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*—Historically, building a system that can answer natural language questions about any image has been considered a very ambitious goal. With the advent of Deep Learning (DL), we have witnessed enormous research progress in Visual Question Answering (VQA), in such a way that systems capable of answering these questions are emerging with promising results. In a general way, we can define a VQA system as an algorithm that takes as input an image and a natural language question about the image and generates a natural language answer as the output. This is, by nature, a multi-discipline research problem. This report highlights a solution to the problem in the form of a simple architecture that is fully symmetric between visual and language representations. It also highlights the results obtained using this approach which represent the current state of the art in the VQA community.**

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 an energy map for the corresponding question highlighting the relevant words in the question to account for when spitting an answer.

The approach highlighted in this paper [17] takes advantage of earlier works in the sense that it makes use of feature extraction and refined Co-Attention mechanisms introduced in the earlier papers such as [1] but also introduces optimizations and a novel architecture which allows for more efficient and better representation of the language and visual features. It also uses a superior Multi-Modal Feature Fusion technique along with Dense Attention layers to achieve state of the art result on VQA COCO Image Dataset.

The architecture used in this paper [17] is simple in the sense that it is perfectly symmetric in terms of Language and Image representations and computationally dense as it takes into account the interaction between every Image region and each question word. Specifically, each Image region attends to every question word at three different levels and each level is attended by every image region for each question. This approach incorporating a hierarchy of attention mechanisms for language and visual representations was introduced by Devi Parikh et al. in [1] and proves effective in extracting useful and relevant information from natural language questions eliminating unnecessary words and enabling the model to focus on keywords required to output contextual answers. Furthermore, the fusion technique improves the correspondence between the representations of the two modalities improving the semantic understanding and the overall accuracy of the model.

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.”

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.

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.

Multi-Modal Feature Fusion: In the domain of VQA, the common approach that has been followed by most of the models is to independently extract visual and language feature representations using various feature extractors. Given the representations of the two modalities, a fusion operation is performed at the final layers to fuse the features of the two modalities which is then fed to fully connected layers serving as the prediction layers for the model. Early works used simple fusion techniques such as summation, dot product or concatenation to achieve these fused features.

Observations demonstrated that simple fusion techniques suffer from extracting low semantic correlation between different modalities and also that fusion techniques had a significant impact of the overall accuracy of the model. This meant that a better fusion technique will result in improved model accuracy on unseen data because the model has access to more effective semantic understanding of the data. Motivated by these inferences, more sophisticated fusion methods such as bilinear pooling and Multi-Modal compact bilinear pooling (MCB) were introduced and used in recent works. These methods make use of outer product between the features to get a high-dimensional representation of the feature space of different modalities. Given the high-dimensionality of the representations, compression techniques such as low-rank bilinear pooling using Hadamard product and Multi-modal Factorized Bilinear (MFB) pooling were introduced. MFB computes a fused feature representation with a matrix factorization trick to reduce the number of parameters and improve convergence rate and is also the fusion technique used in this paper [17].

IV. Dataset And Training Process

For the purpose of our experiments and the demonstration of results of the approach highlighted in this paper [17], We are using the MSCOCO dataset available publicly at VQA official website - http://www.visualqa.org. Two versions of the dataset are available -

- VQA v1 (Currently used for training, validation and testing)
  - 82,783 training images, 248,349 training questions and about 2,483,490 annotations for the questions.
  - The answers are of a maximum of 6 words and include both free and MCQ type answers.
  - The test set comprises of test and test-dev dataset splits.
- VQA v2
  - Similar to VQA v1 with 443,757 training questions and about 4,437,570 training annotations.
  - Added complementary training and validation pairs.
  - Abstract image pairs have also been added but for the purpose of this research, those were not used during training or testing.

These two datasets are the most popular ones in the domain of VQA. The question-answer pairs are human annotated and have a predefined *train*, *val* and *test-standard* splits. There is also a *25%* subset of test-standard known as test-dev. All the annotations are categorized into 3 types: *number*, *Yes/No* and *Other.* Along with that, each question can have a MCQ type answer and 10 free-responses.

For the purpose of final model training, the training and validation sets were merged together to form a single train dataset. The test and test-dev splits were used for evaluating the model.

V. Proposed Approach

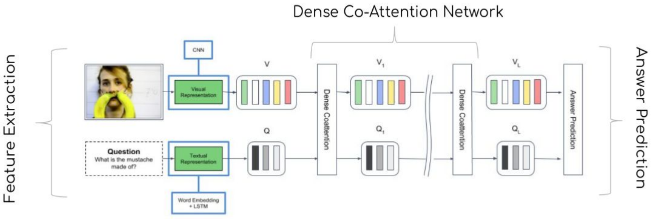
In this section, we will summarise the approach used in this paper [17] along with the architecture of the model and key concepts used in the process.

The architecture basically consists of dense Co-Attention layers bundled on top of each other into a stack, repeatedly fusing visual and language features , initial feature extraction layers and a final prediction layer to predict answers in a multi-label setting.

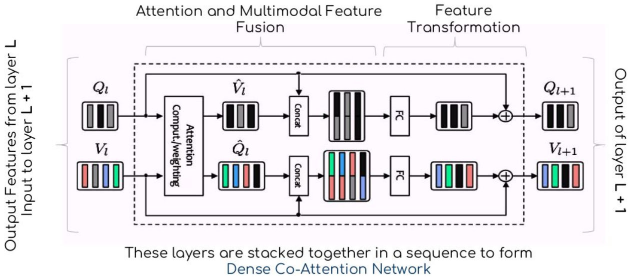
Approach:

- Generate an attention map on every image region for each question word and an attention map on every question word for each image region.
- Redo the process in a hierarchical fashion for phrase level and sentence level for each question.
- With the obtained image and question feature maps perform computation of attended features, concatenation of multimodal representations, and their transformation by a fully connected network with ReLU and a residual connection.
- Finally, using a softmax nonlinearity, generate the score for each possible answer and output the answers with the top scores.

Architecture:



- **Feature Extraction Layer:**
  - Bidirectional LSTM with residual embeddings for language feature extraction.
  - ResNet50 pre trained on ImageNet for Image feature extraction.
  - Multimodal Feature Fusion to map image and question features to lower-dimension fused feature representation.
- **Dense Co-Attention Network:**
  - Parallel Attention
  - A sequence of dense co-attention layers stacked onto each other updating fused features at every level.
- **Answer Prediction Layer:**
  - Fully Connected layer for feature transformation and softmax scores for answer prediction.


Attention and Multimodal Feature Fusion
Feature Transformation
Output Features from layer L
Input to layer L + 1
Q_l
V_l
Attention Comput./weighting
V̂_l
Q̂_l
Concat
Concat
FC
FC
Q_{l+1}
V_{l+1}
Output of layer L + 1
These layers are stacked together in a sequence to form
Dense Co-Attention Network

Computation Involved:

- Attention Computation:

Affinity Matrix: $A_l = V_l^\top W_l Q_l$

$$A_{Q_l} = \text{softmax}(A_l) \qquad A_{V_l} = \text{softmax}(A_l^\top)$$

V_L and Q_L are the Visual and Language feature representations respectively.

- Multi-Modal Feature Fusion:

$$v_{(l+1)t} = \text{ReLU}\left(W_{V_l}\begin{bmatrix} v_{lt} \\ \hat{q}_{lt} \end{bmatrix} + b_{V_l}\right) + v_{lt}$$

For $\forall$ ***v*** ε $\mathbf{V}_\mathbf{L}$ and ***q*** ε $\mathbf{Q}_\mathbf{L}$

- Answer Prediction:

$$(\text{score of answers}) = \sigma\left(\text{MLP}(\begin{bmatrix} s_{Q_L} \\ s_{V_L} \end{bmatrix})\right)$$

Where S_Q and S_V are aggregate representations of Attended Question and Visual representations respectively.

VI. Results

The predictions on the *test-standard* and *test-dev* datasets are included in the submitted ***zip file*** along with project code, report and other artifacts with the name, ***results.json***. The scores from the metric, ***Accuracy***, used to evaluate the model are listed in Table 1 along with similar scores from previous approaches for comparison purposes.

Model	Test-dev (Overall Accuracy)	Test-standard (Overall Accuracy)
VQA team [2]	57.75	58.16
HieCoAtt [1]	61.00	62.10
DAN [18]	64.30	64.20
Strong Baseline [19]	64.50	64.60
MFB [20]	65.90	65.80
DCN [17]	**66.89**	**67.02**

Table 1: Results of the highlighted approach along with published results of others on VQA v1 dataset.

It is clear from Table 1 that the highlighted approach in this paper [17] clearly outperforms previous models in terms of accuracy metric and establishes a new state of the art on the MSCOCO dataset v1 for VQA.

Conclusion

In this report, we showcase a simple and effective approach to Visual Question Answering (VQA) in the form of a novel architecture which is symmetric both, in terms of Visual and Language feature representations. We also highlight related work that has been done and made available to the community of VQA and the techniques that have been adopted and improved in this paper [17]. The network used in this paper [17] named as ***Dense Co-Attention Network*** uses Dense Co-Attention layers with improved feature extraction and Multi-Modal fusion techniques combined with parallel attention mechanisms to establish a new state of the art in VQA community. The experimental results on MSCOCO datasets show the efficiency and effectiveness of this approach.

Acknowledgment

We would like to acknowledge the work done by Nguyen et al. in “Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering” [17] as this is the approach we use in our project work. We would also like to acknowledge the work done by Devi Parikh 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 appreciate and 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.

[17] Nguyen, Takayuki, Duy-Kien and Okatani, “Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering”, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[18] H. Nam, J. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. arXiv preprint arXiv:1611.00471, 2016.

[19] V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162, 2017.

[20] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bi-linear pooling with co-attention learning for visual question answering. in International Conference on Computer Vision (ICCV), 2017.