Visual Question Answer using Co-Attention

DIP Project Report - Building a VQA Model

Sahaj Agarwal
Indian Institute Of Information
Technology, Sri City, Chittoor
CSE S20160010077
sahaj.a16@iiits.in

Garvit Kataria
Indian Institute Of Information
Technology, Sri City, Chittoor
CSE S20160010028
garvit.k16@iiits.in

Anubhav Ujjawal
Indian Institute Of Information
Technology, Sri City, Chittoor
CSE S20160010005
anubhav.u16@iiits.in

Anurag Gupta
Indian Institute Of Information
Technology, Sri City, Chittoor
CSE S20160010006
anurag.g16@iiits.in

Laisha Wadhwa
Indian Institute Of Information
Technology, Sri City, Chittoor
CSE S20160010048
laisha.w16@iiits.in

***Abstract*—A number of works before "Hierarchical Co-Attention paper" [1] proposed attention models for Visual Question Answering (VQA) that generate spatial maps highlighting image regions relevant to answering the question. The Co-Attention mechanism proposed by Dhruv Batra et al. [1] highlights the importance of Hierarchical Question-Image Co-Attention and improves on the then state of the art. We aim to build up on this approach to improve the results stated in the paper [1].**

Keywords—Co-Attention, Visual Question Answering, VQA

I. Introduction

Visual Question Answering (VQA) [1, 2, 3, 4, 5, 6] has emerged as a prominent multi-discipline research problem in both academia and industry. To correctly answer visual questions about an image, the machine needs to understand both the image and question and based on that understanding, come up with a relevant answer. In Recent works, visual attention [7–10] and Co-Attention [1] based models have been explored for VQA, where the attention mechanism typically produces a spatial map highlighting image regions relevant to answering the question and generates a energy map for the corresponding question highlighting the relevant words in the question to account for when spitting an answer. We aim to build up on the latter approach and improve the results stated in the paper.

II. Problem Statement

The problem statement assigned to our group (Group Serial - 23 ["DIP Groups"- Google Sheet]) as per the understanding of the team is stated as -

"To build and/or enhance a model or come up with a process with or without using the existing technologies with the language and framework of choice capable of VQA (Visual Question Answering), i.e, given an image and a question in relation to the image, it is capable of producing a relevant (subjective) and correct (subjective) response to the question with strong correlation to the image."

Any questions, corrections and/or comments to our understanding are welcome.

III. Literature Survey

Many recent works [2, 3, 11, 4, 5, 12, 13, 14] have proposed models for VQA. We compare and relate some of them.

Image Attention: Instead of directly using the holistic entire-image embedding from the fully connected layer of a deep CNN (as in [2, 4, 5, 15]), a number of recent works have explored image attention models for VQA. Earlier works added spatial attention to the standard LSTM model for pointing and grounded QA. Later works builded up on those by proposing a compositional scheme that consists of a language parser and a number of neural modules networks. The language parser predicts which neural module network should be instantiated to answer the question. Some other works perform image attention multiple times in a stacked manner. In [7], the authors generate image regions with object proposals and then select the regions relevant to the question and answer choice. Note that all of these approaches model visual attention alone, and do not model question attention. In [1], authors model both visual as well as question attention using Hierarchical Question-Answer Co-Attention mechanism.

Language Attention: Although relatively few number of works have explored question attention in VQA, there are some related works in natural language processing (NLP) in general that have modeled language attention. In order to overcome difficulty in translation of long sentences, Bahdanau et al. [16] propose RNNSearch to learn an alignment over the input sentences. Some recent works also proposed an attention model to circumvent the bottleneck caused by fixed width hidden vector in text reading and comprehension and a word-by-word neural attention mechanism to reason about the entailment in two sentences.

Co-Attention: Co-Attention [1] attends to the image and question simultaneously. Similar to [9], it connects the image and question by calculating the similarity between image and question features at all pairs of image-locations and question-locations.

00-01 ©2018 IEEE

Hierarchical Question-Image Co-Attention for Visual Question Answering: The main contribution by Dhruv Batra et al. through this paper [1] is listed below:

- A novel co-attention mechanism for VQA that jointly performs question-guided visual attention and image-guided question attention. This mechanism is explored through two strategies, parallel and alternating co-attention.
- A hierarchical architecture to represent the question and consequently construct image-question co-attention maps at 3 different levels: word level, phrase level and question level. These co-attended features are then recursively combined from word level to question level for the final answer prediction.
- A novel convolution-pooling strategy at phrase level to adaptively select the phrase sizes whose representations are passed to the question level representation.

IV. Overall Plan And Current Status

A. Overall Plan

- Re-implement the approach stated by Dhruv Batra et al. in "Hierarchical Question-Image Co-Attention for Visual Question Answering" [1] and re-establish the results stated by them in the paper.

- Try different architectures and learning hyper-parameters to optimize and improve the results stated in the paper [1].

B. Current Status

We are in the process of re-implementing the approach stated by Dhruv Batra et al. [1]. The original source on GitHub - https://github.com/jiasenlu/HieCoAttenVQA for the paper is implemented in Lua. We are using PyTorch (Deep Learning Framework by Facebook) as our default. So far, we have remodeled the architecture for the Co-Attention Network in Pytorch. Next, we plan to acquire the dataset from - http://visualqa.org/download.html , load and process it and train our network. Once we re-establish the results stated in the paper [1], we will employ techniques to improve them.

Acknowledgment

We would like to acknowledge the work by Dhruv Batra et al. in "Hierarchical Question-Image Co-Attention for Visual Question Answering" [1] as their approach has provided us with an excellent intuition on Co-Attention models to build upon. We would like to acknowledge all others as well who have provided their valuable contribution to the domain of VQA. Finally, we would like to thank Dr. Snehasis Mukherjee for providing us with this excellent opportunity to work and learn about and beyond the domain of VQA and the existing technologies used for modeling it.

References

Various papers and works have been referenced in the project report. The references used in the report are cited below:

[1] Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering" Virginia Tech, Georgia Institute of Technology, arXiv:1606.00061v5 [cs.CV], Jan 2017.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In ICCV, 2015.

[3] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In NIPS, 2015.

[4] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In ICCV, 2015.

[5] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In NIPS, 2015.

[6] C Lawrence Zitnick, Aishwarya Agrawal, Stanislaw Antol, Margaret Mitchell, Dhruv Batra, and Devi Parikh. Measuring machine intelligence through visual question answering. AI Magazine, 37(1), 2016.

[7] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In CVPR, 2016.

[8] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In ICML, 2016.

[9] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. arXiv preprint arXiv:1511.05234, 2015.

[10] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In CVPR, 2016.

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332, 2016.

[12] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. arXiv preprint arXiv:1511.05099, 2015.

[13] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. arXiv preprint arXiv:1606.01455, 2016.

[14] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847, 2016.

[15] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In AAAI, 2016.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In ICLR, 2015.