Exploring Models for Image Question Answering

Luyao Zhou, Yujia Li
Department of Computer Science, University of Virginia, Charlottesville, VA 22904
[lz6nf, yl9ap]@virginia.edu

Abstract

The problem of jointly learning image and text through a question-answering task has become really popular nowadays. Several remarkable deep learning models are proposed to solve this problem. In our work, we first re-implement one model called VIS+LSTM proposed in one previous Image Question Answering (Image QA) paper [3], Exploring Models and Data for Image Question Answering. We then refer and modify the model to create two more models, and test all three models on the dataset, COCO-QA, proposed in the same paper. We present analysis about these models, their weaknesses and potential ways to improve performance.

1. Introduction

With the advances of natural language processing (NLP) and computer image understanding, Image Question Answering (Image QA) has become one of the most active research areas nowadays. Image QA problems require to answer given questions based on the context of a given image. Neural networks are widely used as the solution to Image QA problems, and the most common approach is the combination of a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN). CNN is used to extract a global image feature vector of the given image, and the RNN is used to encode the given questions. We implement, test and analyze three Image QA models based on the combination of CNNs and RNNs.

In this work we first re-implement one model called VIS+LSTM proposed in the previous Image QA work [3], Exploring Models and Data for Image Question Answering. This model is a generic end-to-end QA model using visual semantic embeddings to connect a CNN and a Long Short Term Memory (LSTM) network. Here LSTM is one kind of RNN.

Second, we think that by providing the image features to the LSTM as a reference at every timestep may help the model find the correct answer to the question. Inspired by another Image QA paper [1], Ask Your Neurons: A Neural-based Approach to Answering Question about Image, we modify our model and test on the same dataset to see if we can gain any improvement.

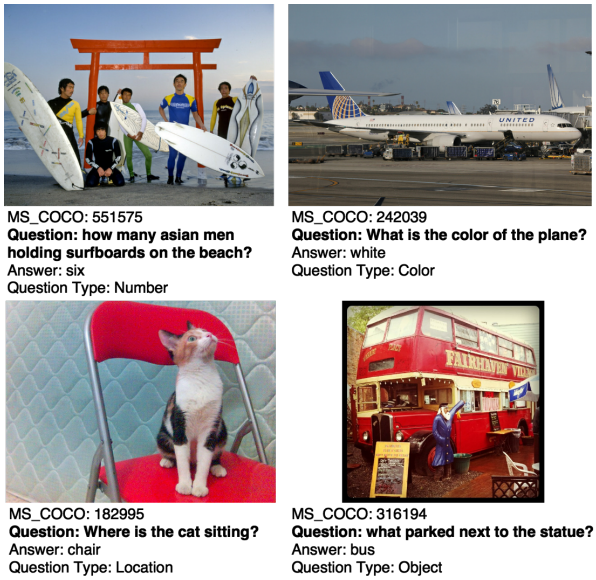


Figure 1. Here we show some sample images from the COCO-QA Dataset that we are using in our work.

Third, we use the idea of adding attention layers to the network, proposed by the paper, Stacked Attention Networks for Image Question Answering [6], and modify the VIS+LSTM model again to include attention layers. The model we construct is inspired by the model proposed in the paper, Stacked Attention Networks (SANs), which allows multi-step reasoning for Image QA.

In this work we assume that the answers consist of only a single word, which change this task into a classification problem. This also makes the evaluation of the models easier. We test all three models on the dataset, COCO-QA, proposed in the original paper. All the models are implemented using PyTorch, the newly released deep learning package written in Python.

2. Related Work

Early works on Image Question Answering include [3, 1] where CNN and LSTM jointly work together to predict answers. In these works, they use pre-trained CNN to extract features from each image, and use LSTM to process the question sequence. Some recent works add extra layers to their model to improve the model complexity and make better predictions. In some previous works [2] researchers add an extra layer to dynamically predict the parameters for classification layer. In stacked attention networks [6], researchers stack multiple attention layers on top of the CNN to search image regions for the answer.

3. Model

In this project, we experiment three models, VIS+LSTM, VIS+LSTM-2, and SANs. Detailed descriptions of the models are below.

3.1. VIS+LSTM

We use the last layer of VGG-19 network [5] pertained on ImageNet [4] to extract image features. Then we use a linear layer to map the 4096 dimension image features to the same size (512) as word embeddings. We first preprocess all the words in the questions dataset. Each unique word in the question dataset is assigned a unique index. We use the embedding layer from PyTorch as the embedding layer. The image is treated as the first word of the question. We feed one word into the LSTM at each time step. The output of the LSTM at the last time step is fed into a linear classifier layer and we use softmax to predict the answer. The vocabulary size of COCO-QA answers is 430. In this model, the pretrained VGG-19 is kept frozen during training. Image linear layer, word embedding layer, LSTM, and linear classifier are trained together.

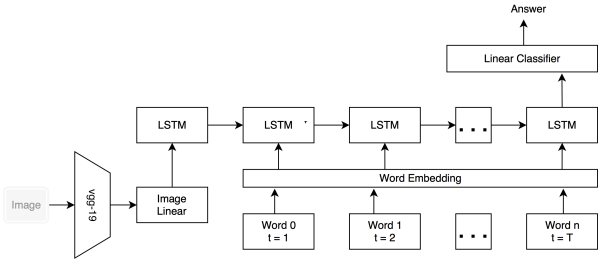


Figure 2. VIS+LSTM Model

3.2. VIS+LSTM-2

This model is a variation of VIS+LSTM model. We think that providing the image features as a reference at every time step when the LSTM receives a word may help the model to predict the answer. Instead of treating the image as the first word of the question, we concatenate the image features to each word in the question sequence. So we feed the word embedding as well as the image features into the LSTM at every time step.

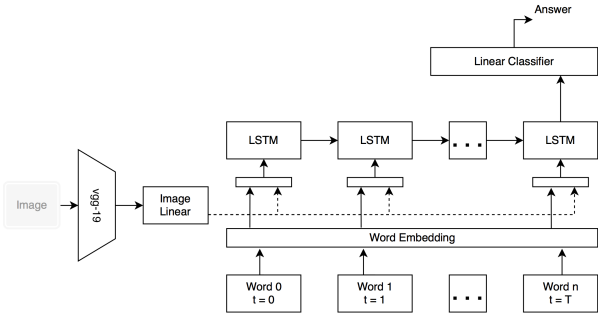


Figure 3. VIS+LSTM-2 Model

3.3. SANs

We then explore the third model to improve the performance on the image question answering tasks. In this model, we use pretrained VGG-19 on ImageNet to extract image features again. But instead of using the output from last fully connected layer, we use the output before the fully connected layer. The image is then scaled to 448 x 448, so the output features have a dimension of 512 x 14 x 14. 14 x 14 represents the number of regions in the image, so each number represents a 32 x 32 region in the original image. 512 is the number of filters in the convolutional network. Then we use a linear layer to transform the feature vector of each image region from 512 to be the same dimension as the question vector, which is 1024 in this case. We use v_I to denote the 196 x 1024 image features matrix.

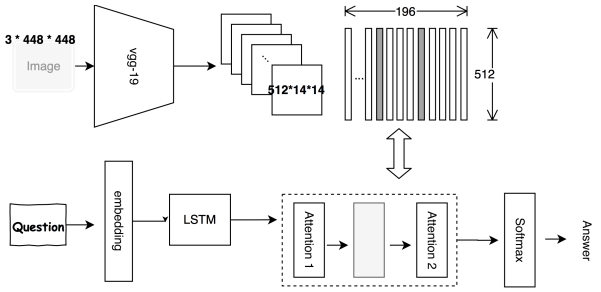


Figure 4. SANs Model

We use the same technique to handle the question embedding and the LSTM. Each word in the question sequence is assigned a word id, and we embed the word using embedding layers. Every embedded word is fed into LSTM at each time step. We take the output from LSTM at the last time step as the representation vector for the question, denotes as v_Q.

In most of the cases, an answer of a question only relates to some regions in the image. We stacked some attention

layers to do multi-step reasoning using the image feature matrix v_I and question feature vector v_Q [6]. Since we already have the image feature vector of 196 regions in the original image, the idea is to use the question vector to highlight certain image regions to find the answer. The image feature matrix v_I and question feature vector v_Q are both fed into an attention layer to generate the attention probability of 196 regions. In the attention layer, image features and question features are sent into linear layers as shown in (1) (no bias for image linear layer). Next, we do a $\oplus$ operation between image features matrix and question features vector. Question vector is added to every column of the image matrix. Then we use softmax to calculated the attention distribution p.

$$h_a = tanh(W_{I,A} v_I \oplus (W_{Q,A} v_Q + b_A)) \tag{1}$$

$$p_I = softmax(W_P h_A + b_P) \tag{2}$$

Based on the attention distribution, we calculated the weighted sum of image vectors of each region i. The output is then combined with question vector to form a refined query vector u [6]. u is treated as the refined question vector which is fed into the next attention network.

$$\widetilde{v_I} = \sum_i p_i v_i \tag{3}$$

$$u = \widetilde{v_I} + v_Q \tag{4}$$

In the k^{th} attention layer, u^{k-1} and v_I is treated as the input, and similar operation is performed to get the new query vector u^k. Finally probability of answer is calculated by performing softmax on u^k.

4. Experiments and Results

We evaluate all three model on the COCO-QA dataset. COCO-QA is proposed in [3]. Based on MS-COCO image caption dataset, the authors use a question generation algorithm to generate question answer pairs using the caption of each image. There are 78736 and 38948 question answer pairs in the training set and testing set respectively. There are four different question types, including object, number, color and location. All answers are single word in order to keep this a classification problem. There are 430 unique words in the answer set. The distribution of the most common answer in the dataset is 2.65%, and the distribution of the least common answer is 0.02%. We consider this dataset to be balanced.

The best test accuracy we get using VIS+LSTM model is 52.3%, which is slightly lower comparing to the original published result (53.3%). We think this result is comparable to the original implementation, although better result can be obtained if we train this model on better hyper-parameters.

The best test accuracy we get using VIS+LSTM-2 is 51.1%. In the paper Neural VQA [1], which implements a similar model, the researchers train and test their model on DAQUAR. DAQUAR is another Image-QA dataset believed to have more difficult questions comparing to COCO-QA [3]. The Neural-Image-QA has a test accuracy of 19.43% on DAQUAR single word image-question pairs. We originally design the model to believe it would have better performance comparing to VIS+LSTM model, as we think it delivery more information of images to the model which helps answering the question. The result turns to be both of them have about the same performance.

We randomly sample some images from the test set, and analyze the reasons why the above two models make mistakes. (i) The model finds the correct object, and the prediction is correct, but the answer use a different word with same meaning. In Figure61932 9 (a), the model prediction is bird, but the correct answer is pigeon. (ii) The model finds the correct object, but it does not correctly identify the object. In Figure 9 (c), the model predicts cow, but the correct answer is bull. In this image, it is very hard to tell whether the object laying in the hay is bull or cow, especially when the question mentions cow. (iii) The question is too vague as there might be multiple correct answers. In Figure 9 (f), the answer station and train can both be correct to the question "what is photographed in black and white". (iv) The model does not correctly identify the object to answer the question. Sometimes, the model needs some logic sense to answer tough questions. The evaluation method for our model is not perfect. Other evaluation methods like top 5 prediction or WUPS score can be used to avoid (i). Since COCO-QA dataset is automatically generated using algorithm, there exists many vague questions that are hard to answer even by human. The quality of the dataset can be improved by human post-processing or apply stronger algorithms to generate the questions.

We also re-implemented the SANs model. We notice that this model overfits very fast when no dropout layers are included. We then add dropout layers ($p = 0.5$) in word embedding, image feature transfer, and attention layer. One interesting fact we notice is that the test accuracy is always higher than the train accuracy. The train data and test data all come from independent dataset. We believe one reason can be that the dropout layers are deactivated when the train mode of our module is set to False. The best test accuracy achieved on this model is only 47.23%, which is much lower than 61.00% reported from the original paper. We also tried to analyze what is the reason behind. Since we don't have the access to the details implementation of the original paper, we might have done things differently. We added some test code that highlights the attention area in the input image, we notice that the attention is focus on wrong areas in some cases. There are still much room of improve-

ments, for example, improving the attention layers, trying other optimization method, applying hyper-parameter searching, and adding tokenizer to question-answer pairs in pre-processing section to reduce the vocabulary size.

Method	Best Accuracy
GUESS	0.073
LSTM	0.368
BOW	0.375
VIS+LSTM [3]	0.533
SANs [6]	0.610
VIS+LSTM	0.523
VIS+LSTM-2	0.511
SANs	0.4723

Table 1. Best accuracy for models on COCO-QA dataset.

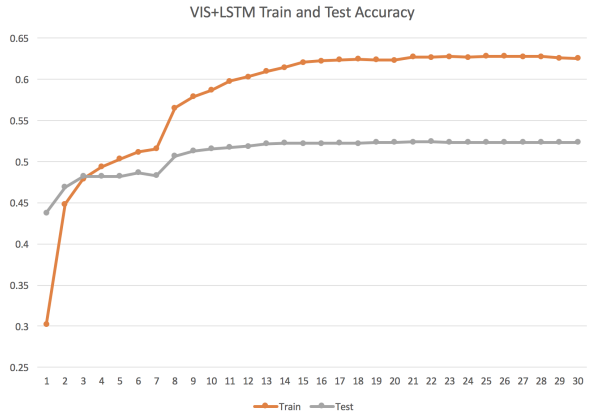


Figure 5. Train and test accuracy for VIS+LSTM model.

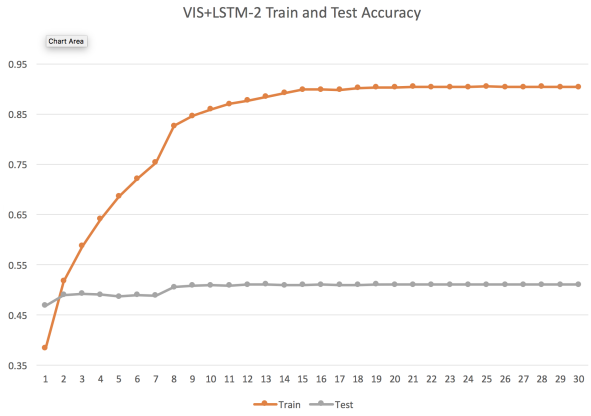


Figure 6. Train and test accuracy for VIS+LSTM-2 model.

5. Conclusion

In this project, we exploring three different models based on previous works to solve Image Question Answering problem. The models includes VIS+LSTM, VIS+LSTM-2, and SANs. We implement all models using PyTorch, a newly released deep learning package written in Python, and train and test three models using COCO-QA dataset. The best accuracy we gain is 52.33% using VIS+LSTM model.

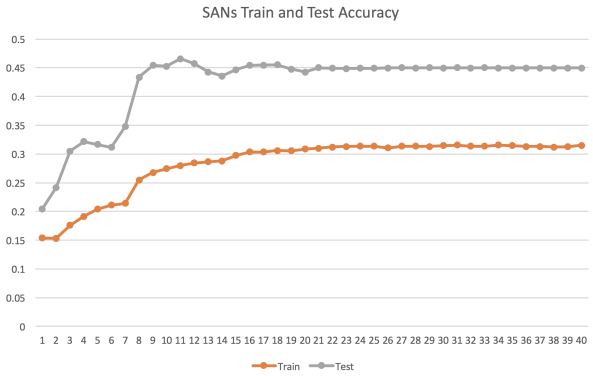


Figure 7. Train and test accuracy for SANs model.

References

[1] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. *CoRR*, abs/1505.01121, 2015.

[2] H. Noh, P. H. Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. *CoRR*, abs/1511.05756, 2015.

[3] M. Ren, R. Kiros, and R. S. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *CoRR*, abs/1505.02074, 2015.

[4] 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.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[6] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015.

Image id: 571437
Question: what topped with the mix of vegetables and fruit
Correct Answer: plate
Prediction: plate

Image id: 359303
Question: what are flying in an open sky
Correct Answer: airplanes
Prediction: airplanes

Image id: 198486
Question: where does the person take a selfie
Correct Answer: mirror
Prediction: mirror

Image id: 298762
Question: two horses pulling what down the road
Correct Answer: carriage
Prediction: carriage

Figure 8. Here we show some correctly answered image question pairs by VIS+LSTM-2 model.

(a) Image id: 300725
Question: what perched on arm of park bench
Correct Answer: pigeon
Prediction: bird

(b) Image id: 288062
Question: where do the colorful toilet seat open
Correct Answer: stall
Prediction: bathroom

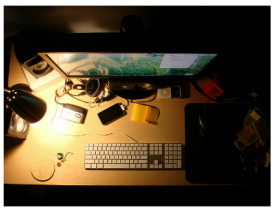
(c) Image id: 183390
Question: what is laying in the hay with another cow
Correct Answer: bull
Prediction: cow

(d) Image id: 407149
Question: what is shown with the couch and table
Correct Answer: scene
Prediction: room

(e) Image id: 464706
Question: what does the lone pilot fly upward in the sky
Correct Answer: jet
Prediction: sky

(f) Image id: 496765
Question: what is photographed in black and white
Correct Answer: station
Prediction: train

(g) Image id: 52947
Question: what filled with all types of different types of donuts
Correct Answer: box
Prediction: plates



(h) Image id: 234341
Question: assorted what on desk with large monitor
Correct Answer: equipment
Prediction: computer

Figure 9. Here we show some incorrect answered image question pairs by VIS+LSTM-2 model.

Image id: 338600
Question: what is the color of the field
Correct Answer: green Prediction: green

Image id: 354537
Question: what is the color of the car
Correct Answer: blue Prediction: blue

Image id: 71027
Question: what are there pulling a man and a cart
Correct Answer: horses Prediction: horses

Image id: 488374
Question: what stand in the middle of a dirt path
Correct Answer: sheep Prediction: sheep

Figure 10. Here we show some correct answered image question pairs by SANs model.

Image id: 21639
Question: what is sitting on a decorative table cloth
Correct Answer: cake
Prediction: cat

Image id: 232511
Question: what does the man drive
Correct Answer: scooter
Prediction: bicycle

Image id: 468577
Question: where is the young girl lying
Correct Answer: crib Prediction: bed

Image id: 516828
Question: what topped with mushroom and barbecue on a disposable plate
Correct Answer: slices
Prediction: pizza

Figure 11. Here we show some incorrect answered image question pairs by SANs model.