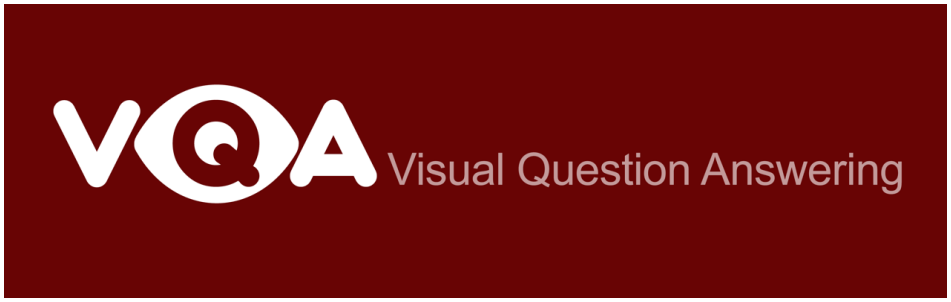
SYNOPSIS

TITLE: Visual Question Answering through Modal Dialogue

TEAM MEMBERS: Somin Wadhwa, Raunaq Jain

PRINCIPAL INVESTIGATOR: Mr. Neeraj Garg

Department of Computer Science & Engineering
Maharaja Agrasen Institute of Technology, Sec -22, Rohini



Resources	Details
Language (Environment)	Python (3.6.x)
Area/Domain	Deep Learning
Sub-Domains	Image Processing, NLP
Reference	visualqa.org
External Resources Used	MSCOCO Image Set, Word Embeddings
Primary Framework	TensorFlow

ABOUT:

Visual Question Answering is a task that has emerged in the last few years and has been get- ting a lot of attention from the machine learning community. The task typically involves showing an image to a computer and asking a question about that image which the computer must answer. The answer could be in any of the following forms: a word, a phrase, a yes/no answer, choosing out of several possible answers, or a fill in the blank answer.

Visual question answering is an important and appealing task because it combines the fields of computer vision and natural language processing. Computer vision techniques must be used to understand the image and NLP techniques must be used to understand the question. Moreover, both must be combined to effectively answer the question in context of the image. This is challenging because historically both these fields have used distinct methods and models to solve their respective tasks.

MSCOCO-QA:

Among many, COCO-QA dataset is another dataset based on MS-COCO. Both questions and answers are generated automatically using image captions from MS-COCO and broadly belong to four categories: Object, Number, Colour and Location. There is one question per image and answers are single-word. The dataset contains a total of 123,287 images. Evaluation is done using either accuracy or WUPS score.

EXISTING APPROACHES (DL + NONDL BASED):

(Kafle and Kanan, 2016) propose a Bayesian framework for VQA in which they predict the an- swer type for a question and use this to gener- ate the answer. The possible answer types vary across the datasets they consider. For instance, for COCO-QA they consider four answer types: ob- ject, color, counting, and location.

(Ma et al., 2015) propose a CNN-only model that we refer to here as Full-CNN. They use three dif- ferent CNNs: an image CNN to encode the image, a question CNN to encode the question, and a join CNN to combine the image and question encoding together and produce a joint representation.

Attention based techniques are some of the most popular techniques that are being used across many tasks like machine translation (Bahdanau et al., 2014), image captioning (Xu et al., 2015) etc. For the VQA task, attention models involve focusing on important parts of the image, question or both in order to effectively give an answer.

PROPOSED WORK:

As has been the trend in recent years, deep learn- ing models outperform earlier graphical model based approaches across all VQA datasets. How- ever, it is interesting to note that the Answer Type Prediction (ATP) model performs better than the non-attention models, which proves that simply introducing convolutional and/or recurrent neural networks is not enough: identifying parts of the image that are relevant in a principled manner is important. ATP is even competitive with or better

than some attention models like Where to Look (WTL) and Stacked Attention Networks (SAN). As we have seen, novel ways of computing attention continue to improve performance on this task. This has been seen in the textual question answering task as well (Xiong et al., 2016) (Seo et al., 2016), so more recent models from that space can be used to guide VQA models. A study providing an estimated upper bound on performance for the various VQA datasets would be very valuable as well to get an idea for the scope of possible improvement, especially for COCO-QA which is automatically generated. Finally, most VQA tasks treat answering as a classification task. Only the VQA dataset allows for answer generation in a limited manner. It would be interesting to explore answering as a generation task more deeply, but dataset collection and effective evaluation methodologies for this remain an open question.

Signature of Mentor
(Mr. Neeraj Garg)