Exploring Models for Visual Question Answering

Yifan Xu YIX081@UCSD.EDU
Department of Computer Science
Department of Cognitive Science
University of California at San Diego (UCSD), La Jolla, San Diego, 92093

ABSTRACT

This work explores image-based question-answering (QA) with different neural network models and reproduces the experimental conclusions presented by previous researchers in this area. Without object detection and image segmentation, the model can generate accurate answers about the images content by given any simple open-ended questions. In order to answer the question based on the image, a very high level of understanding of the relation between those two is necessary. Thus, two kind of neural networks, convolutional neural network (CNN) [7] and recurrent neural network(RNN) [8], are used together to address this problem.

Keywords— VQA, CNN, RNN, LSTM, GRU, VISUAL TURING TEST

I. INTRODUCTION

Passing Turing Test is the ultimate dream of artificial intelligence. In the past decade, many AI areas has achieved striking development. Among all of those areas, Computer Vision(CV), Natural Language Processing(NLP), and Knowledge Representation & Reasoning (KR) are the three major individual research fields. However, most of the real world problems are hard to solve by only using one single sub-domain knowledge. Even to understand a simple question, it commonly requires object recognition ability, language processing, and knowledge base reasoning. Many works done in area such as image captioning have shown that we can achieve descent result in many multi-discipline AI problems.

In last year, Aishwarya Agrawal and his colleague propose a new task of free-form and open-ended Visual Question Answering (VQA) [1]. Given an image and a question based on the image content, the task is to generate a short answer. In their work, the best performing model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [3] to encode the images. In the same year, Mengye Ren explores new models and datasets addressing the problem of VQA. His model is 1.8 times better than Aishwarya Agrawal and his colleague's work [2]. In this paper, I try to reproduces the experimental conclusions presented by him.

However, Visual question and answering is still not fully studied. Even through there are many strong evidences suggest that long short term memory (LSTM) is much better than regular RNN, RNN may still achieve a good result since the question is normally very short.
Gated Recurrent Units (GRUs), a simpler variant of LSTMs, were first proposed in 2014 [4]. It mainly serves the same purpose with LSTMs [5]. Therefore, an empirical evaluation of VQA with different neural network models mentioned above is studied in this paper.

II. METHODOLOGY

II.1 Mathematical formulation

Recurrent Neural Network:
The output of RNN depends not only on the current input, but also on the previous computations. It has a "memory" which captures information about what has been calculated so far. The mathematic formula of RNN is as follows:

$$x(t) = w(t) + s(t-1) \tag{1}$$

$$s_j(t) = f\left(\sum_i x_i(t) u_{ji}\right) \tag{2}$$

$$y_k(t) = g\left(\sum_j s_j(t) v_{kj}\right) \tag{3}$$

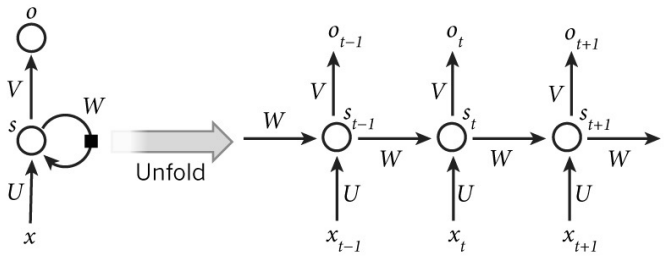
where $f(z)$ is sigmoid activation function:

$$f(z) = \frac{1}{1+e^{-z}} \tag{4}$$

and $g(z)$ is softmax function:

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \tag{5}$$

Here is what a typical RNN looks like:

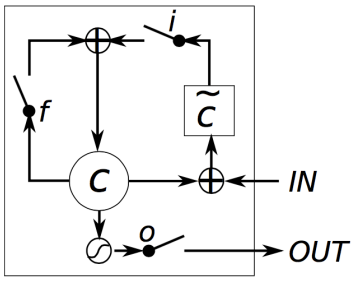


A recurrent neural network and the unfolding in time of the computation involved in its forward computation. Source: Nature

Long-Short Term Memory:
LSTMs were designed as a special RNN to combat vanishing gradients through a gating mechanism. It basically has the same structure as standard RNN besides it adds three gates to control information flow as follows:

$$
\begin{aligned}
i &= \sigma(x_t U^i + s_{t-1} W^i) \\
f &= \sigma(x_t U^f + s_{t-1} W^f) \\
o &= \sigma(x_t U^o + s_{t-1} W^o) \\
g &= \ tanh(x_t U^g + s_{t-1} W^g) \\
c_t &= c_{t-1} \circ f + g \circ i \\
s_t &= \tanh(c_t) \circ o
\end{aligned}
$$

Here is what a typical LSTM looks like:

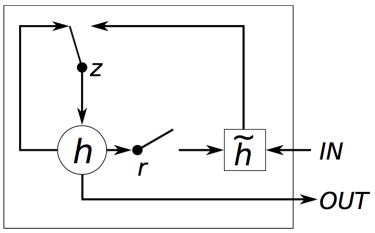


LSTM Gating. Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." (2014)

Gated Recurrent Unit:
A GRU has two gates, a reset gate r, and an update gate z. Intuitively, the reset gate determines how to combine the new input with the previous memory, and the update gate defines how much of the previous memory to keep around. The idea behind a GRU layer is quite similar to that of a LSTM layer, as are the equations.

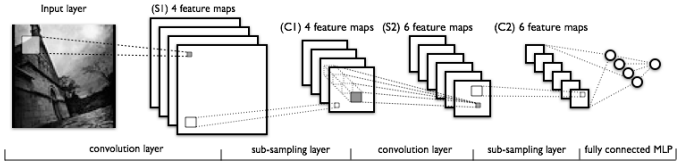
$$
\begin{aligned}
z &= \sigma(x_t U^z + s_{t-1} W^z) \\
r &= \sigma(x_t U^r + s_{t-1} W^r) \\
h &= tanh(x_t U^h + (s_{t-1} \circ r) W^h) \\
s_t &= (1 - z) \circ h + z \circ s_{t-1}
\end{aligned}
$$

Here is what a typical GRU looks like:



GRU Gating. Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." (2014)

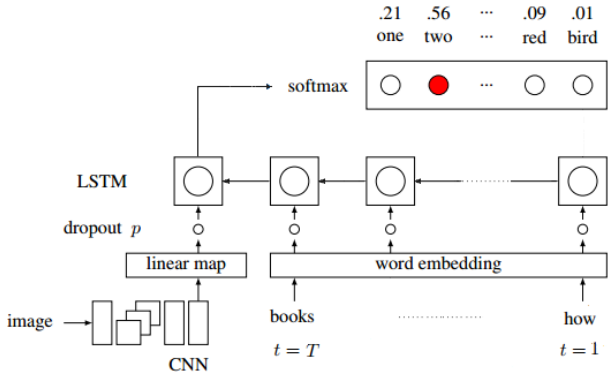
Convolutional Neural Network:
A convolutional neural network forms by convolutional layer, max-pooling layer, and fully-connected layer. CNN is not the main focus in this paper, so I borrow the model from Oxford VGG Conv Net [3]. Here is what a typical CNN looks like:



LeNet.http://andrew.gibiansky.com/blog/machine-learning/convolutional-neural-networks/

II.2 Learning Model

This section summarizes the neural network model and parameters settings I used. The basic logic follows Aishwarya Agrawal and his colleague's work. I tried to use RNN/LSTM/GRU with different layers to encode the questions, and the last hidden layer of the 19-layer Oxford VGG Conv Net trained on ImageNet 2014 Challenge [3]as my image embedding. After that, the LSTMs outputs are fed into a softmax layer at the last time step



Learning Model In General:

Optimization	Value	Detail
Learning Rate	4e-4	learning rate
Decay	0.95	learning rate decay
Decay Period	15	number of epochs to start decaying the learning rate
Alpha	0.8	alpha for adam
Beta	0.999	bata used for adam
epsilon	1e-8	epsilon that goes into denominator for smoothing
Batch size	200	batch size
Max epochs	50	Number of full passes through the training data
Dropout	0.5	Dropout

1. **CNN＋RNN** : The first model is 19-layer CNN and 2-layer RNN with 512 internal states.
2. **CNN＋deep-RNN** : The second model is 19-layer CNN and 3-layer RNN with 768 internal states.

3. **CNN＋LSTM**: The third model is 19-layer CNN and 2-layer LSTM with 512 internal states.
4. **CNN＋deep-LSTM:** The fourth model is 19-layer CNN and 3-layer RNN with 768 internal states.
5. **CNN＋GRU:** The fifth model is 19-layer CNN and 2-layer GRU with 512 internal states.
6. **CNN＋deep-GRU:** The sixth model is 19-layer CNN and 3-layer RNN with 768 internal states.

II.3 Evaluation

To evaluate performance, we simply count the number of correct prediction with the ground truth in the evaluation class.

$$Accuracy = \frac{Number\ of\ Correct\ Prediction}{Total\ number\ of\ Sample}$$

II.4 Data Sets

The datasets in this experiment uses 123,287 images from MS COCO dataset [6]. 82,782 images are used as training set, and the rest images are used for testing. The dataset consists of three parts: images, questions and sample answers. Three questions were collected for each image. Each question was answered by ten subjects along their confidence [2]. The sample of the data set is provided in Appendix.

III. RESULTS

III.1 Overall performance

Table I summarizes the learning performance of each model on COCO-GA. I notice both CNN+LSTM and CNN+GRU achieves around 45.7% accuracy in the 50th iteration. This is really close to the VIS+LSTM result (~48%) achieved by Mengye Ren [2]. Considering the super large image dataset and different kinds of questions we test with, I believe both model achieves reasonably good result. Surprisingly, all deep recurrent neural networks achieve worse result than the shallow one.

Model	Accuracy	Model	Accuracy
CNN+RNN	0.441511	CNN+deep RNN	0.435681
CNN+LSTM	0.457488	CNN+deep LSTM	0.454385
CNN+GRU	0.457661	CNN+deep GRU	0.445394

Table 1. COCO-QA accuracy per model (Epoch 50)

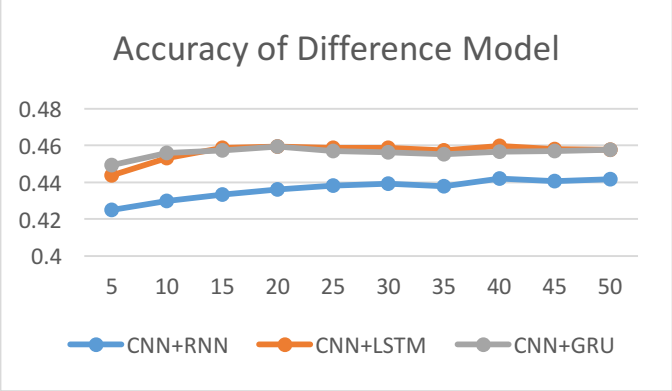
III.2 Effect of Sample Size

Table II records the learning result of each model during the different iteration. Number of epoch indicates the number of full passes through the training data.

	Epoch Period				
CNN + RNN	Epoch 5	Epoch10	Epoch15	Epoch 20	Epoch25
	0.424937	0.429696	0.433491	0.436281	0.438019
	Epoch 30	Epoch 35	Epoch 40	Epoch 45	Epoch 50
	0.439138	0.437903	0.442135	0.440521	0.441511
CNN + deep RNN	Epoch 5	Epoch10	Epoch15	Epoch 20	Epoch25
	0.41093	0.421044	0.423324	0.425571	0.431688
	Epoch 30	Epoch 35	Epoch 40	Epoch 45	Epoch 50
	0.432421	0.432000	0.435145	0.435039	0.435681
CNN + LSTM	Epoch 5	Epoch10	Epoch15	Epoch 20	Epoch25
	0.443625	0.453034	0.458632	0.459530	0.458805
	Epoch 30	Epoch 35	Epoch 40	Epoch 45	Epoch 50
	0.458854	0.457439	0.459637	0.457934	0.457488
CNN + deep LSTM	Epoch 5	Epoch10	Epoch15	Epoch 20	Epoch25
	0.401810	0.422484	0.427673	0.444382	0.444826
	Epoch 30	Epoch 35	Epoch 40	Epoch 45	Epoch 50
	0.449198	0.451133	0.453907	0.453627	0.454385
CNN + GRU	Epoch 5	Epoch10	Epoch15	Epoch 20	Epoch25
	0.449462	0.455957	0.457201	0.459250	0.456813
	Epoch 30	Epoch 35	Epoch 40	Epoch 45	Epoch 50
	0.456319	0.455175	0.456739	0.456797	0.457661
CNN + deep GRU	Epoch 5	Epoch10	Epoch15	Epoch 20	Epoch25
	0.445881	0.451767	0.453108	0.448984	0.448441
	Epoch 30	Epoch 35	Epoch 40	Epoch 45	Epoch 50
	0.446160	0.445211	0.447050	0.445913	0.445394

Table 2. COCO-QA accuracy during different epoch period

From Table 2, we can plot the learning curves along the increasing number of full passes through the training data. I will address the discussion on the DISCUSSION section.



Plot 1. COCO-QA accuracy during different epoch period

IV. DISCUSSION

Examine the results from the Table I and Table II, the combination of deep convolutional neural network and gated recurrent unit achieves the best result. Even the long-short term memory achieves very similar result, it should be noticed that GRU model is much easier to implement than the LSTM. As a new recurrent gated model, GRU is still not fully studied. It may achieve even better result in the future in a simpler way. The combination of deep CNN and traditional

recurrent neural network achieves worst result among those three as expected. However, it only draws down the result by 1-2%. This is because the question in general is quite short. Simply RNN is also capable to capture the relation between questions and images. Most importantly, CNN+RNN has a much shorter running time; it runs 50% time faster than LSTM and GRU. To be fair, LSTM and GRU are much more computational expensive.
Unfortunately, the deep recurrent neural networks do not boost the performance as expected. The reason should be further studied. An intuition guess is a deep recurrent neural network may overfit the data since the sentence is very short in general.

Answering questions based on the image may requires different kind of knowledge. Fine-grained recognition (e.g. Appendix question 2: "is there a shallow") and Object detection (e.g. Appendix question 3: "Is this one bench or multiple bench") are the easiest question that can be answered, which achieves excellent result. Knowledge based reasoning (e.g. Appendix question 1: "what shape is the bench seat.") are really hard to answer since it requires background knowledge and common sense in general. For more information, reader can check the sample given in Appendix (I), or download the source code and try it yourself.

V. CONCLUSION

In this paper, different models are used to address the image question and answering problem. All the tested model shows reasonable understanding of the question. Different kinds of recurrent neural network are becoming more and more popular for natural language processing problem. I have shown the combination between recurrent neural network and deep convolutional neural network might be the right idea to solve image question answering problem eventually. Even gated recurrent unit model as mentioned above achieves 45.7% accuracy for answering different kinds of questions. We still have to admit this QA model is still pretty naïve, and is far away from turning it into the real world application.

VI. REFERENCES

[1]Antol, Stanislaw, et al. "VQA: Visual question answering." Proceedings of the IEEE International Conference on Computer Vision. 2015.
[2]Ren, Mengye, Ryan Kiros, and Richard Zemel. "Exploring models and data for image question answering." Advances in Neural Information Processing Systems. 2015.

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[4] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[5] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory."Neural computation 9.8 (1997): 1735-1780.

[6] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft ´ COCO: Common Objects in Context," in ECCV, 2014.

[7] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[8] Mikolov, Tomas, et al. "Recurrent neural network based language model."INTERSPEECH. Vol. 2. 2010.

VII. Appendix (I) – Sample Question & Answers

This section includes six interesting question answering results generated by six different models. The result is shown in page 5.

- The red highlight indicates the wrong prediction
- The green highlight indicates the correct prediction.
- The light green highlight indicates the reasonable prediction

VIII. Appendix (II) — Source Code

https://github.com/shawnxu1318/Image-Question-Answering

IX. Appendix (III) — Fun Part (Turing Test)

This part includes a simple Turing test. The survey is shown in page 6. The test is very simple. I label 6 different models as player 1 – 6, and present the corresponding answers to people. I simply ask people, among those six players, how many human players are in those players? The real answer is of course 0.

The result is following:

Prediction: Number of Human Player						
0	1	2	3	4	5	6
5	16	7	1	1	0	0

Only 16.7% people answered correctly. Most of the people thought player 2, 3, 6 are human players, so results generated by CNN + LSTM, CNN + GRU and CNN + dGRU could be deceiving. : D

Question			
What shape is the bench seat?			
True Answer		Human Predict	
Oval		Oval	
Predict Answer			
CNN+RNN	Oval	CNN+dRNN	Rectangle
CNN+LSTM	Triangle	CNN+dLSTM	Round
CNN+GRU	Rectangle	CN+dGRU	Circle
Question			
Is there a shadow?			
True Answer		Human Predict	
Yes		Yes	
Predict Answer			
CNN+RNN	Yes	CNN+dRNN	Yes
CNN+LSTM	Yes	CNN+dLSTM	Yes
CNN+GRU	Yes	CN+dGRU	Yes
Qestion			
Is this one bench or multiple benches?			
True Answer		Human Answer	
Multiple		2	
Predict Answer			
CNN+RNN	Commercial	CNN+dRNN	Low
CNN+LSTM	2	CNN+dLSTM	Yes
CNN+GRU	1	CN+dGRU	1

Qestion			
Is this a modern train?			
True Answer		Human Predict	
No		No	
Predict Answer			
CNN+RNN	No	CNN+dRNN	No
CNN+LSTM	No	CNN+dLSTM	No
CNN+GRU	No	CN+dGRU	No
Qestion			
What is the color of the train?			
True Answer		Human Predict	
Green		Indigo	
Predict Answer			
CNN+RNN	Red	CNN+dRNN	Green
CNN+LSTM	Blue	CNN+dLSTM	Red
CNN+GRU	Blue	CN+dGRU	Green
Qestion			
What is on the other side of the train?			
True Answer		Human Answer	
Trees		Trees	
Predict Answer			
CNN+RNN	Trees	CNN+dRNN	Trees
CNN+LSTM	Fan	CNN+dLSTM	Tree
CNN+GRU	Right	CN+dGRU	Trees

Question			
What shape is the bench seat?			
Predict Answer			
1	Oval	4	Rectangle
2	Triangle	5	Round
3	Rectangle	6	Circle

Question			
Is there a shadow?			
Predict Answer			
1	Yes	4	Yes
2	Yes	5	Yes
3	Yes	6	Yes

Qestion			
Is this one bench or multiple benches?			
Predict Answer			
1	Commercial	4	Low
2	2	5	Yes
3	1	6	1

Qestion			
Is this a modern train?			
Predict Answer			
1	No	4	No
2	No	5	No
3	No	6	No

Qestion			
What is the color of the train?			
Predict Answer			
1	Red	4	Green
2	Blue	5	Red
3	Blue	6	Green

Qestion			
What is on the other side of the train?			
Predict Answer			
1	Trees	4	Trees
2	Fan	5	Tree
3	Right	6	Trees