CS698O: Project Proposal
Modified Stacked Attention Networks for VQA using Visual grounding of Phrases

Abhishek Verma(14026) , Ayush Tulsyan(14167)

November 25, 2017

1 Introduction: What is the problem. Why solve it?

Computer vision is the science and technology of obtaining models, meanings, and control information from visual data. This field has influenced the field of Artificial intelligence greatly. The two important fields of computer vision are computational vision and machine vision amongst which computational vision deals with recording, analyzing and trying to understand a visual perception.

One of the important problem is the field of computer vision is visual question answering(VQA). VQA is the area which tries to find answer to a problem given an image and a question using a combination of natural language processing and computer vision techniques. One of the major steps in solving the VQA problem is visual grounding of phrases and features because to identify objects in the image, the subject is interacting with, one has to identify/visually ground the subject first along with the set of objects. Visual grounding problem takes image-sentence pairs as input and learns to visually ground (i.e., localize) arbitrary linguistic phrases. One may think that this problem is similar to the object detection problem. But that's not the case. In object detection, given a set of labels, we need to locate them in the visual whereas in visual grounding problem, given a natural language phrase, we need to analyze it and localize its context in the visual. In this project, we'll deal with images.

Visual grounding has a direct use in many fields such as surveillance AIs (to check if a certain activity has taken place in a video), daily household AIs(to find the location about which AI has been given instructions), etc.

2 Related Work

Image question answering is closely related to image captioning [1], [2], [3], [4], [5], [6], [7]. In [4], the system first extracted a high level image feature vector from GoogleNet and then fed it into a LSTM to generate captions. The method proposed in [2] went one step further to use an attention mechanism in the caption generation process. The approach proposed in [3] first used a CNN to detect words given the images, then used a maximum entropy language model to generate a list of caption candidates, and finally used a deep multimodal similarity model (DMSM) to rerank the candidates. Instead of using a RNN or a LSTM, the DMSM uses a CNN to model the semantics of captions.

Several data sets have been constructed in [8], [9], [10], [11] either through automatic generation based on image caption data or by human labeling of questions and answers given images. Among them, the image QA data set in [9] is generated based on the COCO caption data set.

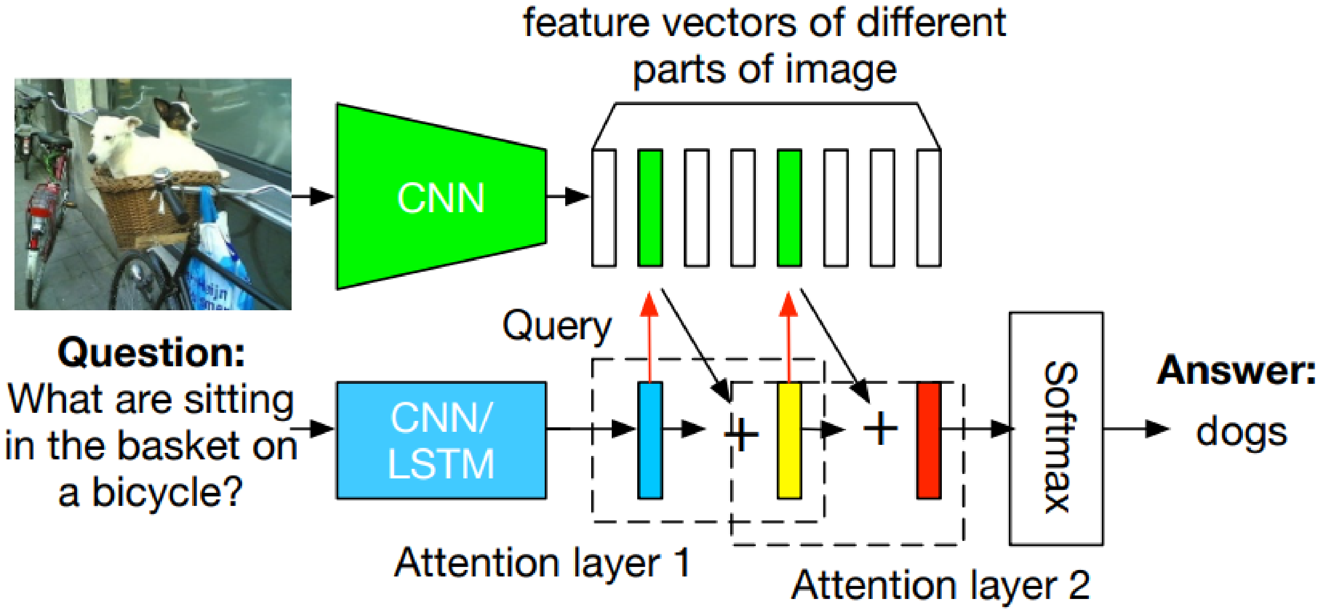
Several image QA models were proposed in the literature. [12] used semantic parsers and image segmentation methods to predict answers based on images and questions. [8] and [10] both used encoder-decoder framework to generate answers given images and questions. They first used a LSTM to encode the images and questions and then used another LSTM to decode the answers. [11] first encoded questions with LSTMs and then combined question vectors with image vectors by element wise multiplication. [13] used a CNN for question modeling and used convolution operations to combine question vectors and image feature vectors[14].

In this project, we'll be using Stacked Attention Networks(SANs) introduced in [14]. We'll also modify [14] by the methods suggested in [15] and check if it'll lead any significant improvement in image question answering problem.

3 Existing solution and novel extensions planned

Problem statement: Given an image and different query phrases, return an answer to the query phrases in context to the image.
Existing solution: We'll follow the method described in [14], i.e., we'll use Stacked Attention Networks(SANs) that learn to answer natural language questions from images. SANs use semantic representation of a question as query to search for the regions in an image that are related to the answer[14]. The SAN contains three major components:

- **Image model:** uses a CNN to extract high level image representations.
- **Question model:** uses a CNN/LSTM to extract semantic vector of a question.
- **Stacked attention model:** locates, via multi-step reasoning, the image regions that are relevant to the question for answer prediction.


feature vectors of different parts of image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/ LSTM
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

(a) Stacked Attention Network for Image QA

We'll also experiment with some modifications in the model. These modifications are mostly inspired by [15]. These include:

- **Modification in Image model:** We plan to replace the max pooling layer of VGGnet by average pooling layer as average-pooling better preserves location information since it is forced to localize in order to maximize its response over the relevant image regions[16].

- **Modification in Question model:** To better model long phrases, we adopt LSTM cells [17] in a two-layer RNN, with a Dropout module [18] inserted in between to prevent over-fitting. We'll pre-train the weights of our language encoder on a combined set of Googles Billion Words dataset and COCO captions in the training set with the next word prediction task.
- **Introduction of Semantic embedding module:** The output of the visual encoder would be projected into the semantic space by the semantic embedding module, which is a two-layer-perceptron with Dropout inserted in between the layers.
- If time permits, we'll also try to introduce structural loss and discriminative loss(described in [15]) in our model.

4 Expected Timeline

- **October 29th 2017(mid-project review):** We plan to implement the paper[14] in PyTorch by this time. Currently, the only codebase available for this paper is implemented in Torch and Theano. This would be the base of the modifications we'll be making in the model afterwards.
- **November 12th 2017(end-project review):** We are looking forward to incorporate as many of the above mentioned modifications possible by this time. We'll then experiment with different datasets such as DAQUAR-AL[12], DAQUAR-REDUCED, COCO-QA[9] and VQA[11].

References

[1] X. Chen and C. L. Zitnick. *Learning a recurrent visual representation for image caption generation.* arXiv preprint arXiv:1411.5654, 2014.

[2] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. *Show, attend and tell: Neural image caption generation with visual attention.* arXiv preprint arXiv:1502.03044, 2015.

[3] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, et al. *From captions to visual concepts and back.* arXiv preprint arXiv:1411.4952, 2014.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. *Show and tell: A neural image caption generator.* arXiv preprint arXiv:1411.4555, 2014.

[5] R. Kiros, R. Salakhutdinov, and R. S. Zemel. *Unifying visual-semantic embeddings with multimodal neural language models.* arXiv preprint arXiv:1411.2539, 2014.

[6] A. Karpathy and L. Fei-Fei. *Deep visual-semantic alignments for generating image descriptions.* arXiv preprint arXiv:1412.2306, 2014.

[7] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. *Deep captioning with multimodal recurrent neural networks (m-rnn).* arXiv preprint arXiv:1412.6632, 2014.

[8] M. Malinowski, M. Rohrbach, and M. Fritz. *Ask your neurons: A neural-based approach to answering questions about images.* arXiv preprint arXiv:1505.01121, 2015.

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

[10] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. *Are you talking to a machine? dataset and methods for multilingual image question answering.* arXiv preprint arXiv:1505.05612, 2015.

[11] 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.

[12] 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 16821690, 2014.

[13] L. Ma, Z. Lu, and H. Li. *Learning to answer questions from image using convolutional neural network.* arXiv preprint arXiv:1506.00333, 2015

[14] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng and Alex Smola. *Stacked Attention Networks for Image Question Answering.* arXiv:1511.02274

[15] Fanyi Xiao, Leonid Sigal and Yong Jae Lee. *Weakly-supervised Visual Grounding of Phrases with Linguistic Structures.* arXiv:1705.01371

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. *Learning deep features for discriminative localization.* In CVPR, 2016.

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

[18] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. *Dropout: a simple way to prevent neural networks from overfitting.* Journal of Machine Learning Research, 2014.