# Attention Based Model in Visual Question Answering

**Kan Chen**[*] kanchen@usc.edu

**Jiang Wang** wangjiang03@baidu.com

**Wei Xu** wei.xu@baidu.com

## 1 Approach

Visual question answering task (VQA) automatically generates an answer for a given image and an image-related question [1]. Attention is of significant importance in VQA because different questions inquire about different image regions. We propose an attention model for VQA that explicitly exploits the questions to guide the attention to generate appropriate answers.

We propose a configurable convolutional neural network to learn question-guided attention. The framework is illustrated in Figure 1. A long-short term memory (LSTM) model [3] is applied to extract the semantic information from the given questions as question embeddings. The question embeddings determine the convolutional kernels $\mathbf{k}$, which are utilized to generate question-guided attention maps. For example, if the question is "what is the color of the car", the model should focus its attention on the regions of the cars. Thus, the convolutional kernel should correspond to the car features. An input image is represented as an $N \times N$ feature map, where each feature vector in the feature map is generated by the *Vgg-19* network [2]. The convolutional kernel $\mathbf{k}$ is applied to the image feature map $\mathbf{I}$ to generate the question-guided attention map $\mathbf{m}$. After filtering out the noise of the image feature map $\mathbf{I}$ by multiplicatively applying the attention map $\mathbf{m}$, we can generate a question-guided image feature $\mathbf{I}'$, which focuses on the locations related to the input question. Finally, we generate the answer by jointly projecting the reduced image feature, the original image feature map $\mathbf{I}$ and the question embeddings, and learning an answer classifier based on the projected features.

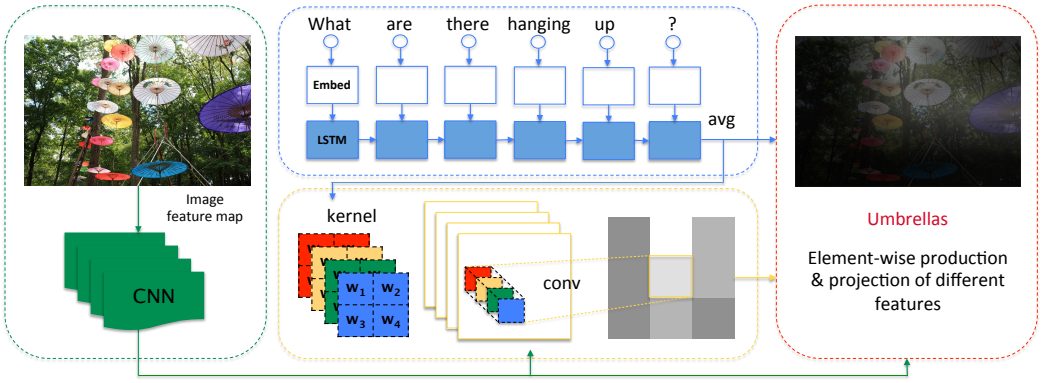


Figure 1: The framework of the attention based VQA model: The green box denotes image feature map extraction; the blue box is query processing part; the yellow box illustrates question-guided attention map learning; the red box is the answer classification procedure.

[*]The work was done while the author was an intern at Baidu Research.

| Model | COCO-QA ACC. | COCO-QA WUPS 0.9 | COCO-QA WUPS 0.0 | DAQUAR-QA ACC. | DAQUAR-QA WUPS 0.9 | DAQUAR-QA WUPS 0.0 |
|---|---|---|---|---|---|---|
| LSTM [5] | 0.3676 | 0.4758 | 0.8234 | 0.3273 | 0.4350 | 0.8162 |
| IMG [5] | 0.4302 | 0.5864 | 0.8585 | - | - | - |
| IMG+BOW [5] | 0.5592 | 0.6678 | 0.8899 | 0.3417 | 0.4499 | 0.8148 |
| VIS+LSTM [5] | 0.5331 | 0.6391 | 0.8825 | 0.3441 | 0.4605 | 0.8223 |
| 2-VIS+BLSTM [5] | 0.5509 | 0.6534 | 0.8864 | 0.3578 | 0.4683 | 0.8215 |
| FULL [5] | 0.5784 | 0.6790 | 0.8952 | 0.3694 | **0.4815** | 0.8268 |
| ATTENTION | 0.5548 | 0.6568 | 0.8890 | **0.4276** | 0.4762 | **0.8304** |
| ATT+HSV | **0.5803** | **0.6814** | **0.8966** | - | - | - |
| HUMAN | - | - | - | 0.5020 | 0.5082 | 0.6727 |

Table 1: Results on Toronto COCO-QA [5] and DAQUAR [4] datasets; "ATTENTION" is our attention model and "ATT+HSV" is our attention model with color features.

## 2 Experiment Results

We explore several attention models and evaluate them on two VQA datasets: Toronto COCO-QA [5] and DAQUAR reduced dataset [4]. Toronto COCO-QA dataset consists of QA-pairs with single-word answers. It has 78736 training and 38948 validation QA pairs. Similarly, we work on DAQUAR reduced dataset, which only contains QA-pairs with single-word answers. The dataset has 3825 and 286 training and test QA-pairs, respectively.

We employ accuracy and WUPS score at threshold of 0.9 and 0.0 [6] as evaluation metrics. Table 1 shows that our model achieves significant accuracy improvements on both datasets compared to state-of-the-art methods in [5]. In Table 2, we demonstrate that our attention based model outperforms methods in [5] in terms of accuracy in three out of four sub-categories. We further visualize some selected images, their corresponding attention maps and the generated answers in Figure 2, which shows that our generated attentions can focus on the regions that are relevant to the VQA task.

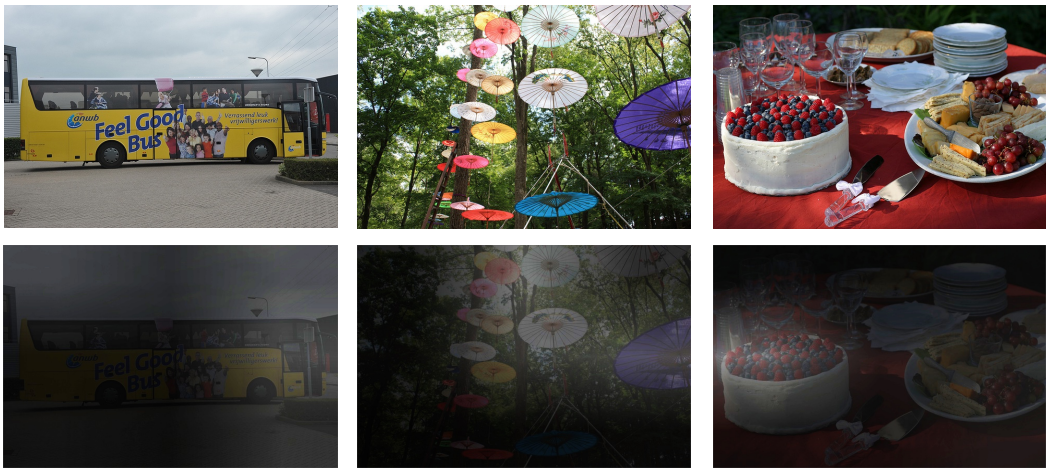


Figure 2: Selected attention maps after element-wise production and answers generated by attention based model (Q = Question, GT = Ground Truth, A = Answer)

| Model | Object | Number | Color | Location |
|---|---|---|---|---|
| IMG+BOW [5] | 0.5866 | 0.4410 | **0.5196** | 0.4939 |
| VIS+LSTM [5] | 0.5653 | 0.4610 | 0.4587 | 0.4552 |
| FULL [5] | 0.6108 | 0.4766 | 0.5148 | 0.5028 |
| ATTENTION | 0.5977 | 0.4693 | 0.4359 | 0.4911 |
| ATT.+HSV | **0.6217** | **0.4799** | 0.4727 | **0.5194** |

Table 2: Per category accuracy on Toronto-QA dataset [5] ; "ATTENTION" is our attention based model and "ATT+HSV" is our attention model with color features.

## References


[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. *arXiv preprint arXiv:1505.00468*, 2015.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

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

[4] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690, 2014.

[5] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *arXiv:1505.02074*. 2015.

[6] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.