Combining Visual and Contextual Information for Question Answering

Kexin Yi
Harvard University
kyi@g.harvard.edu

Aditya Thomas
MIT
adityat@mit.edu

Abstract

In this project we generalize the task of visual question answering (VQA) by combining both image and contextual inputs of a visual scene as the information source for question answering. We introduce a new framework of this generalized task and assess its performance on the COCO-QA dataset. Our result shows a large increase in performance on the addition of contextual information, i.e. image captions, to the question answering pipeline. Under our new framework, the stacked attention model is able to achieve a test accuracy of up to 80% on the COCO-QA dataset. We demonstrate through our experiments that combining multiple information sources can lead to better understanding of a scene, and hence more accurate reasoning. This document serves as the final report for the MIT course 6.869 during Fall 2017.

1. Introduction

Computer vision and natural language processing are two fundamental tasks of artificial intelligence. Recent advances in deep neural networks have greatly benefited both fields and boosted the performance in these tasks by a large margin [11, 19, 5]. What lies at the core of a deep neural network is its power to learn a latent representation for any multi-modal input data, which empowers effective information extraction from both language and image inputs. Recent work has demonstrated that the latent information from both inputs can be combined to perform high-level interdisciplinary tasks such as image description generation and question answering [2, 8].

In natural language processing, and more generally artificial intelligence, an important task is to build an automated system that can understand human language and generate responses based on contextual information. Recently, along with the introduction of the Stanford Question Answering dataset (SQuAD) [15], which defines the benchmark task for context based question answering, a lot of progress has been made in this area such as open domain question answering and large scale reading comprehension [3].

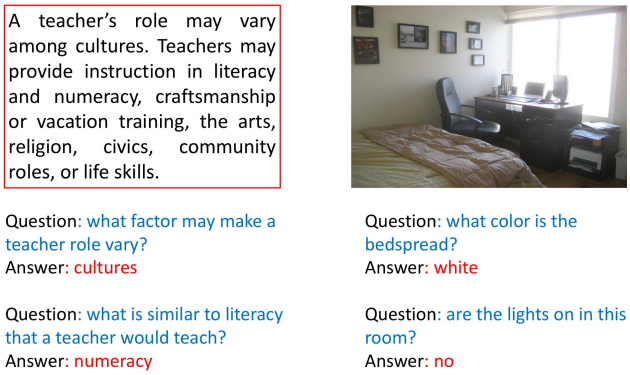

Figure 1. Examples of contextual question answering (left) and visual question answering (right). The examples are selected from the SQuAD and the VQA dataset.

Another important problem in computer vision is visual question answering (VQA), which is closely related to understanding fundamental machine perception and has huge potential applications. Unlike in contextual question answering where the only input and output are natural languages, the visual question answering task bases the questions on the content of natural images. Examples for both baseline tasks can be found in figure 1. As a benchmark, the Visual Question Answering (VQA) dataset [2] has been introduced for the task. Other similar datasets for VQA include the DAQUAR [14] and the COCO-QA [16] dataset.

Visual and contextual information are both key components of human perception. As demonstrated in figure 1, question answering can be performed based on both input sources separately. However from a high level, a human can easily perceive visual information aligned with its contextual descriptions to form a comprehensive understanding of the scene she sees, and is able to perform reasoning based on the full picture. Motivated by this fact, we explore a generalization of question answering by incorporating both visual and contextual information. An example of this generalized task is shown in figure 2, in which the image is associated with both a question and a set of independent contextual descriptions. If we only look at the image shown in

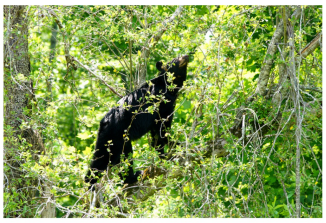

Figure 2. Example of question answering task that combines visual and contextual input. Example taken from the COCO-QA dataset.

the example, we can barely see a dark animal hiding behind the bush, while the text captions provide important complementary information that the animal is actually a bear. As a result, an agent would need information from both the image and the captions in order to answer all questions correctly. Our major goal in this project is to explore models that can leverage both information sources. The rest of this report is arranged as follows: in section 2 we review some related work on similar topics. In section 3 we introduce the models we study. We put emphasis on the stacked attention model [20], which is currently the prominent model for visual question answering, and compare it with other baselines. In section 4 we present and analyze our experiment results. Finally we conclude this project and discuss future research directions in section 5.

2. Related Work

Since the introduction of the tasks, a lot of progress has been made on both contextual and visual question answering. Yang et al. proposed the stacked attention model [20], which uses the question to attend the input image through a multi-layer attention network. The current best performing model, introduced by Anderson et al. further incorporates this top-down attention mechanism with bottom-up object detectors [1]. The current state of the art for contextual question answering is set by the combination of LSTM encoder/decoders with a self-matching attention model [18].

Relating to our generalized task, the baseline for image captioning is set by [9] where the model combines regional CNN and bidirectional LSTM decoder to generate descriptions for images. More recently, a similar task has been introduced that focuses on combined vision-language comprehension, whose evaluation is mainly based on identifying decoy captions of images [4]. [13] leverages the power of visual question answering as a feature extraction module for image-caption ranking.

3. Approach

3.1. Baseline Models

LSTM: For sequential data, such as language and time series, the learning algorithm should have the capability of keeping information from earlier steps. This is also the way humans learn, we base our understanding of events on past events and experiences. This capability of persisting with earlier information is possessed by neural networks that have loops in them known as the recurrent neural networks (RNNs). Long short-term memory (LSTM) [7] is a special kind of RNN that can learn long-term dependencies over sequential inputs. Just like a traditional RNN, LSTM has a chain structure formed by repetitively joining a basic unit, whose structure is shown in figure 3. The most important idea in the architecture of a LSTM lies in the cell state, which preserves memory of all previous state down the chain of repeating modules. The LSTM cell also incorporates a forgetting mechanism that enables removal of information from the cell state, through regulated structures know as gates. Specifically, the "forget gate" f_t determines how much of the cell state C_{t-1} is to be forgotten at the current step

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{1}$$

where σ is the sigmoid function. The "input gate" i_t adds the input information at step t, x_t to the cell state.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{2}$$
$$C'_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{3}$$

The value of the cell state C_t at the next step is given by

$$C_t = f_t \odot C_{t-1} + i_t \odot C'_t. \tag{4}$$

Finally, the "output gate" determines the output of the unit cell at current step

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{5}$$
$$h_t = \tanh(C_t) \odot o_t. \tag{6}$$

For the question answering task, each word from the questions and answers is encoded by an embedding through a fully-connected neural net layer followed by a $\tanh$ nonlinearity. The hidden state at the final step is treated as the output of the LSTM encoder. It is then passed to a multilayer perceptron (MLP) that outputs a probability distribution over all possible answers in the vocabulary, and thus predicts the answer to the question. We note that this model uses no image information so it can only model question-conditional bias without further contextual information.

CNN+LSTM: The CNN+LSTM model sets the baseline for visual question answering, which is a two channel model with both the image and the question as inputs. The image

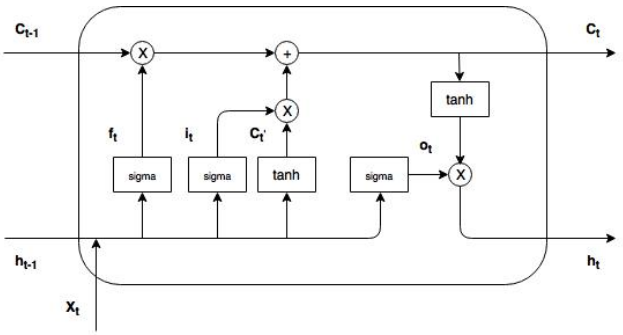

Figure 3. Architecture of a LSTM cell.

feature is extracted by a CNN, while the questions are still encoded by a LSTM which takes its final hidden layer as output similar to the previous model. The image features and the hidden vector for the question are concatenated to form one single feature vector and passed to an MLP that predicts a probability distribution over the answers.

3.2. Stacked Attention Model

The main model that we focus on studying for our generalized question answering task is the stacked attention (SA) model [20]. Similar to the CNN+LSTM baseline, the stacked attention model takes inputs from both the image and question, and encode them using a CNN and a LSTM respectively. The difference between the models lies in the way in which the hidden vector of the question and the image features are combined. Instead of directly concatenating them together, the SA model combines the features through a multi-layer structure known as the attention layers.

In most cases, the input question is related to only a specific object or local region of the image. For example, in figure 4.2, there are multiple objects such as the shed, trees, bushes, grass besides the bench, while the answer to the question only relates to the region of the image around the shed. So using the global image feature vector might not be optimal for question answering. The layers of the stacked attention network progressively filter out regions of the image and gradually focus on the specific elements of the feature map according to the input question as well as other contextual signal.

The way the attention mechanism works is described as follows. Given an image feature map f_I output from a fully convolutional network (FCN), and an encoded question vector f_Q output from the LSTM, the model first feeds them to a single-layer fully connected network

$$h_A = \tanh(W_{I,A} f_I \oplus (W_{Q,A} f_Q + b_A)) \quad (7)$$

$$p_I = \text{softmax}(W_P h_A + b_P) \quad (8)$$

and outputs a probability distribution p_I over different spatial locations on the feature map, where $W_{I,A}, W_{Q,A}, b_A, W_P and b_P$ are all weights to be trained. This output distribution further allows us to compute a weighted average of the feature map elements.

$$\tilde{f}_I = \sum_i p_i v_i \quad (9)$$

$$u = \tilde{f}_I + f_Q \quad (10)$$

where index i scans over all spacial locations on the feature map.

What a single attention layer does can be understood as generating some weights according to the input language vector f_Q. This process can be interpreted as re-weighting different local regions on the input image. Compared to the CNN+LSTM baseline which simply combines the question vector and the image features, the attention layer constructs a more informative input from the language source. Higher weights are assigned to elements of the feature map that corresponds to regions that are more relevant to the question.

The attention mechanism can be further generalized to multiple layers by treating the output u from the previous attention model as the language input for the new layer. Furthermore, multiple layers can be stacked together to create a deep attention mapping. The general expressions for all following attention layers can be written as

$$h_A^k = \tanh(W_{I,A}^k f_I \oplus (W_{Q,A}^k u^{k-1} + b_A^k)) \quad (11)$$

$$p_I^k = \text{softmax}(W_P^k h_A^k + b_P^k) \quad (12)$$

$$u^k = \sum_i p_i^k v_i + u^{k-1}. \quad (13)$$

Each layer outputs an attention mapping p^k that tells the model where on the image to focus at. The final output u^K is then fed to an MLP followed by a softmax function to output the probability distribution over the answers. A sketch of the entire architecture can be found in figure 4.

One quick way to add extra contextual information to the question answering task is to directly feed it to the LSTM encoder following the input question. This simple generalization applies to all three models that we study. We adopt this new framework in our experiments and treat the image captions as extra contextual information.

4. Experiments

In this section, we implement and evaluate various baseline models for question answering and present experiment results of these models on the COCO-QA dataset under our generalized framework. We focus on the stacked attention model and compare it to the baseline LSTM and CNN+LSTM models. The experiments are conducted under 3 different regimes which correspond to different levels of contextual signal input, from no contextual input to strong contextual signal. Details of the model can be found

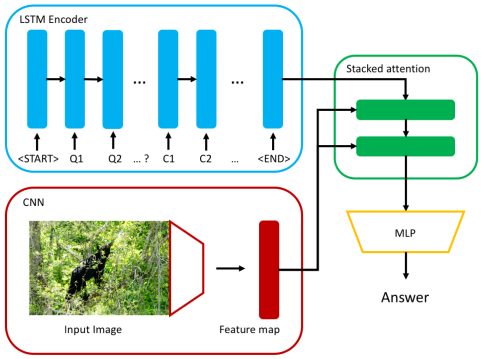

Figure 4. Architecture of the stacked attention model.

in our code at `https://github.com/kexinyi/6.869_final_project`.

4.1. Implementation Details

In all the following experiments, we use Adam [10] with a learning rate equal to 5×10^{-4} as the optimizer, and use the cross entropy as the loss function. The number of attention layers is set to be 2, each layer containing 512 hidden units.

Feature extraction with CNN: In both the CNN+LSTM baseline and the CNN+LSTM+SA model, we use ResNet-101 [6] pretrained on the ImageNet dataset [17] for feature extraction from the image. The last fully-connected layer and pooling layer is removed so the output feature map preserves spatial locality. All input images are color-normalized and rescaled to 224 by 224 before entering the CNN.

LSTM encoder: We use the same double-layer LSTM for all our models. The dimension of the word embedding is 300 and the hidden state has 256 dimensions. We use the same network to process both questions and image captions while separating the two types of contextual inputs by a question mark. In case there are multiple captions, the captions are separated by a period mark. A full input sequence begins with a "<START>" token and ends with an "<END>".

Data preparation: The COCO-QA dataset [16] includes 123287 images, 78736 training questions and 38948 test questions. The answers are all in single words, which allows us to treat the problem as a word-level classification problem. For all experiments we use the full training set for training and choose the first 10000 questions from the test set for validation. All test results are computed from the remaining 28948 test questions. The questions are of 4 types: object, number, color and location. The maximum question length is 55, and average is 9.65. Across the entire test set, there is a 23.3 % overlap in the training questions and a 18.70 % overlap in the training question-answer pairs.All images, including those from the test set, has 5 independent captions provided by the original COCO dataset [12]. To control the level of contextual signal, we select different number of captions as input: 0 captions represents no contextual information; 1 and 5 captions corresponds to normal and strong contextual signal.

4.2. Visualizing Attention Layers

As explained in the previous section, the stacked attention layers recursively re-weight the feature map output from the CNN, and thus guides the model to focus on spatial regions relevant to the input question and captions. The attention mapping can be visualized by rescaling the attention weights to the size of the original image. These mappings include important information of the model's functionality and effectiveness.

The attention map from a baseline stacked attention model without contextual input is shown in figure 4.2. The first attention layer takes the input from the LSTM encoder, which stores information of the question and generates a weight proposal on the feature map. In the example shown, the question asks "What are on the plate near the large stuffed peppers", and the first attention layer is able to find the pepper and focus on the local region. When we go one step further to the second layer, the model expands the horizon and searches for objects that are nearby. The correct answer is then generated from the re-weighted features. We can see here that not only can the attention layers guide the model to the correct local region, but also they are able to perform some sort of high level logical reasoning through the depth.

We also study the effect of contextual input to the attention mechanism. We observe that a moderate contextual signal can help the model attend better to answer the question correctly. In figure 4.2, we compare the baseline SA model without caption with the same model with 1 caption as contextual input. It is observed that without captions, the model only attends to the bottom right bench whereas with captions, the model pays attention to multiple objects including the shed, bench and tree. In this particular case, it is the more widely-spread attention mapping that leads to the correct answer to the question.

4.3. Test Accuracy on COCO-QA

We evaluate the test accuracy on the COCO-QA dataset for the three models, the CNN+LSTM+SA model and the baseline LSTM and CNN+LSTM models with 0, 1 and 5 image captions as contextual inputs. The result is shown in table 4.3. All models are trained for 30000 iterations with a batch size 64 (24 epochs) and the best performing model on the training set is selected for evaluation on the test set. The numbers represent the top-1 accuracy of the predicted answers.

The result shows a strong increase in test accuracy with

Figure 5. Attention mapping for baseline stacked attention model. The question associate with the images is "What are on the plate near the large stuffed peppers?" and the answer is "vegetables".

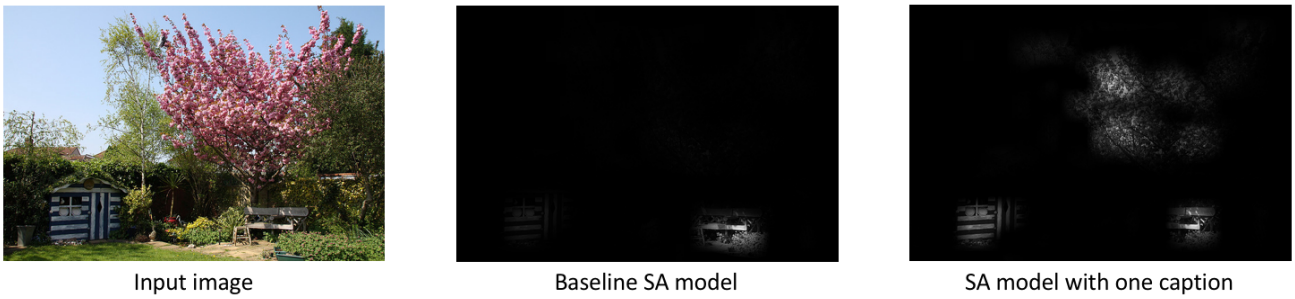

Figure 6. Attention mapping with and without contextual input. The question associate with the images is "What is close to the log cabin looking shed" and the answer is "tree" as predicted by the model with one caption, instead of "bench" as predicted by the baseline. Both attention mappings are the outputs from the second attention layer.

contextual input for all models, which justifies our original intuition. There is also a clear difference in the models' response to the contextual signal. Among all three models, LSTM benefits most from the contextual input but its performance also heavily relies on the captions since they are the only information source other than the question itself. The performance of the CNN+LSTM baseline is also improved by the captions, but it lacks the ability to make use of a stronger input to achieve a higher performance. The CNN+LSTM+SA model outperforms the other two baselines under most conditions. And unlike CNN+LSTM, the model has the flexibility to overfit to a strong contextual signal just like the LSTM model, and is able to achieve state of art question answering performance (0.82 in test accuracy).

One strange behavior that we observed in the results is the unreasonably high performance in the baseline LSTM model, which even beats the stacked attention model under strong contextual input. To our understanding, this behavior might be caused by the fact that the dataset on which we evaluate the models has a strong bias, that the questions are generated based on the image captions. As a result, there is an extremely high correlation between the content of the captions and the input questions, so that it would be more preferable for the model to overfit the captions with a simple language model. On the other hand, a strong contextual input (5 captions) is able to form a complete representation of an image, therefore it is still justifiable to base a heavier part of question answering on the captions when these information is available. After all, our most interesting finding lies in the weak context regime (1 caption), where the model is not overfit by the captions and the contextual input contributes to the attention mechanism in a interpretable way.

Model	0 caption	1 caption	5 captions
LSTM	0.3657	0.6300	**0.8632**
CNN+LSTM	0.5167	0.5942	0.5879
CNN+LSTM+SA	**0.5869**	**0.6462**	**0.8240**

Table 1. Test accuracy on COCO-QA dataset

5. Conclusions and Discussions

In this project, we introduce a generalized question answering task that combines visual information, i.e. an image, along with contextual information, i.e. a text description to assist in the task. We justify the motivation of this task and post a simple framework of joining the question and image captions together. The performance and generalizability of various models for the traditional visual ques-

tion answering task is studied under this new framework.

Overall we found that contextual information can improve the question answering performance for all models, but that the performance is also very sensitive to the input signal. Specifically, our best performing model suffers from overfitting to the contextual information. In order to deal with this issue and make our framework more generalizable, an important future direction is to define a benchmark for this generalized task, where image and contextual information are complementary (or at least not inclusive) for question answering. A good starting point would be the VQA dataset [2], for example by adding a separate LSTM encoder for the captions, and using other annotations such as instance segmentation as additional sources for contextual information for question answering.

Acknowledgements

We thank Chollette Olisah, who is a listener of 6.869, for helpful discussions.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *arXiv preprint*.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[3] D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.

[4] N. Ding, S. Goodman, F. Sha, and R. Soricut. Understanding image and text simultaneously: a dual vision-language machine comprehension task. *arXiv preprint arXiv:1612.07833*, 2016.

[5] J. Gehring, M. Auli, D. Grangier, and Y. N. Dauphin. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*, 2016.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[8] A. Karpathy and L. 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.

[9] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[14] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc., 2014.

[15] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[16] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[18] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[19] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[20] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.