CIS 522 – Final Project – Technical Report

Dying Neurons

April 2022

Team Members:

- William C Francis; willcf; Email: willcf@seas.upenn.edu
- Rashmi Phadnis; rphadnis; Email: rphadnis@seas.upenn.edu
- Vandana Miglani; vmiglani; Email: vmiglani@seas.upenn.edu

Abstract

Visual Question Answering (VQA) refers to the concept of taking in an image and a question as inputs and producing an answer as an output. The question fed in has to be relevant to the image provided and the model has to use the image to generate the correct answer to the question asked. A well-functioning VQA system depends on the questions that are asked being answerable and if they are answerable then the answer being contained in the image itself. Since a lot depends on the multimodal nature of the inputs to the system - both images and natural language - as well as the nature of the question, we aim to use a multimodal architecture for our baseline which will consist of a CNN and an LSTM. This architecture is then improved by the addition of a stacked attention layer in the advanced DL model. The baseline architecture produced an accuracy of 35.36% and the Stacked Attention Network (SAN) outperformed the baseline, producing an accuracy of 54.82%. In particular, the SANs produced visually reasonable and logically sound answers among its top-5 predictions. A key outcome in this paper is how the stacked attention layers help in multi-step reasoning and focusing on relevant portions of the image to detect the answer.

1 Introduction

The recent advancements in computer vision and natural language processing (NLP) have brought several research topics in AI to limelight. Visual Question Answering (VQA) is one such active research area that has witnessed a renewed excitement in the past couple of years [1], [2]. Part of this excitement stems from the fact that VQA is an interesting application in the field of multi-discipline Artificial Intelligence (AI) research problems particularly in

Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation Reasoning. In this project, we introduce the task of open-ended VQA. The system considers an image and a free-form question about the image together and outputs an answer. The main difference between a normal image captioning task and VQA is that a specific question about an image requires the model to identify certain objects and their characteristics (activity of a person, color of an object). What attracted us to this problem statement was this very multi-modal nature.

Figure 1: Fig. 1. Examples of open-ended questions

VQA can be used in a variety of applications, which include aiding visually-impaired users in understanding their surroundings ("What temperature is this oven set to?"), analysts in making decisions based on large quantities of surveillance data ("What kind of car did the man in the red shirt drive away in?"), and interacting with a robot ("Is my laptop in my bedroom upstairs?"). This project has the potential to fundamentally improve the way visually-impaired users live their daily lives, and revolutionize how society at large interacts with visual data.

To answer an open-ended question an AI model requires to develop a vast array of capabilities such as language processing, object detection, activity recognition, knowledge-based reasoning, and commonsense reasoning. VQA is also problematic when it comes to evaluation of the results. While the answer to most questions would be a simple yes or no, the right answer is sometimes subjective or even not discernible from the image (e.g. "is this person crying?", "how fast is the train moving?") and thus makes it difficult to evaluate. VQA is also a highly computationally expensive task which is often tackled by neural networks with millions of parameters. In our advanced DL approach, we implemented a Stacked Attention Network (SAN) that introduces an attention mechanism. Attention mechanism has been successfully applied in image captioning [8] and machine translation [9], showing promising results.

Stacked attention networks focus on regions of the image that are relevant to the question posed through multi-step reasoning. The SAN model locates the objects in the image to the concept referred in the question and through layers of stacked attention, it rules out irrelevant regions to finally pinpoint the regions

that are most indicative to answer the question. The basic VQA model consists of two major components, namely, the image model and the question model. The image model uses a CNN to extract high level features from the images and the question model uses an LSTM to extract the semantic vector of the question. The SAN model consists of a stacked attention component in addition to the these two components. The stacked attention component uses the question embedding to query the image embedding in the first attention layer. The question embedding and the retrieved image embeddings are combined to form a refined query for the second attention layer. In this project, we explore this novel approach and compare it with the state of the art VQA models.

2 Related Work

The task of Image Captioning is highly related to Visual Question Answering. In [7], the proposed model extracts high level image feature vectors from GoogleNet, which are then fed into a LSTM. [8] extended this approach and used an attention mechanism in the caption generation process. While there are similarities between the tasks of Captioning and VQA, it is important to note that in Image Question answering, the question is supplied and the task is to infer the answer based upon the understanding of the image.

Several papers have explored the intersection of Vision and Language for the purposes of VQA. Unlike some of the earlier work in which settings were fairly restricted and the approaches only considered answering questions that came from closed world settings, our proposed task aims to answer free-form and open-ended questions.

We discuss other research work most pertinent to our proposed task of free-form and open-ended question answering. In [3], the authors propose combining an LSTM for question and a CNN for the image, to finally generate an answer. In their model, at each time step, the LSTM question representation is conditioned on the CNN image features. In the final step, the last LSTM hidden state is used to decode the answer.

This is different from the approach proposed in [4] which forms the baseline architecture of our project. The baseline approach proposed in [4] uses the technique of "late fusion" - i.e. image and question embeddings are computed independently and fused via element-wise multiplication. These fused embeddings are then passed through a Multi Layer Perceptron to generate a probability distribution over the answer classes.

Other methods include the encoder-decoder architecture as presented in paper [5]. The proposed approach used an LSTM to encode questions and images and then, used another LSTM to decode answers. Image features are fed to every LSTM cell. Another approach [6] used CNN for question modelling and convolutional operators to combine question vectors and image feature vectors.

3 Dataset and Features

The project utilizes the latest version of a standard Visual Question Answering dataset - Visual VQA dataset v2.0. This release consists of 82,783 MS COCO training images, 40,504 MS COCO validation images and 81,434 MS COCO testing images (images are obtained from the MS COCO website). Along with the images, this release also has 443,757 questions for training, 214,354 questions for validation and 447,793 questions for testing. This dataset comprises of 4,437,570 answers for training and 2,143,540 answers for validation (approximately 10 responses per question).

The questions and answers within the dataset are collected by the authors of the VQA paper which we use as a baseline. The authors chose MS COCO images to elicit a more interesting set of questions and answers since the MS COCO images have a lot of diversity and richness. The authors used a user interface called Amazon Mechanical Turk to collect the questions and the answers. To make sure that the users who frame the questions create a set of interesting and complex questions, they told the users that the questions have to fool a smart robot. The authors did this to make sure there weren't any simple yes/no questions or questions which could be answered easily and weren't dependent on the image such as "What is the color of the banana?". In all, they collect 3 unique questions per image from each user.

For the answers, they use the same user interface. They collect 10 answers from each user and ask for short phrases or single words as answers. The collected answers for a particular question might have multiple possible correct answers such as "red", "marroon", "dark orange". Even for yes/no questions, both the 'yes' or 'no' might be correct. The authors choose the most common answer as the correct answer.

From our analysis of the dataset, most questions comprise of around 4 to 10 words. Most of the answers (89.32%) are one word answers and almost all answers are at most 3 word answers.

For our project, due to limited compute, we restrict the dataset to 80,000 images for train and test and 40,000 images for validation. The vocabulary for the answers was generated by choosing the 1000 most common answers from the dataset.

4 Methodology

4.1 Non-DL Benchmark

The problem that we have chosen has a complex multi-modal nature which requires dependence on both the question asked (NLP) and the image provided (Computer Vision). In order to formulate a non-DL approach, we decided to treat it as only an NLP problem. Even so, it is difficult to generate an answer from only the question and make sure that the answer is relevant to the image provided. We first tried a random selection method as our non-DL approach.

From the answers provided in the dataset, we generate a vocabulary of answers in which we choose the 1000 most common answers. From this vocabulary, we randomly choose a word as the correct answer. This gave very bad results. We then tried a method wherein we introduced a bias towards the most popular answer - 'yes'. The model always gives yes for any question.

4.2 Baseline DL Approach

In our baseline approach, we implemented a 2-channel vision + language model, which fused embeddings from both these channels via point-wise multiplication. These fused embeddings were then passed through fully-connected layers. To obtain the answer, we generated a probability distribution over the range of possible answer classes.

The different components of architecture are outlined as follows:

1. Image Channel This channel provides l_2 normalized activations from the last hidden layer of VGGNet for the images. The image embeddings are transformed to 1024-dimensions by using a fully connected layer and tanh non-linearity.

2. Question Channel For the embedding of the question, an LSTM with 2 hidden layers is used. The 2048-dimension embedding obtained from the LSTM is a concatenation of the last cell state and last hidden state representation. To transform the 2048-dimension encoding to 1024-dimensions, we used a fully connected layer and tanh non-linearity.

3. Multi-Layer Perceptron - The 1024-dimension image and question embeddings are fused together via point-wise multiplication to obtain a single embedding. This fused embedding is then passed through fully-connected layers. The Multi-Layer Perceptron unit consists of 2 linear layers, each with 1000 hidden units. Each of the hidden layers was also followed by a dropout layer (set to 0.5) and tanh nonlinearity. The MLP unit is then followed by a softmax layer to obtain a probabiltiy distribution over the 1000 possible words from the answer vocabulary.

The model was trained end-to-end using Cross Entropy loss. During training, the parameters from VGGNet were frozen to the pretrained weights obtained from the ImageNet Classification. The model was trained for 40 epochs using a Step Learning Rate scheduler where the initial learning rate was 0.001. This was decayed with a step size of 10 and a decay rate of 0.1. We used an Adam optimizer for the training.

The Figure 2 below illustrates our baseline architecture.

4.3 Advanced DL Approach

For the advanced DL approach, we chose to focus on the way the two embeddings from the Image and the Question channel are fused together. In our baseline DL

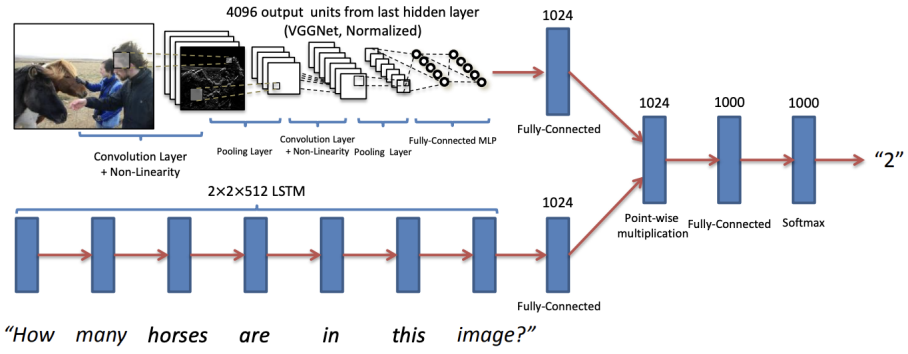

Figure 2: Baseline DL Architecture

approach, this is done using a simple element-wise multiplication. We replaced this with a Stacked Attention module. Thus, the underlying architecture of the advanced DL approach mirrors the baseline DL approach with changes made to the Image channel and the way the two embeddings are fused together. The different components of architecture are outlined as follows:

1. Image Channel Instead of using the last hidden layer of a pre-trained VGGNet, we use the last pooling layer this time. This gives us an embedding of 14x14x512. An intuitive way of thinking about this is that the image was divided into 196 regions (14*14) and each region was encoded in an embedding vector of size 512. These embedding vectors are sent through a linear layer to reshape them to be the same size as the Question embedding obtained from the Question Channel.

2. Question Channel The Question Channel remains the same as the baseline DL approach.

3. Attention Layer (Fusion) The underlying intuition of using the attention mechanism is to query the 196 regions with the question embedding multiple times to find the regions in the image relevant to the question. Using multiple attention layers to perform the fusion leads to a more detailed attention distribution which helps in locating the regions that are relevant to the potential answers. We use a two layer attention mechanism. In the first visual attention layer, the question vector is used to query the image vectors. In the second layer, we concatenate the question vector and the first layer output to form a refined query vector to query the image vectors again in the second attention layer. The output is the attention distribution over these 14*14 or 196 regions. This attention distribution is used to weight the original image feature vector. Finally, we combine

the weighted image features with the last query vector to give a single embedding vector. The figure below shows how the focus is narrowed down on a relevant region after each of the 2 attention layers for the question 'What is sitting in the basket on the bicycle?'.

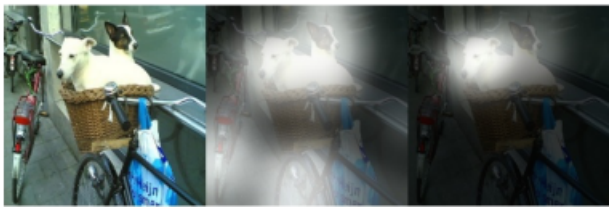

Figure 3: Visualization of Attention for First and Second Layers

4. Multi-Layer Perceptron The fused embedding is then passed through fully-connected layers with the same architecture as the baseline DL model.

The model was trained using a Cross Entropy Loss since we need to look at the difference between probability distributions. We trained it for 40 epochs using a Step Learning Rate scheduler where the initial learning rate was 0.001 which was decayed with a step size of 10 and a decay rate of 0.1. We used an Adam optimizer for the training.

The figure below illustrates the advanced architecture:

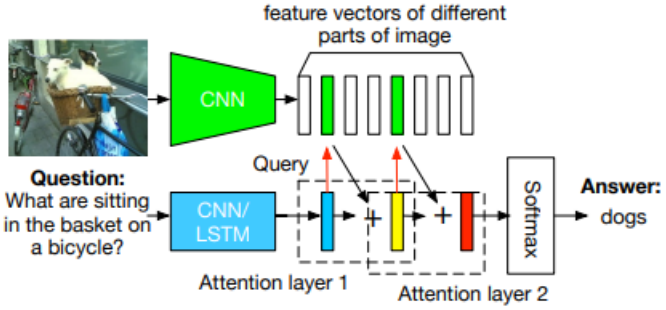

Figure 4: Advanced DL Architecture

5 Results

The following table shows the accuracy of the three approaches. As expected, our non-DL approach performs poorly, the basic VQA model which does not utilize an attention model performs better and the advanced DL model performs the best. The accuracy was calculated based on whether the model was able to predict the correct answer or not. The loss function that was minimized was the Cross Entropy Loss.

The following table shows the results obtained for a few test images and the questions from the baseline DL and the advanced DL model. The predicted probability for the top 5 answers is shown.

Methods	Open-Ended Questions			
	All	Yes/No	Numbers	Other
Prior-Yes Answers (Non-DL Approach)	19.10%	71.84%	0.00%	01.25%
Baseline Deep Learning Approach	35.36%	82.32%	23.20%	28.61%
Stacked Attention Network Approach	54.82%	80.50%	33.45%	47.68%

Table 1: Accuracies of the three approaches on different types of questions

What is the shape of the clock?

What kind of room is in the image?

What animal is it?

Baseline Approach Results	**Predicted - Probability** '<unk>' - 0.3413 'yes' - 0.2894 'no' - 0.0445 'green' - 0.0397 'grey' - 0.0353	**Predicted - Probability** '3' - 0.3071 'no' - 0.2403 'cloudy' - 0.0424 'green' - 0.0289 'mirror' - 0.0284	**Predicted - Probability** 'bird' - 0.2248 'yes' - 0.1613 'yellow' - 0.0434 'blue' - 0.0434 'green' - 0.0367
SAN Approach Results	**Predicted - Probability** 'circle' - 0.2575 'square' - 0.1877 'round' - 0.1162 'summer' - 0.0804 'fall' - 0.0803	**Predicted - Probability** 'urban' - 0.2826 'city' - 0.2759 'indoor' - 0.1654 'london' - 0.0564 'new york' - 0.0493	**Predicted - Probability** 'crow' - 0.3461 'bird' - 0.1851 'water' - 0.1566 '<unk>' - 0.0498 'boat' - 0.0375

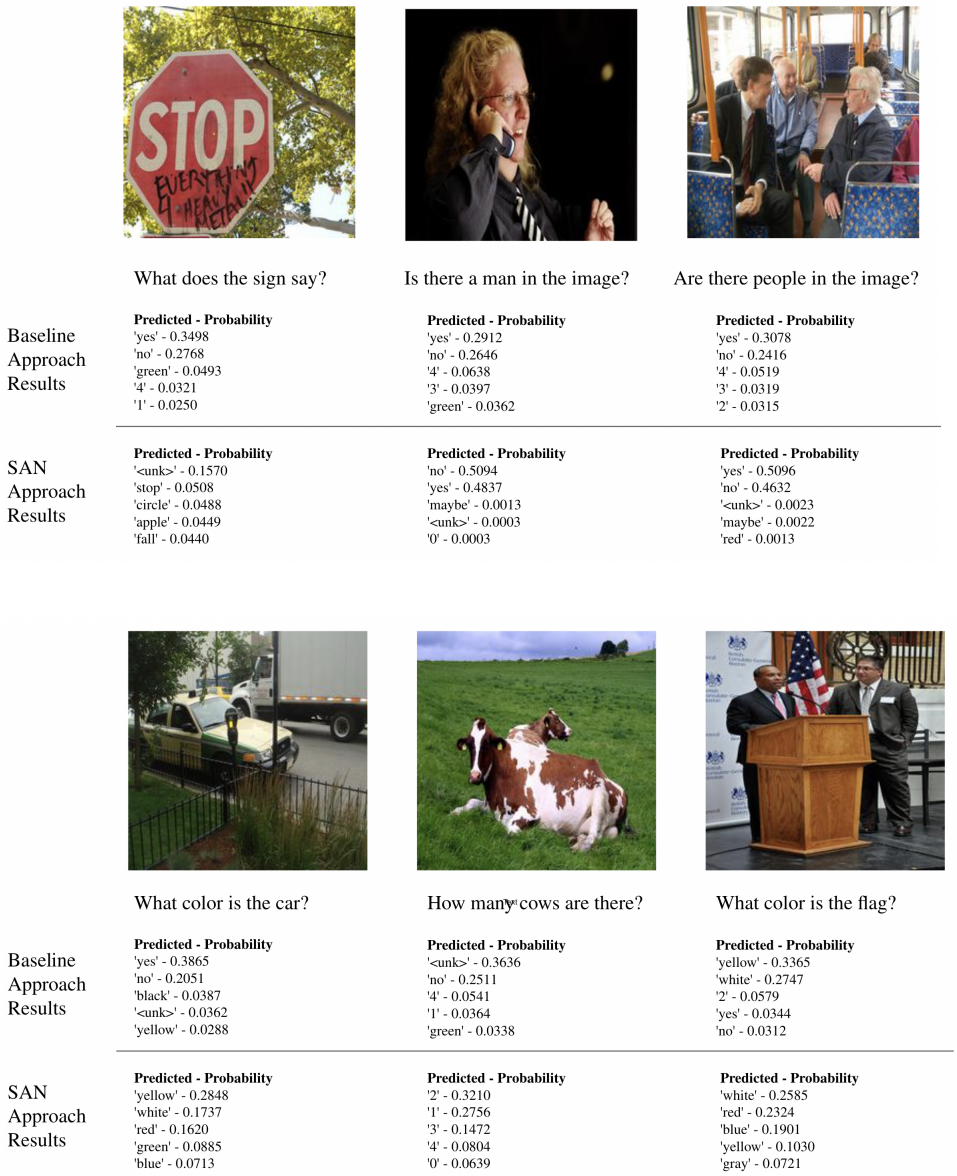

Figure 5: Comparison of answers generated by the baseline approach and the Stacked Attention Network for different questions

6 Discussion

While the Non-DL approach wasn't able to effectively capture the image and question to get the relevant answer, both the deep learning approaches worked

well. Within the deep learning models, the stacked attention model consistently outperformed the basic VQA model, particularly for questions which required multi-step reasoning to get to the answer. This can be explained by the fact that the attention mechanism helped the model to focus on the relevant regions.

6.1 Findings

The basic VQA model seems to have overfit on the answer dataset, learning that 'yes'/'no' will usually be the right answer and learning which color or number will usually be right (most commonly occurring number and color in answer dataset). The stacked VQA model results seem to display more reasoning to get to the correct answer from the model's side. For example, the shape of the clock in Figure 5 was accurately identified as 'circle' and 'round' by the stacked attention model but the baseline model failed to predict even a single response indicating a shape. Whenever the baseline model is unsure of its answers, it picked up the most common words in the vocabulary such as 'yes', 'no', 'green' and '<unk>' as observed from the image of the clock, birds, and the stop sign. A possible hypothesis is that the baseline model has not trained well enough to be generalizing well on the test images. On the other hand, the SAN model requires much lesser training to produce similar results by focusing on the relevant regions in the images, thus creating much better results for the same amount of training.

The SAN model shows surprisingly good results on an indoor image of the room, even predicting that it could be an urban environment and relating it to urban cities such as New York and London. Although the baseline model includes 'mirror' in its predictions, it predicts this with a very low confidence. We can also observe that the 'yes' and 'no' answers are almost always predicted with high confidence values often ranging from 0.4 to 0.5. This shows how well the model has learned to answer 'yes' and 'no' questions.

Another interesting observation can be drawn from the image where the color of a flag is queried. The baseline model overemphasizes on the word 'color' and produces results which includes the colors of background objects such as the white color of the wall, and the yellow color in the background. The answer of '2' might be attributed from the presence of two people in the image. On the other hand, the SAN model identifies all the colors of the flag correctly: white, red, and blue as the top predictions.

Our methodology of creating a concise vocabulary before predicting the answers seems to have worked well as compared to predicting the answers from a larger vocabulary. The models learned faster as they need only consider the words in the reduced vocabulary. In addition, this gives the models a higher success rate by pure luck even if the model gives a wrong prediction as there is a high probability the the answer would correlate to the question given it is one of the most common answers.

The results show how effective VQA can be used in multiple applications that could aid human lives in a multitude of domains. For example, the identification of the stop sign from a language prompt such as "what does the sign say?". This

could be life changing to people suffering from vision impairment [10] as they could get information from the surroundings, almost like an AI walking stick. VQA can also be integrated into AI personal assistants like Siri and Alexa to grab relevant information quickly and effectively. With VQA, an AI personal assistant would be able to sort through your photos and bring you images that has a beach in the background. Or even asking your Siri upstairs whether your laptop is on the table could leverage VQA. VQA could also be used to sift through large quantity of information in an instant. The police could use it to track down a criminal given his/her identification details from the thousands of surveillance cameras across the nation with just a query. With the advent of technology that deal with massive visual data, an advanced VQA model such as the SAN model could save time and energy, while effectively carrying out its task.

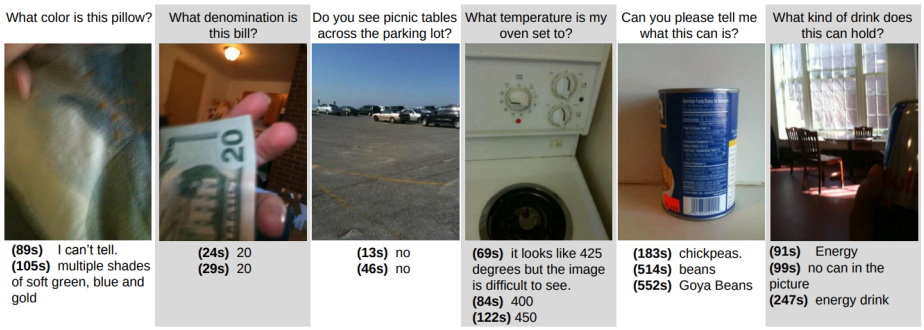

Figure 6: VQA results [10] demonstrating its application for the visually-impaired

6.2 Limitations and Ethical Considerations

Firstly, we were limited by both computing power and time, thus we could only train on a limited subset of the VQA dataset. Access to sufficient computing resources will let us train on the entire dataset and thus assess whether the basic VQA model gives comparable performance to the attention VQA model.

If we assume that we have access to adequate compute and can train for longer, the concept of VQA is limited by the kind of data we train on. In the dataset that we used, most questions required a yes/no answer making the words 'yes' and 'no' the most common answers. We can see from the results that the basic DL model learned that giving a yes or no would usually give it a correct answer.

Another thing to consider is how the dataset was curated. Since the questions and answers were created by humans, it is very possible that their biases carried through when they asked questions or answered them. The application for Visual Question Answering can be found in any instances where humans have to elicit situationally-relevant information from visual data or where humans and machines must collaborate to extract information from pictures. If

we take the example of security analysts using this on surveillance data, there is a very high chance that the model's answers will reflect the biases of the people who provided the dataset for questions such as 'Did the man steal from the store?'.

The last thing is the variety of the images curated for the task of VQA. Our current dataset used MS COCO images which has a high variety of images. But these images might not be relevant if the application of the VQA system is for security analysts where surveillance images and data might make more sense. It might be beneficial to thoughtfully curate the images so that they are most relevant to the application at hand.

6.3 Future Research Directions

In our stacked attention model, we considered a different approach to fusing the image and question embeddings, rather than the point-wise multiplication in our baseline approach. For future work, we would like to explore other ways to combine the image and question embeddings into a single embedding. One such way could be to fuse the embeddings using autoencoders. Variational Autoencoders are useful in learning latent space representations and we want to explore if their addition would be helpful in learning image and question embeddings.

Another direction we would like to explore is to consider other CNN architectures for generating the image embeddings. At present both our approaches use pre-trained weights from VGGNets. We would like to study how the results change when we use a different architecture such as Regional Convolutional Neural Networks (R-CNNs) instead. R-CNNs are popular for object detection tasks and are based on the key concept of 'region proposals'. Region proposals are used to localise objects within an image. Given that the responses to questions are usually based on the properties of some object within an image, we believe that using R-CNNs could yield better results.

We would also like to conduct an ablation study to explore the effects of increasing/decreasing the number of attention layers in the stacked attention network. Another area we would like to analyse further is determining if images are really necessary for answering some questions. For instance, it may be possible to determine answers to some questions like 'What is the color of the grass' or 'What is the color of the banana the monkey is eating' based on questions alone, without necessarily accounting for the image. This would help us in understanding which portion of the questions can be answered by commonsense alone and how the visual information encoded in images is critical to the task of Visual Question Answering.

7 Conclusions

In this project, we explored 2 different approaches to the Visual Question Answering task - a baseline approach where we fused image and question embed-

dings via point-wise multiplication and an advanced DL approach, in which we used stacked attention layers, following from the intuition that question answering is generally a multi-step reasoning process. Both the architectures produced reasonable answers and were able to generate logical results to the posed questions. We observed that Stacked Attention Networks produced better results. Using multi-step reasoning in the stacked attention layers, SANs focused on the relevant details in the image to detect an answer and yielded good results.

References

1. X. Chen and C. L. Zitnick. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In CVPR, 2015.
2. A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In CVPR, 2015.
3. M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neuralbased approach to answering questions about images. In ICCV, 2015.
4. 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.
5. 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.
6. L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network
7. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator.
8. 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.
9. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
10. J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. VizWiz: Nearly Realtime Answers to Visual Questions. In User Interface Software and Technology, 2010.