Visual Question Answering – A challenge

Kowshik Thopalli, Perikumar Javia

Abstract—In this project we would like to come up with a potential solution to the well-defined problem in Computer Vision i.e., Visual Question Answering(VQA). This task of free-form and open-ended VQA is proposed as an open challenge in Late 2015 through Virginia University. Since then this challenge has attracted great attention. The task is defined as to provide an accurate natural language answer when given an image and a natural language question about the image. The dataset is provided through their website. The dataset is an extensive collection of about ~0.25M images (Real and Abstract), ~0.76M questions and ~10M answers. The present leaderboard is topped with an overall accuracy over four different categories by Jongryul.lee et al (https://competitions.codalab.org/competitions/6961#results) with 67.64%. We would like to apply techniques learned in class/ through our own study and try to beat/come close to the state-of-the art results.

Index Terms— Deep Learning, Visual Question Answering, Computer vision, Natural Language Processing

1 Introduction

Object recognition is a fundamental part of perception in robotics and computer vision. However, with new devices pouring into the market and gaining popularity among the specially-abled like Amazon's Alexa, there is a significant interest in trying to solve multi-discipline AI challenges. These devices mainly respond to speeches with high accuracy. Taking motivation from this we want to develop some state of art deep learning techniques that answers questions from the images.

Visual Question Answering (VQA) is a task that proposes to connect Natural Language Processing (NLP) and Computer Vision (CV) and stretch their limits. Computer Vision is that field which concerns itself with acquiring, processing and understanding images while NLP studies the methods of Human Computer Interaction through natural language. Combining these two fields through the proposed task of VQA can be potentially utilized in various applications like answering questions such as "who is standing at the door"? etc. This is made possible by studying the given scene and running queries given by user on the image through a representation of the image using NLP techniques. Potential applications like creating a new paradigm shift in the way we interact with the computers, bi-directional image-sentence retrieval, etc. has thus made this problem a hot research area in both academia and industry. Both the fields CV and NLP have seen significant developments in their respective goals through deep learning in the recent past and the growth trend is only exponential because of the large amount of data available in both the contexts. The first steps in an effort to marry these two fields can be seen in attempts like automatic video-captioning or image description where efficient methods have been developed to learn or map inputs from images and text to a combined space and thus making inferences from that joint space. This was achieved by a combination of Convolutional Neural Networks (CNNs) that are trained on object recognition with word embeddings[5] trained on a very large text corpus.

VQA is a difficult task because in many cases it requires information that is not there in the image. The range of additional information required can be anything from simple common sense to expert-level knowledge on the objects present in the image. The below example summarizes the complexity of the problem and gives a visual understanding of the description above

An example of VQA is shown in below figure

Figure 1:- A sample VQA task. The picture is sourced from WWW.Visualqa.org

As mentioned earlier due to its significant complexity and potential uses this problem has gained much traction in academia/ industry. Thus there have been a number of datasets proposed for this task. Each dataset will at least have three things- the image, a question

about the image and the correct answer for the question. There are other variants such as having multiple choice questions or Fill in the Blank questions.

We wish to implement/compare our algorithms in one of the most used datasets- the VQA dataset[1] that consists of ~0.25M images which includes ~0.2M MS COCO dataset[b] and ~0.05M abstract scene images. Three questions were collected for each image and each question was answered by ten subjects along with their confidence. The dataset thus contains over 10M answers for 760K questions. This dataset also has two different kinds of questions one being open ended and free form questions and the other being multiple-choice questions.

This is our proposed timeline and project milestones defined for every 2-2.5 weeks

We have identified these sub-tasks to solve this problem.

They are:

- Extensive Literature Survey on Visual Question Answering Problem with emphasis on gaining comprehensive knowledge on CNN's[2] for Object recognition, RNNs[3] and LSTM for language representation etc.(Kowshik)
- Dynamic Memory Networks(DMN's):- Question answering is a well-studied problem in the context of NLP. There were significant improvement in results of the complex textual QA problem using DMN's. Complex textual QA as in creating a need to answering general textual QA by understanding the entire question , its context and pay attention to relevant details to answer a question. This is made possible by modelling interaction between multiple parts of data over many passes. However these networks which have gained good popularity among NLP community have not been explored much by CV community except for [4] . This is the only paper that uses DMN's for VQA.

  Due to its proven record in answering complex problems by using a memory augmented modular architecture we want to explore their application to this problem.(Perikumar)

So our **first milestone** will be to do sub-task 1, develop and explore DMN's for VQA.

Most of the present State of the art results are obtained through a combination of CNNs and RNNs/LSTM.
We wish to improve the results by exploring and implementing newer architectures and other methods learned in the class. (Kowshik and Perikumar)
This will be our milestone #2.
During a brain storming session with Ragav, we understood the importance of answering the vital question- "Are these networks not overfit?". We wish to come up with some experiments to either prove/ disprove that these networks overfit (Kowshik). Also, we would like to explore the possibility of taking outputs not from the last layer but from (n-k)th layers where n is the number of layer and k= 1,2..so on and see its prospective uses and its effects on results(Perikumar). **This will be our milestone #3.**

The fourth milestone would be to consolidate our experiments and come up with the best possible architecture and its parameters such that our network results in an accuracy that puts us in top five of the leaderboard and submit the report (Kowshik and Perikumar)
Note:- Names in parenthesis represent the person responsible

References:

[1]Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, VQA: visual question answering, CoRR abs/1505.00468 (2015).

[2]Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia, ABCCNN: an attention based convolutional neural network for visual question answering, CoRR abs/1511.05960 (2015)

[3]Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz, Ask your neurons: A neural-based approach to answering questions about images, CoRR abs/1505.01121 (2015).

[4]Caiming Xiong, Stephen Merity, and Richard Socher, Dynamic memory networks for visual and textual question answering, CoRR abs/1603.01417 (2016).

[5]Mengye Ren, Ryan Kiros, and Richard S. Zemel, Image question answering: A visual semantic embedding model and a new dataset, CoRR abs/1505.02074 (2015).

Dataset:-
[a] http://www.visualqa.org/
[b] http://www.cs.toronto.edu/~mren/imageqa/data/cocoqa