQA in recent years. However, so far, all attention models used in IQA literatures mainly focus on a kind of soft spatial attention mechanism which takes inspiration from the model proposed in [5] for image captioning. The main differences among the methods of this category lie in the functions from which soft attention weights are computed.

Typically, the soft attention mechanism produces a spatial map highlighting image regions relevant to a question's answers. It aligns question sentence/words with candidate visual regions. Depending on questions, it predicts spatial latent coefficients to weight the localized convolutional mappings in convolutional neural network. The resulted weighted representation rather than full image is then used as a basis for answering questions.

Yang *et al.*[6] proposed a stacked attention network to highlight question-relevant regions using multiple attention layers.

In [7], the spatial mappings "inception 5b/output" of GoogLeNet are used as images' representation. Each word embedding is used to perform fine-grained alignment between images and questions in its first attention hop.

Chen *et al.*[8] introduced an attention-based configurable convolutional neural network to locate attentions based on input queries. It generates a configurable convolutional kernel through question embedding, and convolves it with image features to generate attention mappings.

In [9], top-ranked 99 edge boxes together with a full image are used as candidate visual regions related to a question's answers. The region selection layer generates attention weights by applying softmax on the inner product of image features and text features.

The attention used in [10] depends upon the previous hidden state of recurrent unit and the convolutional features. It also accepts a two-layer feed-forward network to predict the attention weights.

Ilievski *et al.*[11] employed off-the-shelf object detector to identify important regions and fuse the infor-

mation from the regions and global features via LSTM (long short-term memory). The criteria for selecting objects are based on similarity scores between the question words and the labels of objects.

Kumar *et al.*[12-13] proposed a dynamic memory network in which the update gate of GRU (gated recurrent unit) is replaced by an attention gate. The computation of the attention gate is similar to traditional multiple layer perception network.

Lu *et al.*[14] proposed a co-attention mechanism which jointly explores visual attention and question attention at different semantic levels.

Fukui *et al.*[15] proposed to utilize multimodal compact bilinear pooling for efficient and expressive multimodal feature fusion, rather than traditional element-wise multiplication, addition or concatenation. Their used attention mechanism is similar to the soft one used in [6].

2.2 Global Features Based Methods

In this category, image features used are all from fully connected layer in convolutional networks. Global features extract image contents in cost of spatial information lost.

Noh *et al.*[16] proposed a dynamic parameter layer dealing with multimodal combination for visual question answering. Hashing trick is employed to predict the weights in the dynamic parameter layer, avoiding explosion of parameters' scales.

Kim *et al.*[17] proposed multimodal residual networks for visual question answering. The visual features used are global.

Andreas *et al.*[18] proposed to answer image-related questions through collections of jointly-trained neural "modules" based on linguistic structure.

Wang *et al.*[19] emphasized the importance of large external knowledge on developing structured representation of image content.

Ma *et al.*[20] employed multimodal convolutional network to learn the interactions between image and question representations. They treated an image as an individual semantic component and question words as consecutive semantic components.

2.3 Recurrent Attention Model with Reinforcement Learning

Mnih *et al.*[21-22] proposed a recurrent attention model for object recognition. The recurrent model extracts information from an image by adaptively selecting a sequence of regions or locations. As we know, humans always focus attention selectively on parts of the visual space to acquire information when and where it is needed, and combine the information from different fixations over time to build up an internal representation of the scene. Therefore, the recurrent attention model has neuroscience and cognitive science basis. Besides, it is also capable of reducing the computation cost by only processing the selected regions at high resolution.

Li *et al.*[23] proposed a object detection model named AC-CNN with multiple stacked LSTM layers. By incorporating multi-level information into the region-based CNN, they reported better object detection performance.

3 Contextually Guided Recurrent Attention Model for IQA

In this paper, we propose a deep reinforcement learning based multimodal recurrent neural network for IQA. It is built around a recurrent attention model with contextual guidance, consisting of several subcomponents. The whole architecture of the proposed model is as shown in Fig.1. For convenience, we use the term "network" to describe its non-linear subcomponents since they are typically multi-layered neural networks in the following content.

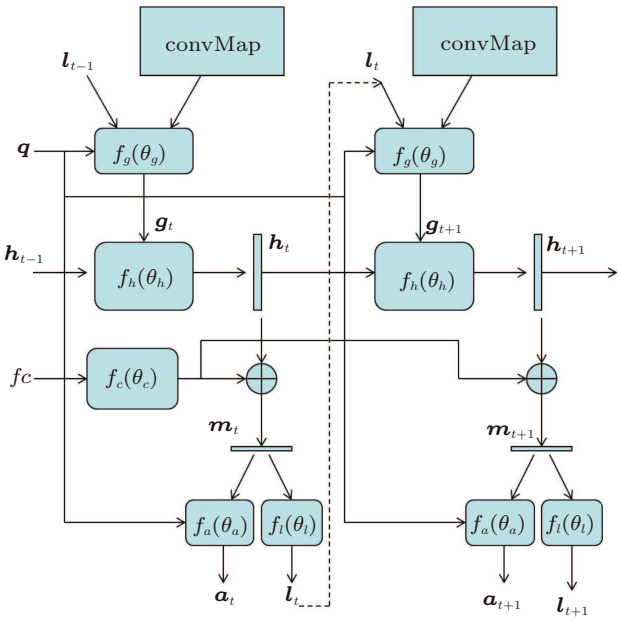

Fig.1. Architecture of deep reinforcement network with contextual-guided recurrent attention.

3.1 Model Architecture

convMap is a spatial feature mapping extracted from CNN's convolutional layer. $\boldsymbol{l}_t$ is the location of regional attention at time t. $f_g(\theta_g)$ is a multimodal glimpse network which combines the visual retina information with question information centered at $\boldsymbol{l}_t$ on convMap. $f_h(\theta_h)$ is a GRU-based recurrent network which is used to memorize the sequentially attended information. $\boldsymbol{h}_t$ is a hidden state accumulating multimodal memories at time t. $f_c(\theta_c)$ is a multilayer network for extracting contextual information. In this paper, it directly takes the global feature "$fc7$" output from CNN's last fully-connected layer. The regional multimodal information and context feature are used to generate compositional feature m_t. $f_l(\theta_l)$ is a location network, making predictions in sequence to next attention location. $f_a(\theta_a)$ is a predicting network, which produces possible answers for visual questions. These subcomponents work collaboratively to solve the IQA problem. In the following parts, we will describe them respectively in details.

3.1.1 Multimodal Glimpse Network $f_g(\theta_g)$

The glimpse network acts as "eyes" of our IQA agent. Its network is as shown in Fig.2.

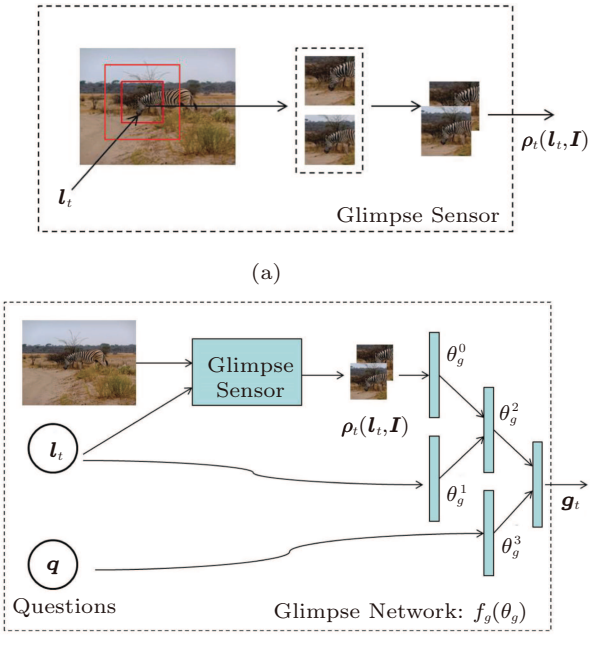

Fig.2. Multimodal glimpse network $f_g(\theta_g)$.

Given location $\boldsymbol{l}_{t-1}$ and input image $\boldsymbol{I}$, the glimpse network uses a glimpse sensor to extract a multi-resolution observation $\boldsymbol{\rho}_t(\boldsymbol{l}_{t-1}, \boldsymbol{I})$ centered at $\boldsymbol{l}_{t-1}$. It extracts patches with pyramid scales and concatenates representations of these patches.

Observation $\boldsymbol{\rho}_t$ and glimpse location $\boldsymbol{l}_{t-1}$ are then mapped into a hidden space using fully-connected layers parameterized by θ_g^0 and θ_g^1 respectively. Another fully-connected layer θ_g^2 is used to combine them. A question is mapped into a space with the same dimension as θ_g^2 defines by using another separate fully-connected layer θ_g^3.

In these cases, we denote θ_g^i to be the weight and bias $\{\boldsymbol{W}_g^i, \boldsymbol{b}^i\}$ of each layer. Information from these two kinds of sources, visual glimpse information and question information, is concatenated to form a question-related glimpse representation $\boldsymbol{g}_t$ through element-wise summation.

$$\boldsymbol{g}^0 = \mathrm{ReLU}\left(\boldsymbol{W}_g^0 \boldsymbol{\rho}_t + \boldsymbol{b}^0\right), \boldsymbol{g}^1 = \mathrm{ReLU}\left(\boldsymbol{W}_g^1 \boldsymbol{l}_{t-1} + \boldsymbol{b}^1\right),$$
$$\boldsymbol{g}_t = \mathrm{ReLU}\left(\boldsymbol{W}_g^2[\boldsymbol{g}^0, \boldsymbol{g}^1] + \boldsymbol{b}^2\right) + \mathrm{ReLU}\left(\boldsymbol{W}_g^3 \boldsymbol{q} + \boldsymbol{b}^3\right),$$

where $\mathrm{ReLU}(x) = \max\{0, x\}$.

3.1.2 Recurrent Attention Network $f_h(\theta_h)$

The recurrent network is the core sub-component, acting as the memory part of our IQA agent. It aggregates information extracted from individual glimpses and combines them in a coherent manner.

We use gated recurrent unit (GRU)[24] as the core network. At each time step t, GRU incrementally combines the glimpse representation $\boldsymbol{g}_t$ with the internal state $\boldsymbol{h}_{t-1}$ at previous time step, and produces a new internal state of model $\boldsymbol{h}_t$.

$$\begin{aligned}
\boldsymbol{h}_t &= (1 - \boldsymbol{z}_t)\, \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \widetilde{\boldsymbol{h}_t}, \\
\boldsymbol{z}_t &= \sigma\left(\boldsymbol{W}_z \boldsymbol{g}_t + U_z \boldsymbol{h}_{t-1}\right), \\
\widetilde{\boldsymbol{h}_t} &= \tanh\left(\boldsymbol{W}_h \boldsymbol{g}_t + U_h\left(\boldsymbol{s}_t \odot \boldsymbol{h}_{t-1}\right)\right), \\
\boldsymbol{s}_t &= \sigma\left(\boldsymbol{W}_r \boldsymbol{g}_t + U_r \boldsymbol{h}_{t-1}\right),
\end{aligned}$$

where σ denotes sigmoid function, and $\widetilde{\boldsymbol{h}_t}$ is candidate activation. $\boldsymbol{h}_{t-1}$ is the previous activation state. The update gate $\boldsymbol{z}_t$ decides how much the unit updates its activation. $\boldsymbol{s}_t$ is a set of reset gate and $\odot$ represents an element-wise multiplication.

The activation of internal state $\boldsymbol{h}_t$ summarizes the information extracted from the past consecutive local observations.

3.1.3 Context Network $f_c(\theta_c)$

Global feature fc provides contextual information. In the viewpoint of hierarchial semantic understanding, it can provide a big-picture on the top level. Herein, it can also provide sensible hints on where the potential interested regions are in a given image.

For convenience, we transform the global feature fc to a space $\hat{fc}$ having the same dimension with $\boldsymbol{h}_t$ using one fully-connected layer. We combine the transformed $\hat{fc}$ with internal state $\boldsymbol{h}_t$ to generate a compositional feature $\boldsymbol{m}_t$ by using element-wise summation.

As a result, $\boldsymbol{m}_t$ contains both globally contextual information and local experiences on sequential attentions. The agent can more purposefully attend to the next glimpse based on the compositional feature $\boldsymbol{m}_t$ and reinforcement rewards.

3.1.4 Location Network $f_l(\theta_l)$

The location network acts as a controller that directs attention according to the currently aggregated global and local information. It consists of a fully-connected hidden layer, and takes the compositional feature $\boldsymbol{m}_t$ as input and makes a prediction on where to extract the next image patch for the glimpse network.

In this paper, we encode the real valued glimpse location tuple $\boldsymbol{l}_t$ by using a Cartesian coordinate that is centered at the middle of the input image.

3.1.5 Answer Prediction Network $f_a(\theta_a)$

The answer prediction network, shown in Fig.3, is composed of one hashing-based dynamic parameters layer, one fully-connected hidden layer and a softmax output layer for classification. It predicts a question's answer based on compositional visual information $\boldsymbol{m}_t$ and question information $\boldsymbol{q}$. Depending on the predicted answer, it receives a reward r as a response.

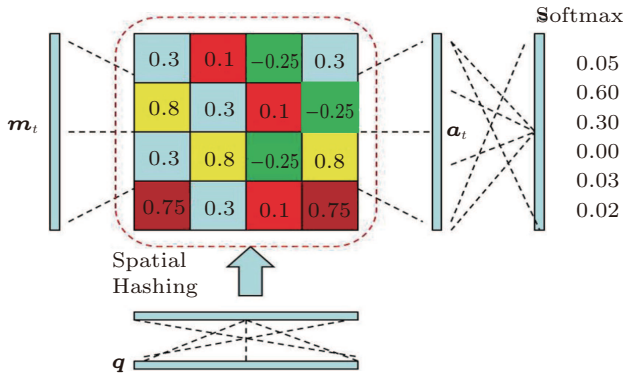

Fig.3. Answer prediction network $f_a(\theta_a)$.

The dynamic parameters layer in this network is similar to the one in [16]. Its weights are determined adaptively by a parameter prediction network. This parameter prediction network is a fully-connected layer. It takes question $\boldsymbol{q}$ as input, and generates a real valued vector, which corresponds to candidate weights for the dynamic parameter layer.

Both the location network and the answer prediction network act as the action network of our IQA agent. After executing an action, the agent receives a new visual observation of the environment at $\{\boldsymbol{l}_{t+1}\}$ and a reward signal r_{t+1}.

The goal of the agent is to maximize the sum of the reward signals $R = \sum_{t=1}^{T} r_t$.

3.1.6 Question Representation

The word sequence of a question is encoded by a recurrent network with GRU cells, as shown in Fig.4.

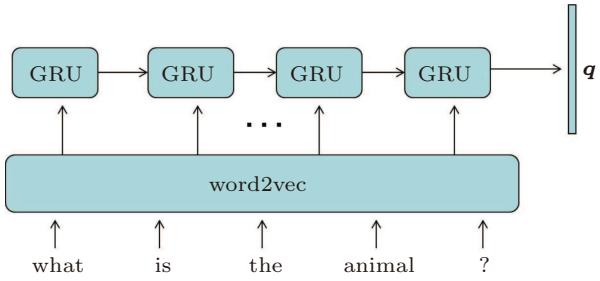

Fig.4. Question representation generation network.

Given question $\boldsymbol{q} = (w_1, ..., w_t, ..., w_N)$, where w_i is the word at position i, we first embed the words to a vector space through a pre-trained word2vec model[25]. Then we feed the embedded word vectors in sequence to the GRU. The last hidden state is taken as the question representation.

3.1.7 Image Representation

Image features are extracted by using a convolutional neural network shown in Fig.5.

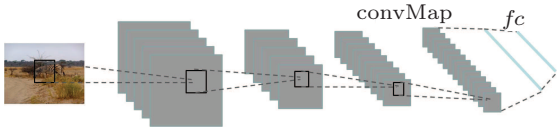

Fig.5. Convolutional neural network.

We apply a glimpse sensor on convolutional mappings, rather than on raw pixels. Compared with the original image in high resolution, convolution mapping has its own advantages. On one side, it reduces the search range for glimpse locator. On the other side, higher level preprocessing avoids noisy information and retains the spatial information of the original image.

The output of the last fully-connected layer retains the global information of the image, and therefore it is taken as the contextual feature fc.

In this paper, we adopt the pre-trained VGG-16 layers network[26] as our CNN model. The glimpse sensor is operated on the last convolution layer $conv5$. Context information is represented by $fc7$ feature.

3.2 Deep Reinforcement Training for the Contextually Guided Recurrent Attention Model

Given an image $\boldsymbol{I}$ and a question $\boldsymbol{q}$, IQA learning is often formulated as a classification problem $a^* = \max_{\boldsymbol{a}} p(\boldsymbol{a}|\boldsymbol{I}, \boldsymbol{q}; \theta)$ with the cross entropy objective function. θ represents network parameters.

In our recurrent attention model, the IQA agent predicts answers conditioned on a series of intermediate latent location variables $\boldsymbol{l} = (l_1, ..., l_t, ..., l_T)$ and corresponding patches. Therefore, we can formulate the learning process as maximizing the likelihood of answers given images and questions by marginalizing over the glimpse locations.

$$\log p(\boldsymbol{a}|\boldsymbol{I}, \boldsymbol{q}; \theta) = \log \sum\nolimits_l p(\boldsymbol{l}|\boldsymbol{I}, \boldsymbol{q}; \theta)\, p(\boldsymbol{a}|\boldsymbol{l}, \boldsymbol{I}, \boldsymbol{q}; \theta).$$

The marginalized objective function can be learned through optimizing its variational lower bound shown as follows.

$$\begin{aligned} &\log \sum\nolimits_l p(\boldsymbol{l}|\boldsymbol{I}, \boldsymbol{q})\, p(\boldsymbol{a}|\boldsymbol{l}, \boldsymbol{I}, \boldsymbol{q}) \\ &\geqslant \sum\nolimits_l p(\boldsymbol{l}|\boldsymbol{I}, \boldsymbol{q}) \log p(\boldsymbol{a}|\boldsymbol{l}, \boldsymbol{I}, \boldsymbol{q}). \end{aligned} \tag{1}$$

Ideally, the deep recurrent attention model should learn to look at locations that are relevant for classifying answers of interest. However, for each glimpse in the glimpse sequence, it is non-trivial to evaluate the exponentially many glimpse locations during training.

One of practical policies is to estimate the locations $\boldsymbol{l}$ by using Monte Carlo samples. Specifically, we assume location $\boldsymbol{l}_t$ is defined by a two-component Gaussian distribution with a fixed variance. The variance is a hyper-parameter, which is set empirically. The output of the location network is taken as the mean $\hat{\boldsymbol{l}}_t$ of the location policy at time t.

$$\tilde{l}_t^m \sim p(\boldsymbol{l}|\boldsymbol{I}, \boldsymbol{q}; \theta) = \mathcal{N}\left(\boldsymbol{l}_t; \hat{\boldsymbol{l}}_t, \Sigma\right), m = 1, ..., M. \tag{2}$$

The attention agent runs the location policy M episodes. At the t-th glimpse of the m-th episode, the location is randomly sampled from the assumed distribution. As a result, the derivatives of the lower bound in (1) with respect to the model parameters θ are formulated as follows.

$$\begin{aligned} \frac{\partial \Omega}{\partial \theta} &= \sum\nolimits_l p(\boldsymbol{l}|\boldsymbol{I}, \boldsymbol{q}) \left(\frac{\partial \log p(\boldsymbol{a}|\boldsymbol{l}, \boldsymbol{I}, \boldsymbol{q})}{\partial \theta} + \log p(\boldsymbol{a}|\boldsymbol{l}, \boldsymbol{I}, \boldsymbol{q}) \frac{\partial \log p(\boldsymbol{l}|\boldsymbol{I}, \boldsymbol{q})}{\partial \theta} \right) \\ &\approx \frac{1}{M} \sum_{m=1}^{M} \left(\frac{\partial \log p\left(\boldsymbol{a}|\tilde{\boldsymbol{l}}^m, \boldsymbol{I}, \boldsymbol{q}\right)}{\partial \theta} + \log p\left(\boldsymbol{a}|\tilde{\boldsymbol{l}}^m, I, \boldsymbol{q}\right) \frac{\partial \log p\left(\tilde{\boldsymbol{l}}^m|\boldsymbol{I}, \boldsymbol{q}\right)}{\partial \theta} \right). \end{aligned} \tag{3}$$

The log-likelihood $\log p\left(\boldsymbol{a}|\tilde{\boldsymbol{l}}^m, \boldsymbol{I}, \boldsymbol{q}\right)$ in the gradient estimator (3) may introduce substantial high variance due to its unbounded range. Especially when the sampled location is off from answer-related regions in the image, the log likelihood will induce an undesired large gradient update that is back-propagated through the rest of the model.

To reduce the variance induced from $\log p(\boldsymbol{a}|\tilde{\boldsymbol{l}}^m, \boldsymbol{I}, \boldsymbol{q})$, we replace it with a 0/1 discrete indicator function R as in [21].

$$\begin{aligned} \frac{\partial \Omega}{\partial \theta} \approx \frac{1}{M} \sum_{m=1}^{M} \Bigg(&\frac{\partial \log p\left(\boldsymbol{a}|\tilde{\boldsymbol{l}}^m, \boldsymbol{I}, \boldsymbol{q}\right)}{\partial \theta} + \\ &\lambda (R^m - \boldsymbol{b}) \frac{\partial \log p\left(\tilde{\boldsymbol{l}}^m|\boldsymbol{I}, \boldsymbol{q}\right)}{\partial \theta} \Bigg). \end{aligned} \tag{4}$$

Hyper-parameter λ balances the scale of the two gradient components. $R^m = \sum_{t=1}^{T} r_t^m$ is the cumulative reward at the m-th episode. In our problem setting, reward R is sparse and delayed. $r_T = 1$ if the answer is predicted correctly after T steps, and $r_T = 0$ otherwise. $\boldsymbol{b}$ is a baseline needed to learn during the recurrently interactions.

It is not difficult to find that the first term in (4) is a normally gradient for classification problems. When viewed as a reinforcement learning update, the second term in (4) is an unbiased estimate of the gradient with respect to θ of the expected reward R under the model action policy. Therefore, we have a practical gradient estimator (4) from (3).

We use a hybrid supervised loss to train the model. The cross entropy loss is optimized to train the answer prediction network f_a. The resulted gradients are back-propagated through the core recurrent attention network and glimpse network. The location network f_l is always trained with REINFORCE[27].

During inference, the model behaves as a feed forward network. As suggested in (1), we use samples of

location sequences $\tilde{\boldsymbol{l}}^m = (\tilde{l}_1^m, ..., \tilde{l}_T^m)$ and average their answer predictions.

$$E_l\left(\log p\left(\boldsymbol{a}|\boldsymbol{I},\boldsymbol{q}\right)\right) \approx \frac{1}{M}\sum_{m=1}^{M}\log p\left(\boldsymbol{a}|\tilde{\boldsymbol{l}}^m,\boldsymbol{I},\boldsymbol{q}\right),$$

where $E_l(\bullet)$ represents expectation.

Therefore, the proposed IQA agent will be evaluated multiple times based on the given question and image, with the predicted answer being averaged.

4 Experiments

4.1 Datasets

We compare our proposed model on two public datasets: COCO-QA[1] and MSCOCO-VQA[3].

4.1.1 COCO-QA

The dataset is automatically generated from captions in the Microsoft COCO dataset. It contains about 78 736 training questions and 38 948 testing questions. These two kinds of questions are based on 8 000 and 4 000 images respectively. There are four types of questions including object, number, color, and location. Each type takes about 70%, 7%, 17%, and 6% of the whole dataset, respectively. All answers in this dataset are single words.

4.1.2 MSCOCO-VQA

The dataset contains 248 349 training questions, 121 512 validation questions and 244 302 testing questions. There are three sub-categories according to answer-types including yes/no, number, and other. Similar to [3], we use the top 1 000 most frequent answers as the possible outputs.

This set of answers covers about 86.54% of the train+val answers. For testing, we train our model on VQA train+val and report the test-dev and test-standard results from the VQA evaluation server. The used evaluation protocol is the same with [3].

Several examples about image question answering are shown in Fig.6.

4.2 Experimental Settings

We implement our proposed model in the framework of the deep reinforcement learning as Mnih *et al.*[21] did on recurrent attention model.

We firstly resize an image to 448×448, and then input the resized one to a CNN. We take the activation from the last convolution layer — *conv*5 of VGG16 Net as input image convMap $\boldsymbol{I}$ for glimpse network. The size of the resulted convolution mapping is 28×28 with 512 channels. We take the activation from the last fully-connected layer $fc7$ as the global context feature.

(a) (b) (c)

Fig.6. Image question answering examples. (a) Q: What color is the signal? A: Red. (b) Q: What animal is this? A: Cat. (c) Q: Are they riding horses both the same color? A: No.

In the glimpse network, given a glimpse location $\boldsymbol{l}_t$, we extract glimpse patches of two different resolution scales $\{x_t^1, x_t^2\}$ on $\boldsymbol{I}$. x_t^1 is the original patch and x_t^2 is a down-sampled coarser image patch. The original patch size is set to 4×4. The scale ratio is set to 2. Patches $\{x_t^1, x_t^2\}$ are max-pooled respectively in space. We use the concatenation of their pooled features $\{\hat{x}_t^1, \hat{x}_t^2\}$ as the glimpse observation, resulting a final glimpse observation vector $\boldsymbol{\rho}_t$ of 1 024 dimensions.

The glimpse observation vector $\boldsymbol{\rho}_t$, the location coordinates $\boldsymbol{l}_t$, and question $\boldsymbol{q}$ pass through a multi-layer perception network as defined in Fig.2. In this paper, the dimensionality of $\boldsymbol{g}^0$ and $\boldsymbol{g}^1$ is 512 while the dimensionality of $\boldsymbol{g}_t$ is 1 024 for the attention model.

The hidden state size of GRU in recurrent attention network $f_h(\theta_h)$ is empirically set to be $d_1 = 1\,000$.

The parameters of GRU for generating a question's representation $\boldsymbol{q}$ are initialized with the skip-thought vector model pre-trained on a book-collection corpus containing more than 74M sentences[28]. Its hidden state size is set to be $d_2 = 2\,400$. During training, the parameters will be fine-tuned accordingly.

The other hyper-parameters in our experiments are the learning rate η and the location variance Σ in (2). They are determined by cross-validation. Empirically, in this paper, we set $\eta = 10^{-4}$ and $\Sigma = 0.11$.

We allow the recurrent attention model taking $T = 4$ glimpses before making an answer prediction. A gradient clip with the threshold 0.1 is locally adopted on each module to handle the gradient explosion. The dropout strategy with 0.5 dropout rate is applied after each nonlinear layer to prevent over-fitting.

During training, the pre-trained CNN model for images is kept frozen. All the other parameters including

the initialized GRU parameters for questions are tuned. The experiments in this paper are run in Torch using NVIDIA GTX TitanX. The mini-batch is set to 100. The optimization used is Adam. The maximum number of training epoch is set to 60 with an early stop. The number of episodes M is empirically set to 5.

Some examples on the glimpse locations attended are shown in Fig.7.

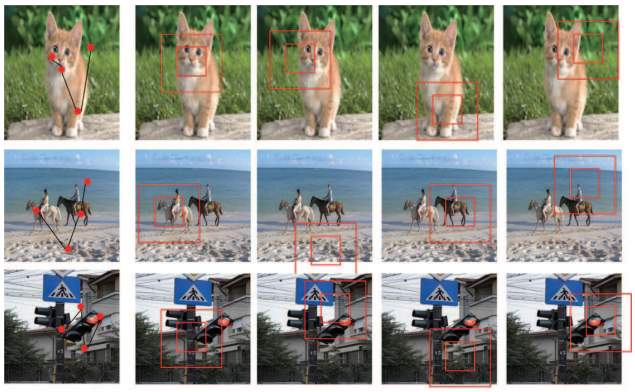

Fig.7. Illustration of sampled glimpse locations learned from questions and answers from Fig.6. (a) Input image with glimpse path overlaid. (b)∼(e) show the respective bounding boxes of the four glimpses the network chooses.

4.3 Results and Analysis

We compare the proposed method with several state-of-the-art methods, including methods using global feature, like VIS+LSTM[1], LSTM Q+I[3], DPPnet[16] and iBOWIMG[29], or methods using soft attention to local regions, such as SAN[6], ABC-CNN[8], IMG-CNN[20] or methods using region box, such as FDA[11].

Classification accuracy and WUPS score[30] are employed to evaluate the performance on COCO-QA. WUPS uses the Wu-Palmer similarity[31] between words based on the WordNet[32] taxonomy.

$$WUPS = \frac{1}{N}\sum_{i=1}^{N}\min\left\{\prod_{\boldsymbol{a}\in A^i}\max_{t\in\Gamma^i}\mu\left(\boldsymbol{a},t\right),\prod_{t\in\Gamma^i}\max_{\boldsymbol{a}\in A^i}\mu\left(\boldsymbol{a},t\right)\right\},$$

where A^i denotes the predicted answer set, and Γ^i denotes the ground truth answer set of the i-th example. $\mu\left(\boldsymbol{a},t\right)$ denotes the Wu-Palmer similarity between the prediction and ground-truth. We use two thresholds 0.9 and 0.0 in our evaluation respectively.

A specifically defined accuracy[3] is employed to reflect human consensus on MSCOCO-VQA, as described in (5). In this criterion, a predicted answer is regarded to be correct in the condition that at least three annotators agree. If the predicted answer is not correct, its evaluation score will depend on the number of agreements.

$$Acc_{VQA} = \frac{1}{N}\sum_{i=1}^{N}\min\left\{\frac{\sum_{t\in\Gamma^i}\sigma\left[a_i=t\right]}{3},1\right\}, \quad (5)$$

where $\sigma\left[a_i=t\right]$ denotes an indicator function.

The evaluation results on COCO-QA and MSCOCO-VQA are shown in Table 1 and Table 2 respectively.

Table 1. Results on COCO-QA (%)

Method	Accuracy	WUPS	
		@0.9	@0.0
SAN(1, LSTM)[6]	59.60	69.60	90.10
SAN(2, LSTM)[6]	61.00	71.60	90.90
DPPnet[16]	61.19	70.84	90.61
IMG-CNN[20]	58.40	68.50	89.67
VIS+LSTM[1]	53.31	63.91	88.25
VIS + $\text{LSTM}_{\text{Full}}$[1]	57.84	67.90	89.52
ABC-CNN[8]	58.10	68.44	89.85
$\text{HieCoAtt}_{\text{VGG}}$[14]	62.90	72.80	91.30
Our method	62.82	71.55	90.75

Table 2. Results on MSCOCO-VQA Open-Ended Task (%)

Method	Test-Dev				Test-STD
	Y/N	Number	Other	All	All
iBOWIMG[29]	76.55	35.03	42.62	55.72	55.89
LSTM Q+I[3]	80.50	36.77	43.08	57.75	58.16
SAN[6]	79.30	36.60	46.10	58.70	58.90
DPPnet[16]	80.71	37.24	41.69	57.22	57.36
$\text{DPPnet}_{\text{FIXED}}$[16]	80.48	37.20	40.90	56.74	-
FDA[11]	**81.14**	36.16	45.77	59.24	59.54
$\text{HieCoAtt}_{\text{VGG}}$[14]	79.60	38.40	**49.10**	60.50	-
Our method	80.65	**38.72**	46.64	**59.76**	**59.94**

Except iBOWIMG[29] with GoogLeNet[33] and FDA[11] with ResNet[34], all the other methods use the pre-trained VGGNet as their CNN model for extracting image features. Generally, features extracted from ResNet have the strongest discriminative power. VGGNet and GoogleNet have similar power, a little weaker than the ResNet. Despite the fact that a model would have better performance if using ResNet, in this work, for fair comparison with most state-of-the-art methods, we still choose to use the prevalent VGGNet as our pre-trained CNN model.

From the experimental results, we can observe that our method similarly achieves the best performance as HieCoAtt[14], and outperforms all the other compared methods in most cases.

In details, it is not difficult to find that the core dynamic parameters prediction layer in DPPnet[16] is embedded as a part of our answer prediction network. If the contextually guided reinforcement attention part is removed, our model degrades to a similar structure to DPPnet. Compared with DPPnet, our method improves the performance by 1.7% on COCO-QA and nearly 2.5% on MSCOCO-VQA. These improvements are owe to the reinforcement training strategy and the complementary information of global context on local regions.

SAN[6] is the newest published method that successfully employs soft attentions to solve the IQA problem. It allows multi-step reasoning on spatial attention. In contrast, we employ a reinforcement strategy to sequentially attend glimpses. The recurrently incorporated "hard" attentions focus more purposefully on regions related to the question's answers, with the guidance of global contextual information. The experimental results also show the effectiveness of our strategy. It outperforms SAN by 1.8% on COCO-QA and 2% on MSCOCO-VQA.

Compared with FDA[11] explicitly using region boxes, our selections of glimpse patches are implicitly related to the question. Owe to the reinforcement reward directly designed for answer prediction, the glimpse network is apt to extract patches at locations closely related to answer-related regions, not limited to objects. Further, in FDA, the sequence order of selected regions input into LSTM is determined according to their corresponding word order in the question. This practice is a little arbitrary, because there do not exist straightforward connections between the two modality data. In contrast, our mechanism is more natural by embedding the sequence selection in a holistic learning process. Our method outperforms FDA about 0.5% on the MSCOCO-VQA open-ended task. Though the improvement is marginal, it should be pointed out that the result of FDA is obtained by using much more discriminative ResNet for visual feature.

The contemporary work HieCoAtt[14] achieves a little better performance than ours. However, we should point out that its contributions are embodied in using a more complex model to fuse vision and text information, rather than attention mechanism, because in its model the used attention is still as same as traditional "static" soft one.

The contributions of HieCoAtt and our work are not in conflict, because we focus on different viewpoints. HieCoAtt utilizes much more question information, while we pay more efforts on proposing a new vision attention model. Moreover, we should also emphasize that our network is less complicated than HieCoAtt's, as our model is capable of reducing computation cost by only processing a few selected regions on high-level convolution mapping. Our model outperforms HieCoAtt in speed on condition of achieving similar performance.

In order to explicitly validate the effectiveness of the proposed method, we have further done an ablation study on two variants. One removes the global contextual feature *fc* from the proposed network, retaining its recurrent "hard" attention parts. The other replaces the recurrent "hard" attention parts with traditional soft one. As a result, the soft attention based variant is shown in Fig.8, where the attention layer is a single layer perception with softmax layer generating attention distribution over the regions as in (6).

$$\begin{aligned} \boldsymbol{h}_I &= \tanh\left(\boldsymbol{W}_I \boldsymbol{v}_I \oplus \left(\boldsymbol{W}_T \boldsymbol{q} + \boldsymbol{b}\right)\right), \\ \boldsymbol{p}_I &= \mathrm{softmax}\left(\boldsymbol{W}_p \boldsymbol{h}_I + \boldsymbol{b}_p\right). \end{aligned} \tag{6}$$

We denote by $\oplus$ the addition of a matrix and a vector. Parameters' size of answer prediction network $f_a\left(\theta_a\right)$ is the same with the one in our full model.

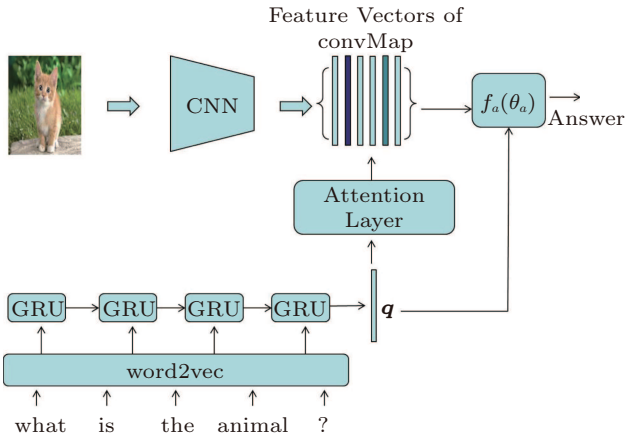

Fig.8. Illustration of the soft attention based variant in ablation studies.

We evaluate the two variants on the COCO-QA dataset. The results are shown in Table 3.

Table 3. Ablation Study on COCO-QA

Method	Accuracy	WUPS	
		@0.9	@0.0
Recurrent hard attention	61.05	70.88	90.20
Soft attention	60.71	70.82	89.96
Our full model	62.82	71.55	90.75

From the ablation study, we can observe that, in the case of our experimental settings, recurrent hard attention gains better performance than the soft attention variant. Our full model in Fig.1 achieves the best results, which demonstrates that for answering visual question, global context and local details indeed have much potential to possess complementary information for different types of answers. Global context can provide helpful guidance for purposefully focusing on informative details.

5 Conclusions

In this paper, we proposed a deep reinforcement learning based multimodal recurrent neural network for image question answering. It is built upon recurrent attention model with the idea of "where to look under contextual guidance". The agent is trained by using reinforcement learning strategy, so that it can attend local regions more purposefully, and at the same time avoid global information lost. We compared the proposed method with several state-of-the-art methods. The experimental results demonstrated that it achieves satisfying performance on public IQA dataset, and outperforms the compared methods in most cases. Currently, our network is mainly focused on single-word answers. In the future, we will do research on the cases of multiple words or open-ended sentences.

References

[1] Ren M Y, Kiros R, Zemel R. Image question answering: A visual semantic embedding model and a new dataset. *arXiv: 1505.02074*, 2015. https://arxiv.org/abs/1505.02074v1, June 2017.

[2] Gao H Y, Mao J H, Zhou J, Huang Z H, Wang L, Xu W. Are you talking to a machine? Dataset and methods for multilingual image question answering. *arXiv: 1505.05612*, 2015. https://arxiv.org/abs/1505.05612, June 2017.

[3] Antol S, Agrawal A, Lu J S, Mitchell M, Batra D, Zitnick L, Parikh D. VQA: Visual question answering. In *Proc. IEEE Int. Conf. Computer Vision*, December 2015, pp.2425-2433.

[4] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A deep learning approach to visual question answering. *arXiv: 1605.02697*, 2016. https://arxiv.org/abs/1605.02697, June 2017.

[5] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. the 32nd IEEE Int. Conf. Machine Learning*, February 2015, pp.2048-2057.

[6] Yang Z C, He X D, Gao J F, Deng L, Smola A. Stacked attention networks for image question answering. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016, pp.21-29.

[7] Xu H J, Saenko K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *arXiv: 1511.05234*, 2015. https://arxiv.org/abs/1511.05234, June 2017.

[8] Chen K, Wang J, Chen L C, Gao H Y, Xu W, Nevatia R. ABC-CNN: An attention based convolutional neural network for visual question answering. *arXiv: 1511.05960*, 2015. https://arxiv.org/abs/1511.05960, June 2017.

[9] Shih K J, Singh S, Hoiem D. Where to look: Focus regions for visual question answering. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016, pp.4613-4621.

[10] Zhu Y K, Groth O, Bernstein M, Li F F. Visual7W: Grounded question answering in images. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016, pp.4995-5004.

[11] Ilievski I, Yan S C, Feng J S. A focused dynamic attention model for visual question answering. *arXiv: 1604.01485*, 2016. https://arxiv.org/abs/1604.01485, June 2017.

[12] Kumar A, Irsoy O, Ondruska P, Iyyer M, Bradbury J, Gulrajani I, Zhong V, Paulus R, Socher R. Ask me anything: Dynamic memory networks for natural language processing. In *Proc. the 33rd Int. Conf. Machine Learning*, June 2016, pp.1378-1387.

[13] Xiong C M, Merity S, Socher R. Dynamic memory networks for visual and textual question answering. In *Proc. the 33rd Int. Conf. Machine Learning*, June 2016, pp.2397-2406.

[14] Lu J S, Yang J W, Batra D, Parikh D. Hierarchical question-image co-attention for visual question answering. In *Proc. Advances in Neural Information Processing System*, Dec. 2016.

[15] Fukui A, Park D H, Yang D, Rohrbach A, Darrell T, Rohrbach M. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv: 1606.01847*, 2016. https://arxiv.org/abs/1606.01847, June 2017.

[16] Noh H, Seo P H, Han B. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016, pp.30-38.

[17] Kim J H, Lee S W, Kwak D H, Heo M, Kim J, Ha J W, Zhang B T. Multimodal residual learning for visual QA. In *Proc. the 30th Conf. Neural Information Processing System*, Dec. 2016.

[18] Andreas J, Rohrbach M, Darrell T, Klein D. Neural module networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016, pp.39-48.

[19] Wang P, Wu Q, Shen C H, van den Hengel A, Dick A. Explicit knowledge-based reasoning for visual question answering. *arXiv: 1511.02570*, 2015. https://arxiv.org/abs/1511.02570v2, June 2017.

[20] Ma L, Lu Z D, Li H. Learning to answer questions from image using convolutional neural network. In *Proc. the 30th AAAI Conf. Artificial Intelligence*, March 2016, pp.3567-3573.

[21] Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In *Proc. Advances in Neural Information Processing Systems*, Dec. 2014.

[22] Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention. *arXiv: 1412.7755*, 2015. https://arxiv.org/abs/1412.7755, June 2017.

[23] Li J N, Wei Y C, Liang X D, Dong J, Xu T F, Feng J S, Yan S C. Attentive contexts for object detection. *arXiv: 1603.07415*, 2016. https://arxiv.org/abs/1603.07415, June 2017.

[24] Chung K, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv: 1412.3555*, 2014. https://arxiv.org/abs/1412.3555, June 2017.

[25] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv: 1301.3781*, 2013. https://arxiv.org/abs/1301.3781, June 2017.

[26] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv: 1409.1556*, 2015. https://arxiv.org/abs/1409.1556, June 2017.

[27] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992, 8(3/4): 229-256.

[28] Kiros R, Zhu Y K, Salakhutdinov R, Zemel R, Torralba A, Urtasun R, Fidler S. Skip-thought vectors. *arXiv: 1506.06726*, 2015. https://arxiv.org/abs/1506.06726, June 2017.

[29] Zhou B L, Tian Y D, Sukhbaatar S, Szlam A, Fergus R. Simple baseline for visual question answering. *arXiv: 1512.02167*, 2015. https://arxiv.org/abs/1512.02167, June 2017.

[30] Malinowski M, Fritz M. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proc. the 27th Int. Conf. Neural Information Processing Systems*, Dec. 2014, pp.1682-1690.

[31] Wu Z B, Palmer M. Verbs semantics and lexical selection. In *Proc. the 32nd Annual Meeting on Association for Computational Linguistics*, June 1994, pp.133-138.

[32] Miller G A. WordNet: A lexical database for English. *Communications of the ACM*, 1995, 38(11): 39-41.

[33] Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S E, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. *arXiv: 1409.4842*, 2014. https://arxiv.org/abs/1409.4842, June 2017.

[34] He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016, pp.770-778.

Ai-Wen Jiang received his Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2010. Currently, he is an associate professor at Jiangxi Normal University, Nanchang. His research interests include vision and language, deep learning, and crossmodal retrieval.

Bo Liu received his Ph.D. degree in computer science from University of Massachusetts Amherst, Massachusetts, in 2015. He is the recipient of the UAI-2015 Facebook Best Student Paper Award. Currently he is a tenure-track assistant professor in College of Computer Science and Software Engineering at Auburn University, Auburn. His primary research area covers machine learning, deep learning, stochastic optimization, and their numerous applications to BIGDATA.

Ming-Wen Wang received his Ph.D. degree in computer software and theory from Shanghai Jiao Tong University, Shanghai, in 2000. Currently, he is a full professor at Jiangxi Normal University, Nanchang. His research interests include natural language processing, information retrieval, and machine learning.