Visual Textbook Network: Watch Carefully before Answering Visual Questions

Difei Gao[1,2]
difei.gao@vipl.ict.ac.cn
Ruiping Wang[1,2,3]
wangruiping@ict.ac.cn
Shiguang Shan[1,2,3]
sgshan@ict.ac.cn
Xilin Chen[1,2,3]
xlchen@ict.ac.cn

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China

[2] University of Chinese Academy of Sciences, Beijing, 100049, China

[3] Cooperative Medianet Innovation Center, China

Abstract

Recent deep neural networks have achieved promising results on Visual Question Answering (VQA) tasks. However, many works have shown that a high accuracy does not always guarantee that the VQA system correctly understands the contents of images and questions, which are what we really care about. Attention based models can locate the regions related to answers, and may demonstrate a promising understanding of image and question. However, the key components of generating correct location, i.e. visual semantic alignments and semantic reasoning, are still obscure and invisible. To deal with this problem, we introduce a two-stage model Visual Textbook Network (VTN), which is made up by two modules to produce more reasonable answers. Specifically, in the first stage, a textbook module watches the image carefully by performing a novel task named sentence reconstruction, which encodes a word to a visual region feature, and then decodes the visual feature to the input word. This procedure forces VTN to learn visual semantic alignments without much concerning on question answering. This stage is just like studying from textbooks where people mainly concentrate on the knowledge in the book and pay little attention to the test. At the second stage, we propose a simple network as exam module, which utilizes both the visual features generated by the first module and the question to predict the answer. To validate the effectiveness of our method, we conduct evaluations on Visual7W dataset and show the quantitive and qualitative results on answering questions. We also perform the ablation studies to further confirm the effectiveness of the individual textbook and exam modules.

1 Introduction

Visual question answering (VQA) in general is to answer questions concerning a specific image which requires capabilities of visual semantic alignments and semantic reasoning. In the past few years, several VQA systems have demonstrated promising results on VQA task.

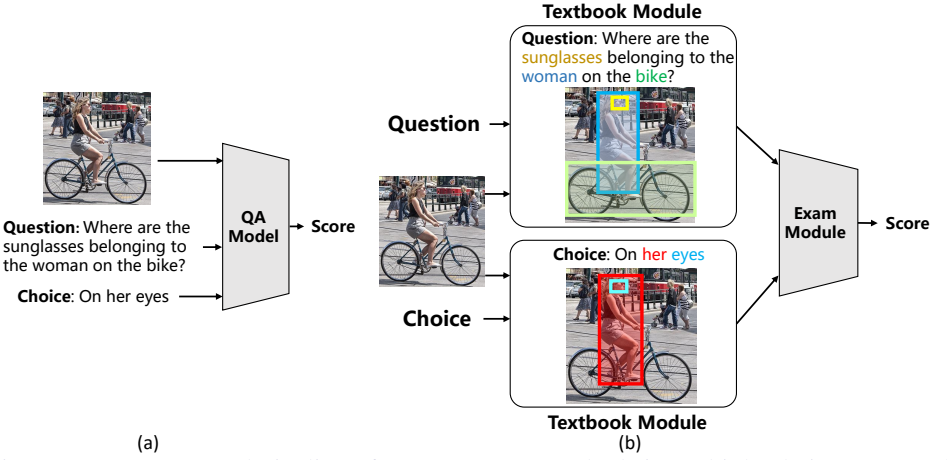


Figure 1: (a) An general pipeline of most common methods in multiple-choice VQA task. These methods start to answer the question at the beginning of the model. (b) Basic pipeline of Visual Textbook Network(VTN). To build a qualified VQA system, VTN first takes triplet <image, question, choice> as input to the textbook module, where we force module to align the content of the image and meaning of the text without concerning about VQA test. After real understanding of the image and the text, exam module utilizes image and text features from previous module to generate answer.

However, getting a high accuracy does not guarantee a VQA system correctly understands the content of an image and the meaning of a question which is what we really want. In addition, many recent works [1, 20, 27] point out the problem that VQA systems easily overuse the language prior to answer the question without truly understanding of images and questions. Some works [15, 24] solve this problem by designing powerful attention mechanisms to firstly focus on regions related to question, then answer it. While a reasonable final attention region may demonstrate a promising understanding of image and text, it is hard to prove that VQA system has correctly achieved both targets of visual semantic alignments and semantic reasoning on both image and text. For example, if a VQA system correctly attends to an image region that contains 'kid', it is possible for three reasons below: 1) the 'kid' is mentioned in question (visual semantic alignments); 2) the attribute, such as 'in a blue shirt', is depicted in question (semantic reasoning); 3) VQA system does not know this is a 'kid', but attention to this region just happen to get a high accuracy. We actually expect that a VQA system can locate an image region only for the first two reasons, rather than the last one. However, concentrating too much on final accuracy would probably cause that VQA system keeps away from truly understanding images and texts.

Inspired by the learning process of human, where people firstly acquire knowledge in book, then practice test skill, our method is designed to firstly focus more on obtaining knowledge from image-text pair, then practice visual answering skill. To this end, as shown in Figure 1, we propose a two-stage architecture Visual Textbook Network (VTN) which is made up by two modules: textbook module and exam module. Specifically, our VTN model first aligns visual information (image region) and texts (questions and choices) by learning a novel task, i.e. sentence reconstruction. In such task, the textbook module is designed to first locate the spatial position of each individual word in the image, and then reconstruct the word itself with the visual representation at the located position. This task forces VTN

to make efforts to learn the correspondence between image regions and texts, thus prevents the model from focusing only on the final answering accuracy. After the textbook module is trained, it is used to locate key information in the image, and extract local visual features accordingly, which are later used as input to the following exam module.

The visual feature extracted above mostly locates the image region that is related to one word. However, to correctly predict the answer, the VQA system needs to locate the region related to the whole question. To this end, the exam module is designed to perform semantic reasoning for question answering. To be specific, the exam module maps word-level image features to sentence-level features through several non-linear operations, and gives answers based on both image feature and question feature.

We evaluate our architecture on multiple-choice question dataset Visual7W which contains rich natural language annotations: questions and choices of multiple-choice question. Quantitative and qualitative results demonstrate that our method has promising capabilities of visual semantic alignments and semantic reasoning. Besides, the visualization of attention maps validates that our method can make the VQA process more transparent.

2 Related Work

Visual Question Answering. In last few years, a variety of works [3, 5, 9, 16, 24, 26] have demonstrated promising results on *free-form open-ended* and *multiple-choice* visual question answering task. We approximately classify these works into two branches. First branch of VQA works [15, 21, 26] concentrates on correctly locating the image region based on question. [21] adopts a multimodel framework in which question features and many candidate image regions features are mapped into a common latent space. Then an image-region selection mechanism will find the proper region. [26] proposes a multi-step attention network to locate the image regions that are relevant to the answer prediction. [15] presents a hierarchical co-attention model that co-attends to both the image and question. Second branch of VQA works [2, 24, 27] mainly focuses on how to build model to effectively realize semantic reasoning which requires model to behave differently for different type of questions. [17] proposes a deep convolutional neural network with a dynamic parameter layer whose parameters are determined dynamically based on a given question. Neural Module Networks [3] decomposes visual question answering model into module components which can dynamically assemble a specific model to simulate a specific reasoning process. [27] views semantic reasoning procedure as verifying if the question tuple which summarizes the content of a question is depicted in the scene or not. In this paper, we propose a two-stage network that can first learn to locate the image region, then learn to realize semantic reasoning.

Multi-stage model in Scene Understanding. The second line of works that are very relevant to VTN is multi-stage model in scene understanding. These methods utilize the objects information of an image to improve the performance on scene understanding, such as scene classification [13], image captioning [7]. Many existing models [7, 12] build a two-stage model to solve scene understanding task. In the first stage, these methods generate the objects or concepts that one image contains, then utilize these results to achieve their own tasks on the second stage. Recently, most VQA works [3, 15, 21, 24, 26] also follow a multi-stage model. Especially for visual question answering on abstract scenes task [23, 27], to avoid noise in recognizing objects, dataset provides the annotations of objects information in the image. [23] only focuses on the semantic reasoning on VQA by designing sophisticate answering model to reason on the scene graph of an image. However, annotations of objects

may lose some important image appearance information which can not be clearly expressed by words or concept labels. In our work, similar to abstract scenes VQA, we first parse the image content, then answer the question. But during two stages, we keep the objects appearance information.

Image Captioning. Sentence reconstruction task in VTN is similar to image captioning task, since they both generate words based on content of an image. So, in this paragraph, we introduce the works on image captioning task. Several pioneering approaches [4, 7, 10, 25] have made promising results on image captioning task. Many papers on image captioning have focused on generating every word based on specific image region. [25] proposes a visual attention model which is able to align words to spatial regions within an image. [11] adopts a model that can learn the correspondence between short phrases within sentences and image region. More recently, there are also works that utilize the results of visual question answering to achieve image captioning task or in opposite way. [14] proposes a model to choose a set of questions, then predicts answers based on image and caption respectively. The VQA results are used to rank images and captions. [8] proposes a baseline method that uses captioning of an image as auxiliary input to VQA model. In addition, similar to ours, [19] proposes a phrase reconstruction model that builds the correspondence between image region proposals and phrases to implement visual grounding. In our paper, instead of using region proposals, we build a sentence reconstruction module based on attention mechanism in image captioning to extract features related to a natural language sequence.

3 Visual Textbook Network

We now give an overview of the architecture of Visual Textbook Network(VTN). Our model is designed to solve multiple-choice visual question answering task which inputs <image, question, choice> triplets, then outputs the matching score of the triplets. The main architecture of our model is illustrated in Figure 2. Our goal is to force VQA system to concentrate more on parsing the image and make answering procedure more transparent. To achieve this, VTN is designed to have two stages: first, we propose the textbook module that performs sentence reconstruction task to pay attention to visual semantic alignments, then the exam module is implemented to predict the probability that one triplet <image, question, choice> is correct.

3.1 Textbook Module

The first stage of Visual Textbook Network is to perform visual semantic alignments. In other words, what we need is aligning every word in sentence to an image region. The textbook module achieves this capability by implementing a novel task, sentence reconstruction, which can be seen as a cross-modal autoencoder [6]. The word in sentence is encoded to a visual feature vector that depicts the context information (region information) of a given image. Then the visual feature is decoded to reconstruct input word. In this section, we will introduce the details how the textbook module achieves this task.

VTN splits input triplet <image,question,choice> into two parts, image-question pair and image-choice pair. Each pair is fed into one textbook module. As illustrated in Figure 2(a). From now on questions and choices are processed with exactly the same operation, so we use word 'sentence' to represent both question and choice. The textbook module takes a single image and a sentence as inputs. The words in sentence are encoded to a set of one-hot vectors

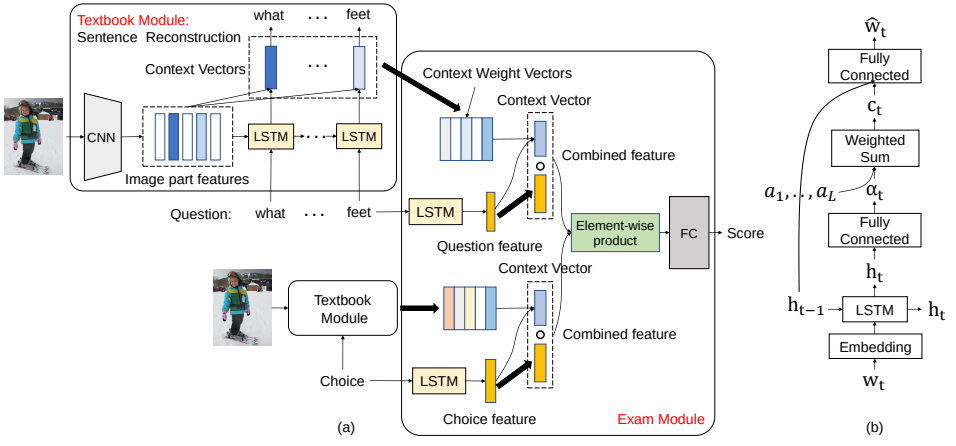


Figure 2: Architecture of Visual Textbook Network (VTN): (a) The framework of VTN. Input triplet <image, question, choice> is split into two parts, image-question pair and image-choice pair. Each pair is fed into textbook module to do sentence reconstruction for learning visual semantic alignments and extracting attention maps as visual features. Then, the exam module predicts two final attention maps for two image-sentence pairs and outputs the <image, question, choice> matching score. $\circ$ denotes vector concatenating. (b) The sentence reconstruction model at every time step. A one-hot word vector w_t is encoded to a visual feature c_t through several layers. Then a fully connected layer decodes the visual feature to a word $\hat{w}_t$.

$\mathbf{w}_1, ..., \mathbf{w}_T$, where T denotes the number of words in sentence. We use a convolutional neural network to extract a set of convolution layer feature vectors $\mathbf{a}_1, ..., \mathbf{a}_L$ as image part features, where L is the number of image parts. Then the module starts to reconstruct sentence vectors $\hat{\mathbf{w}}_1, .., \hat{\mathbf{w}}_T$ given image part vectors and word vectors.

A long short-term memory (LSTM) network is used to reconstruct the entire word sequence. At every time step of LSTM, the module reconstructs one word by given the previous hidden state $\mathbf{h}_{t-1}$ and the current input word $\mathbf{w}_t$. Figure 2(b) illustrates the reconstruction procedure at every time step of LSTM in detail. This can be formulated as

$$\hat{\mathbf{w}}_t = \arg\max_{\mathbf{w} \in \mathbf{V}} p(\mathbf{w} | \mathbf{w}_t, \mathbf{h}_{t-1}; \theta) \tag{1}$$

with parameters θ and vocabulary of words $\mathbf{V}$.

To reconstruct a word, similar to the auto-encoder framework, the word is encoded to a context vector $\mathbf{c}_t$ which is the representation of an image region. The context vector is calculated as a weighted arithmetic mean of the $\mathbf{a}_i$, $\mathbf{c}_t = \sum_{i=1}^{L} \alpha_{t,i} \mathbf{a}_i$, where $\alpha_{t,i}$ is the weight and $\sum_{i=1}^{L} \alpha_{t,i} = 1$. The context weight vector $\alpha_t = (\alpha_{t,1}, ..., \alpha_{t,L})$ is a linear projection of hidden activation of LSTM $\mathbf{h}_t$ followed by ReLU non-linearity and softmax layer. And we abbreviate the computation of hidden activation of LSTM at every time step as $\mathbf{h}_t = \text{LSTM}(\mathbf{w}_t, [\mathbf{h}_{t-1}, \mathbf{c}_{t-1}])$, where $[,]$ denotes concatenating two vectors, $\mathbf{w}_t$ is the input of LSTM and $[\mathbf{h}_{t-1}, \mathbf{c}_{t-1}]$ is the previous hidden state that inputs to LSTM.

Once obtaining the representation of an image region, the context vector $\mathbf{c}_t$ is decoded to a word $\hat{\mathbf{w}}_t$. To let decoder have the capability of putting emphasis on language model, decoder predicts word also based on the hidden state at last time step $\mathbf{h}_{t-1}$, since some

words, such as 'is', 'and', are difficult to reconstruct based on an image. The decoder is formulated as

$$\mathbf{h}_w = \tanh(\mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{W}_c \mathbf{c}_t) \tag{2}$$

$$\hat{\mathbf{w}}_t = \text{softmax}(\mathbf{h}_\text{w} \mathbf{W}_\text{w}) \tag{3}$$

where $\mathbf{W}_h, \mathbf{W}_c$ are the weight matrices, which project $\mathbf{h}_{t-1}$ and $\mathbf{c}_t$ to the same dimension and $\mathbf{W}_w$ is also the weight matrix.

Now we have defined the model to reconstruct a sentence based on an image. The cross entropy loss l_s for reconstructing the sentence is used to train the textbook module. To use the knowledge learned in the textbook module for VQA, we output the set of context weight vectors $\alpha_1, ..., \alpha_T$, which contain visual semantic alignments information, to the exam module. And since image-question pair and image-choice pair are fed into two textbook modules respectively, there are two sets of context weight vectors input to following module. In addition, though the modules processing image-question pair and image-choice pair are the same, the parameters are not shared.

3.2 Exam Module

After acquiring the correspondence between words in sentence and image regions, to correctly predict the answer, VTN has to locate the region related to the whole question and the choice. Therefore, the purpose of the exam module is to map word-level image features to sentence-level features and to predict the matching score of input triplet. Figure 2(a) shows the architecture of the exam module.

More concretely, the inputs to the exam model are two sets of context weight vectors, one input question and one choice. For simplicity, the operations below to process the question and the choice are exactly same, thus we take processing question for example. The exam module first takes a question and context weight vectors from the textbook module for image-question pair as inputs. Words in the question are encoded into one-hot vectors and passed through an embedding layer. Then an LSTM is used to extract the feature of the question $\mathbf{h_q}$ which is the hidden activation of the final time step.

A simple method is proposed to compute attention vector of the image γ_q that is related to entire sentence. γ_q is calculated as a weighted arithmetic mean of $\alpha_1, ..., \alpha_T$:

$$\beta = \text{softmax}(\text{ReLU}(\mathbf{W}_\text{b} \mathbf{h}_\text{q})) \tag{4}$$

$$\gamma_q = \sum_{t=1}^{T} \beta_t \alpha_t \tag{5}$$

where $\mathbf{W}_b$ is weight matrix and $\beta = (\beta_1, .., \beta_T)$ are the weights of arithmetic mean which are generated according to sentence feature $\mathbf{h}_q$. And the context vector of whole sentence is computed as $\mathbf{c}_q = \sum_{i=1}^{L} \gamma_{q,i} \mathbf{a}_i$. Then we concatenate two vectors $\mathbf{z}_q = [\mathbf{c}_q, \mathbf{h}_q]$ to jointly represent the feature of image and question.

Now we implement above operations on both image-question pair and image-choice pair, and get two joint representation vectors, $\mathbf{z}_q$ for image and question, $\mathbf{z}_c$ for image and choice. Finally, the exam module will generate the matching score of <image, question, choice> triplet by calculating the element-wise product of $\mathbf{z}_q$ and $\mathbf{z}_c$ followed by a fully connected layer. Similar to textbook module, the parameters above are not shared between question and choice. When training and testing VTN, we simultaneously calculate several scores

of <image, question, choice> triplets which share the same image-question and only one choice is labelled as correct answer. Then we calculate the softmax cross entropy loss of output scores to train the model.

3.3 Training

We train our model to minimize combination of three cross-entropy loss l_q, l_m, l_a, where l_q is the cross-entropy loss for reconstructing questions, l_m is the loss for reconstructing choices, l_a is the loss for output answers. For l_m, we only calculate the loss for reconstructions of right choices, others will be set to zero. The total loss l can be expressed as:

$$l = l_a + \lambda(l_q + l_m) \tag{6}$$

where λ is a hyper parameter that balances the relative weights of the textbook module and the exam module.

4 Experiment

4.1 Datasets & Experiment Setup

We evaluate our method on the **Visual7W** dataset [28]. A question in Visual7W is one of seven types, *who, what, when, where, why, how* and *which*. Visual7W dataset contains two parts, telling task which involves 6W questions, and pointing task which only involves which questions. In this paper, we evaluate the models on telling task which contains 69,817 training QA pairs, 28,020 validation QA pairs, 42,031 test QA pairs and totally 47,300 MSCOCO images. Each QA pair contains a question, an image and four natural language multiple choices of which only one choice is correct. We choose to evaluate models on Visual7W dataset for the reason that Visual7W's multiple choices contain rich natural language information that can effectively train the textbook module.

Our model is developed by Tensorflow. We use VGG-19 [22] to extract the feature of an image. The $14 \times 14 \times 512$ feature map of conv5 layer before max pooling is used to represent $\mathbf{a}_1, ..., \mathbf{a}_L$, where $L = 14 \times 14$. We use Glove [18] as our word embedding layer, and the words are encoded into a 300-dimensional feature vector. We do not finetune both CNN and word embedding layer. We have empirically tested the influence of hyper parameter λ by setting its value to the range $[0.1, 10]$, and found that our model yields relatively stable accuracy. Here we report the result by setting $\lambda = 1$ in loss function.

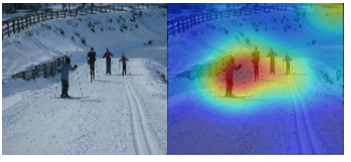
The network is trained using RMSprop optimizer with learning rate of $5e^{-5}$, momentum 0.3 and weight decay $5e^{-5}$. We apply dropout after the LSTM layers and fully connected layers in the textbook module with dropout rate 0.5. We train for up to 50 epochs with batch size 100 and use early stopping if the validation accuracy does not improve in the last 5 epochs.

4.2 Ablation study

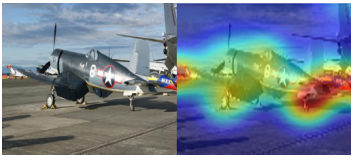
In order to validate the effectiveness of Visual Textbook Network, we are going to answer following questions: **1)** Is the reconstruction of sentence effective in visual semantic alignments? **2)** Is firstly parsing the image content and the text helpful in VQA? **3)** Whether exam module can attend to the image region related to the whole sentence?

hats

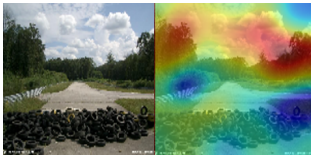
lights



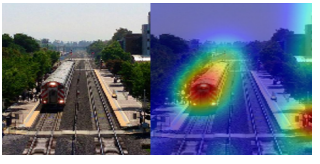
What are the **people** doing?



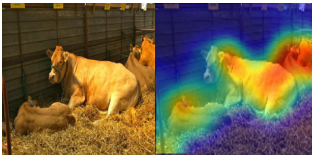
How is the **plane**?



What shines brightly in the sky?

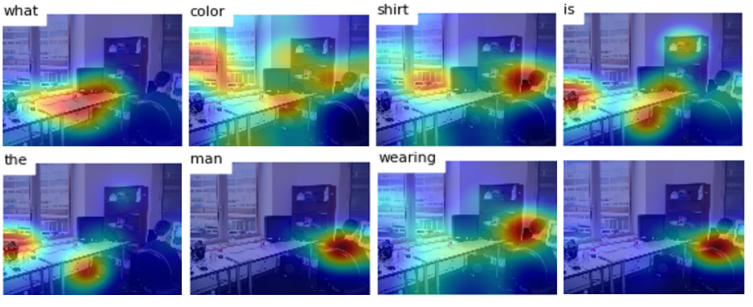


What is the color of train?



Who is in the picture?

Figure 3: Top: attention maps to reconstruct one word in the textbook module for multiple choices on Visual7W. Middle: attention maps to reconstruct one word in the textbook module for questions. Bottom: attention maps for whole questions which are also used to generate answer.



Question: What color shirt is the man wearing?

Choices: black, silver, white, gold **VTN**: **black** **GT**: black

Figure 4: Attention maps for one sample. Top: The first seven attention maps are generated when reconstructing question. The last one attention map is the result of semantic reasoning on whole question. Bottom: Attention maps when generating four choices.

Method	What	Who	Where	When	Why	How	Overall
LSTM-Atten	50.7	63.6	64.1	75.7	57.0	49.7	55.6
Q-Textbook	53.2	64.7	66.3	76.9	58.4	50.5	57.5
M-Textbook	55.2	66.1	67.2	76.9	59.5	52.6	59.2
Mean-Exam	56.5	66.8	68.0	77.2	59.5	53.3	60.1
Full-VTN	58.6	68.3	69.7	78.2	59.3	54.2	61.7

Table 1: Test accuracy(%) of different models for ablation study. Models are trained on the Visual7W train split and tested on test split.

Method	What	Who	Where	When	Why	How	Overall
FRITZ ET AL.[16]	48.9	58.1	54.4	71.3	51.3	50.3	52.1
ZHU ET AL.[28]	51.5	59.5	57.0	75.0	55.5	49.8	55.6
LSTM(Q,A,I')-ATTENTION[20]	59.7	59.7	61.0	74.2	52.8	52.2	54.8
LSTM(Q,A,I)-CONTEXT[20]	61.0	69.5	71.4	81.5	62.7	56.2	63.9
MCB[8]	60.3	69.2	70.4	79.5	58.2	51.1	62.2
VTN	58.6	68.3	69.7	78.2	59.3	54.2	61.7

Table 2: Test accuracy(%) on multiple-choice Visual7W dataset.

In this section, we ablate certain component of our model, or replace it with other alternative component in previous work. These models are as follows:

LSTM-Atten: Instead of using the textbook module to locate image region, we use LSTM model with spatial attention which is similar to work [28].

Q-Textbook: We only apply the textbook module on question image pair. We use LSTM model to encode multiple choices and there is no attention on multiple choices.

M-Textbook: We only apply the textbook module on multiple choices image pairs. We use LSTM model to encode question and there is no attention on question.

Mean-Exam: The textbook module will generate a set of attention regions. Instead of using the exam module to reason on these attention regions, *Mean-Exam model* calculates the mean of attention region features.

Full-VTN: The full Visual Textbook Network introduced in this paper.

Table 1 shows the results of these ablations and full model on Visual7W test set. A comparison of our model with other models is shown in Table 2. The results of comparison methods come from the original literatures.

Effectiveness of the textbook module: From the comparison in Table 1, it is clear that the models contain textbook module (last 4 rows) outperform the model without such module (LSTM-Atten), confirming the effectiveness of the textbook module. Moreover, the top and middle lines in Figure 3 show that textbook module can correctly focus on the image region for a given word. The above results indicate that the sentence reconstruction task can learn the alignments between image regions and words.

Effectiveness of the exam module: The results of *Mean-Exam* and *Full-VTN* show that implementing more sophisticated semantic reasoning model will gain more improvement. Figure 3 and Figure 4 show the attention maps that exam model reasons on a set of attention regions based on a sentence. Qualitative results show that exam model can locate the most likely region related to the whole sentence.

Effectiveness of attention on candidate choices and questions: The performance of *Q-*

Textbook and *M-Textbook* in Table 1 may show that VTN utilizes the information of multiple choices more efficiently. The declarative sentences in multiple choices are more related to an image, while utilizing the questions information needs more semantic reasoning.

Table 2 shows that the accuracy of our model is comparable with the state of the arts. Besides, compared to the state of the arts, our method is easier to concentrate on parsing the image and makes the VQA process more transparent, as shown in Figure 4.

5 Conclusions

To solve the problem that most VQA systems generate answer without truly understanding image, we propose a two-stage model Visual Textbook Network(VTN) that first concentrates on performing visual semantic alignments and then pays attention to answering. To implement visual semantic alignments, we define a novel task named sentence reconstruction to learn it. Moving forward, we are going to design more sophisticated textbook module and exam module to improving the accuracy.

Acknowledgements

This work is partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61379083, 61650202, and Youth Innovation Promotion Association CAS No. 2015085.

References

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *arXiv preprint arXiv:1606.07356*, 2016.

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. In *Proceedings of NAACL-HLT*, pages 1545–1554, 2016.

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[4] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2016.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[6] Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[7] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482, 2015.

[8] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*, pages 457–468, 2016.

[9] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *European Conference on Computer Vision*, pages 727–739. Springer, 2016.

[10] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[11] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[12] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[13] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances In Neural Information Processing Systems*, pages 1378–1386, 2010.

[14] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pages 261–277. Springer, 2016.

[15] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[16] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9, 2015.

[17] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 30–38, 2016.

[18] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

[19] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[20] Abhijit Sharang and Eric Lau. Recurrent and contextual models for visual question answering. In *arXiv preprint arXiv:1703.08120*, 2017.

[21] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4613–4621, 2016.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[23] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. In *arXiv preprint arXiv:1609.05600*, 2016.

[24] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2397–2406, 2016.

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

[26] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[27] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022, 2016.

[28] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.