Visual Question Answering using Attention
ECE 285 - Intro to Visual Learning

Ajit Deshpande
ECE Department
A59003350

Parthasarathi Kumar
ECE Department
A59003519

Abstract

Visual Question Answering (VQA) is a challenging problem at the intersection computer vision and natural language processing where a networks is tasked at answering a text based question based on a input RGB image. We use a combination of image features extracted from a CNN based image encoder and a LSTM based text encoder to fuse the features which are finally used to predict an answer. The model is trained on the VQA 2.0 data-set [3]. Extensive experiments are conducted on the resultant model to analyse how the model learns to focus on different image regions.

1 Introduction

Visual question answering (VQA) is a problem at the intersection of computer vision and natural language processing. It is a step towards human level intelligence where the model has to reason from multiple modalities to successfully answer the question. It has many interesting applications like helping the visually impaired to interact with the world of images via text. It also provides an opportunity to understand how these system interact with the 2 modalities to generate the answer. To elaborate on this, a image can be associated with multiple question and answer pairs. Each pair requires the system to pay attention to a specific region of the image.

A typical VQA system can be broken down into multiple parts as indicated in Figure 1. They tend to have components to extract information from the images and question given via text independently. Then there is a component to fuse the 2 sources of information in an effective way to answer the given question and image pair. Lastly the fused information is used to generate the answer. Like most computer vision tasks, the feature extraction is done via a CNN based backbone while for the text-feature encoding, LSTMs and GRUs are popular options.

The main contribution of this work is the fusion strategy to effectively utilize the text and image features. It is a simple but at the same time effective way of leveraging both modalities to produce the best possible result as described in detail in Section 3.3.

2 Related Work

Visual Question Answering is a subset of computer vision tasks involving generation of answers of questions for a given image. This task was initially proposed in [4] and [5]. The major techniques used in [5] was using a query based Turing test to generate words as answers corresponding to questions. The questions are restricted to yes/no answers which is restricting of the question types.

In [4], the major contributions were to propose a generalized VQA methodology and also created a VQA dataset (VQA 1.0 and later VQA 2.0) which used the MSCOCO dataset and provided annotations for the questions and answers.

Other similar tasks such as image and video caption generation, image tagging and natural language processing problems for text based question and answering have been addressed by works such as

[6] which tackles the task of generating similar images using queries, [7] which generates image descriptions using conditional random fields, [8] which again generated image descriptions but using multimodal recurrent neural networks and [9] that describes in detail video explanation techniques using few shot learning.

3 Method

The two inputs to a VQA framework are the image and the question. We can understand the visual question answering task as a classification task where our model chooses an answer from a fixed number of possible answers. Hence, given N possible answers, our model generates a probability vector for each class and the class with the maximum probability is chosen.

$$a_{pred} = \arg\max_{a} P(a \mid img, ques) \tag{1}$$

The model is shown in figure 1. We use the image and question as elements of our data to train our model. The main methodology of a general VQA task involves four major steps

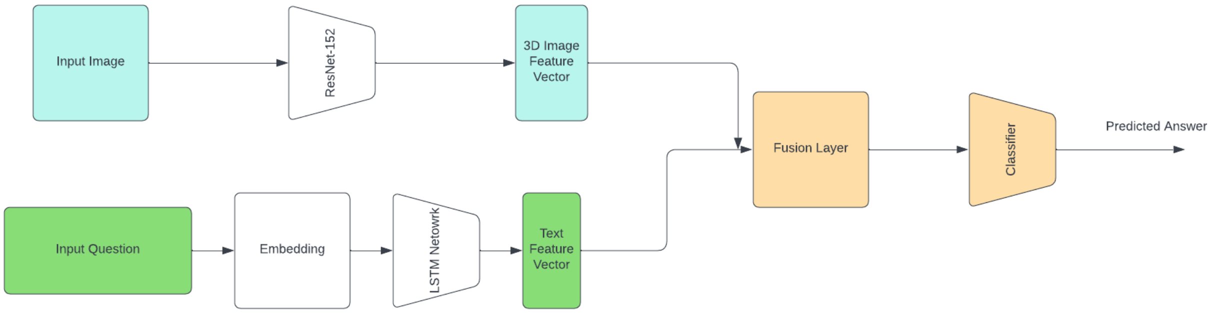


Figure 1: Model Diagram

3.1 Image Featurization

To generate features from the input image, we use a Convolutional Neural Network to generate a feature vector from the same. In our model we use a pretrained ResNet-152 model to extract these features from the training image as a pretraining task.

3.2 Question Featurization

This stage entails converting natural language questions into their embeddings for further processing. We create a vocabulary of the words in the questions and encode each question to a vector of numbers using this encoding. Next we generate embedding vectors for each question which are then used in our fusion layer for further processing. The next step of this stage is to pass this embedding thorugh an LSTM network to generate the feature vector corresponding to each question.

3.3 Fusion Layer

This stage designs ways of combining image features and the question features to enhance algorithmic understanding. We first evaluate the weighted average of features detected in the image at each spatial location and concatenate it with the word embeddings to obtain the joint representations. This process is usually referred to as generation of the Attention layer which helps in the task of VQA. The process of concatenation done is different in different methods which will be explained in Sections 3.3.1 -3.3.2.

3.3.1 Tiled Attention

In this method, we construct the fusion or the attention layer by converting each embedding of the question into a tile equivalent to the size of the image feature generated by the ResNet-152 network and add the two vectors together to generate the tiled attention feature vector which is passed through the classifier network to generate a possible answer. We can summarize the Tiled Attention Method using the diagram in figure 2

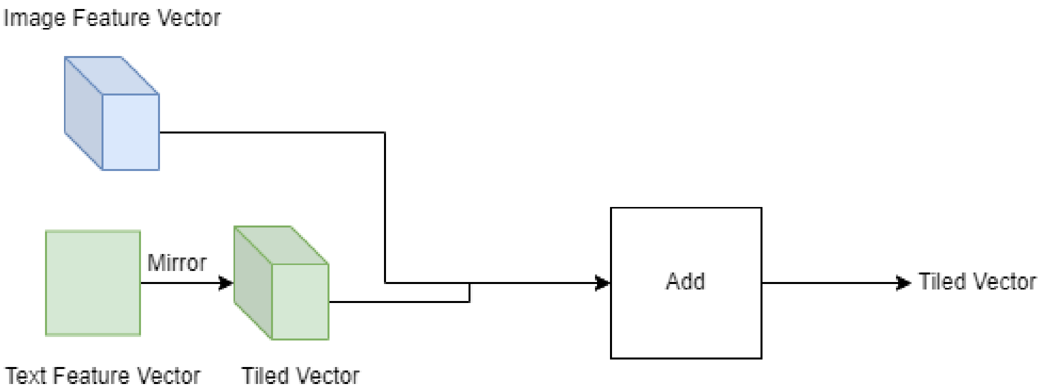


Figure 2: Tiled Attention Block

3.3.2 Stacked Attention

The stacked attention block mimics a attention module. The embedding from the input question is used to query the image features to retrieve image vectors. This is combined with the previous query i.e. the original question to generate a refined query. The process is repeated and the sum of the final query and retrieved image vector is used to predict the final answer via a MLP. The basic idea is to use multiple levels of attention to focus on the specific regions of the image that are relevant to the image.

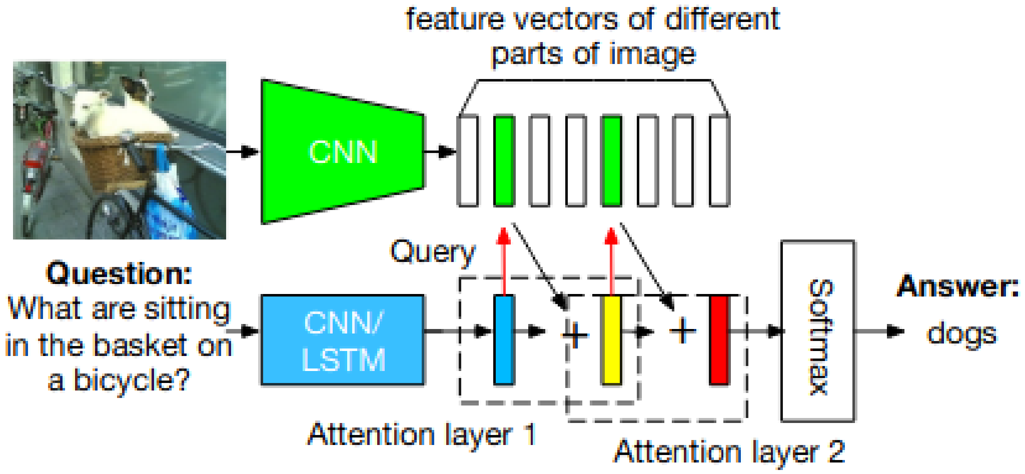

feature vectors of different parts of image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/ LSTM
+
+
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

Figure 3: Stacked Attention Block

3.4 Answer Generation

In this stage, we utilize the fused features to understand the input image and the question asked, to produce a possible answer to the question asked. We apply a non-linearity to generate probabilities for the answer classes. During training, we evaluate the cross-entropy loss using these probabilities to evaluate the difference in the log-likelihood of the predicted answers and the ground truth answers for the images. This stage is also referred to as the Classifier layer.

4 Experiments

Using our models we perform experiments on four different tracks - different datasets, different hyper-parameters such as optimisers and learning rates, the two different models to observe the results for them on some test cases.

4.1 Dataset Details

We perform the task of Visual Question Answering on the VQA 1.0 and VQA 2.0 dataset. The number of images in the training set for VQA_v1.0 is 82,783 images with each image of size 640×480 being provided with 3 questions and each question having 10 possible answers. The validation set has 40,504 images and the test set has 81,434 images with 3 questions for each image.

For VQA 2.0, we have the same number of images although the number of questions for each image vary for each image. Each question again has 10 answers from 10 different annotators.

4.2 Experiments Performed

As described earlier, we perform the experiments on the two different frameworks - the Tiled Attention method and the Stacked Attention method to compare the results of the two frameworks.

For each experiment, we use the Cross Entropy Loss to train our network and also restrict our vocabulary to the 1000 most common single-word answers in the training data. We made this choice because we observed that the top 1000 answers covered almost 86% of the total answers in the dataset. Also, having a possible probability value for each word in the dataset at the classifier output would become too computationally heavy for the network. We also consider the maximum question length to be 15 to establish an upper bound on the length of the embedding generated by the Text Processing phase for the questions.

4.3 Hyperparameter Tuning

We tuned such as hyperparameters learning rate, batch size, dropout probability and also changed the optimizer types to find the best set of hyperparameters for our models. The best hyperparameters we obtained were a learning rate of 0.005 using the cross entropy loss, a batch size of 32, a dropout probability of 0.2 during training. We also found the Stochastic Gradient Descent (SGD) optimizer to work best for the model.

4.4 Quantitative Results

We first observe the training curves for the training of our models and can compare the loss and accuracy curves for them. These training curves are shown in Figures 4. We can conclude that the training for the Tiled Attention Model is slightly better than the Stacked Attention Model.

Next, we observe the validation losses and accuracies for the two models. The curves are shown in Figures 6. Our conclusion can be that the Tiled Attention Model performs slightly better on the validation set as well.

We can tabulate the maximum accuracies achieved by our models in the table below

Model	Training Accuracy	Validation Accuracy
Tiled Attention Model	0.672	0.472
Stacked Attention Model	0.518	0.455

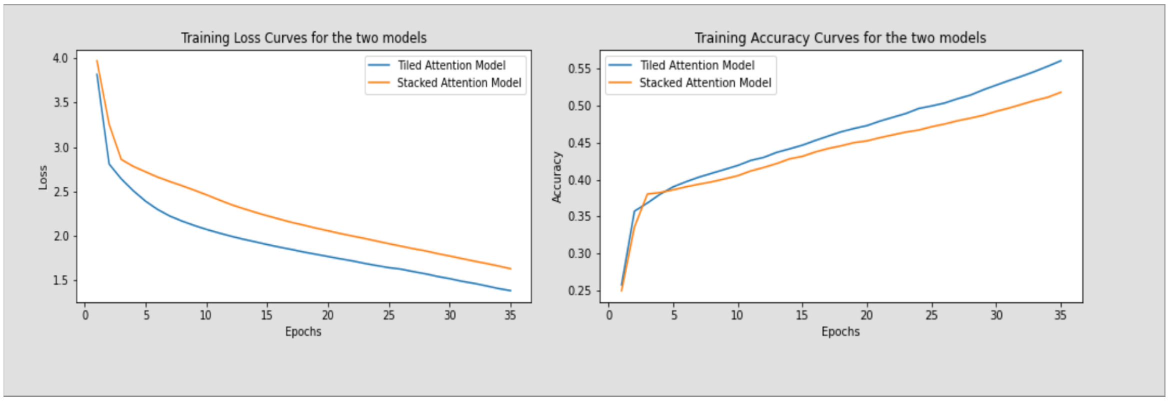


Figure 4: Training Loss curves for the two models. We use a batch size of 32, learning rate of 0.005 and a dropout rate of 0.2.

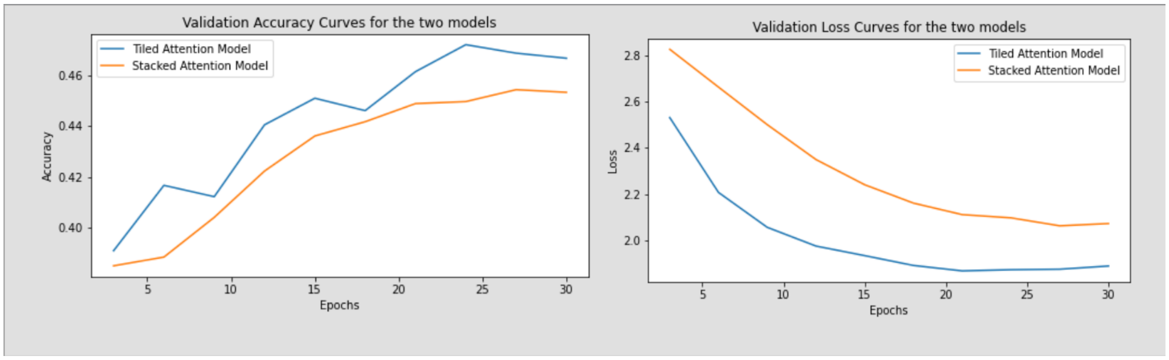


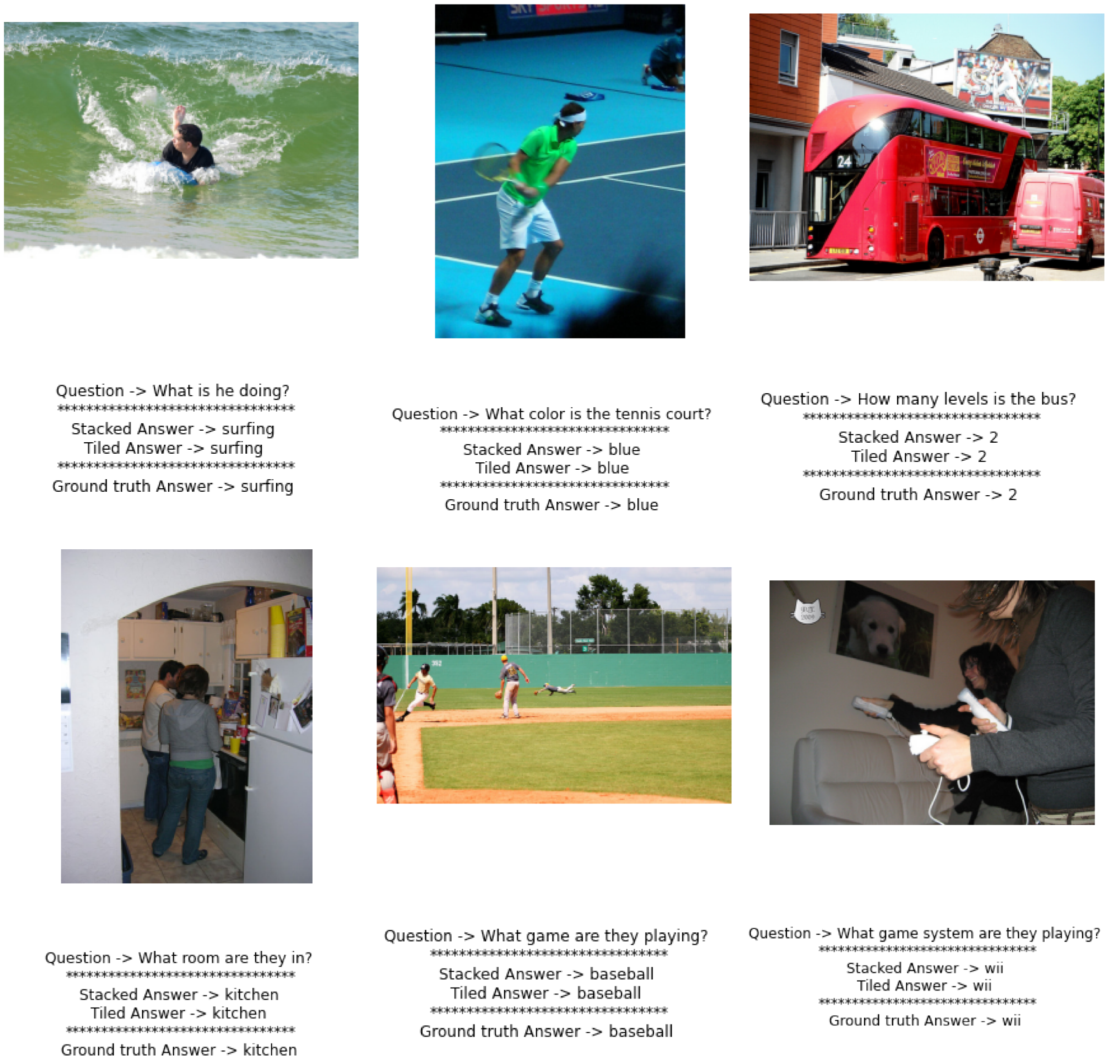
Figure 5: Training Accuracy curves for the two models. We use a batch size of 32, learning rate of 0.005 and a dropout rate of 0.2.

Since there are 1000 possible answers in our vocabulary, an accuracy of 0.47 on the validation set indicates that the model has been trained to understand the context of both the image and the question to generate a logical answer in most cases. Also note that in the validation set, there are possible answers which contain words not in the vocabulary thus resulting in a decrease in the accuracy of the models.

4.5 Qualitative Results

In Figures 6 highlight some of the positive cases where the model predicted the answer correctly as per the given ground truth. It can be seen that the questions are generally "what" type of questions and have a some object relation in the image to associate with the question.

In Figures 7 we show some of the negative cases where the answers didn’t match the given ground truth. Most of these failure cases are on questions that are ambigous or too subjective for the model to associate with the image.



```
Question -> What is he doing?
*********************************
Stacked Answer -> surfing
Tiled Answer -> surfing
*********************************
Ground truth Answer -> surfing
```

```
Question -> What color is the tennis court?
*********************************
Stacked Answer -> blue
Tiled Answer -> blue
*********************************
Ground truth Answer -> blue
```

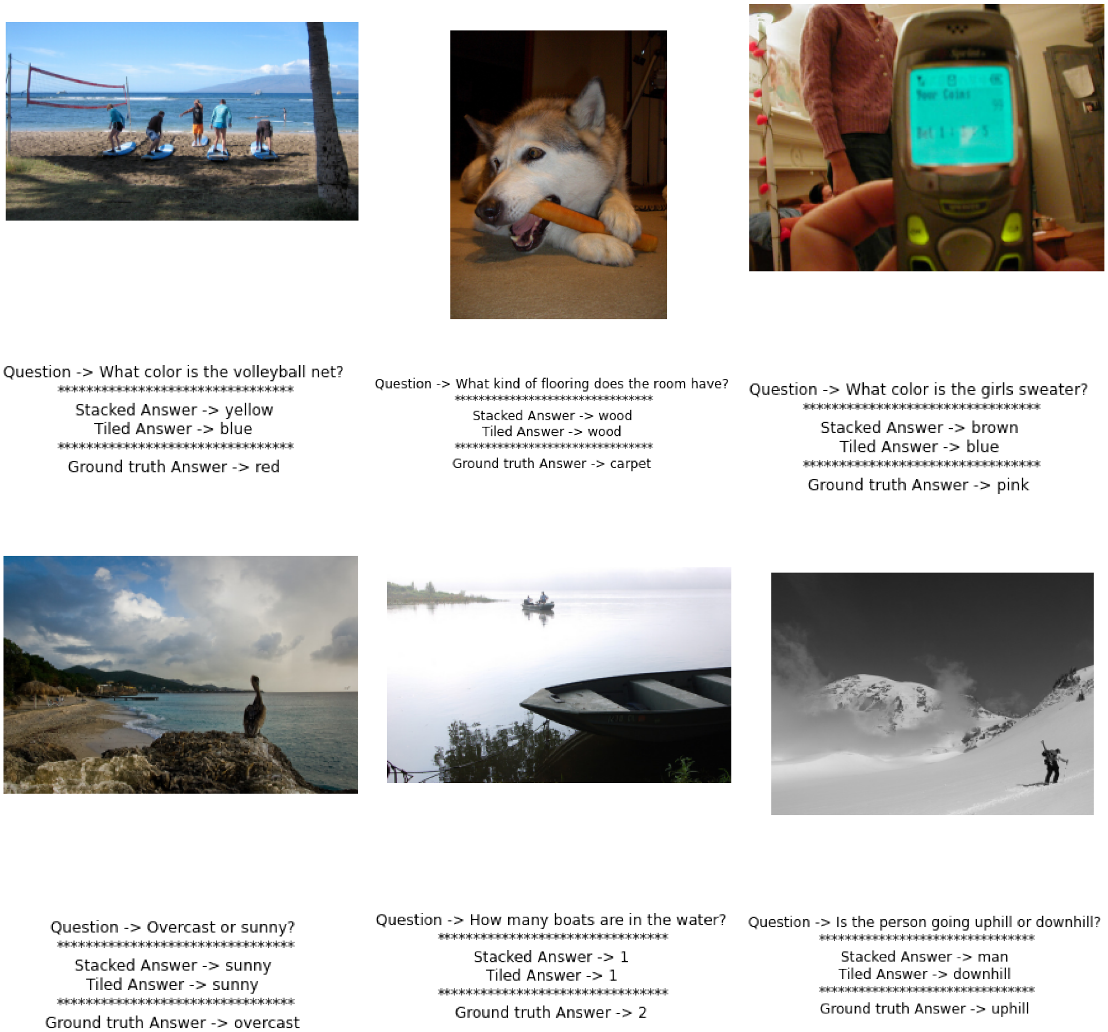
```
Question -> How many levels is the bus?
*********************************
Stacked Answer -> 2
Tiled Answer -> 2
*********************************
Ground truth Answer -> 2
```

```
Question -> What room are they in?
*********************************
Stacked Answer -> kitchen
Tiled Answer -> kitchen
*********************************
Ground truth Answer -> kitchen
```

```
Question -> What game are they playing?
*********************************
Stacked Answer -> baseball
Tiled Answer -> baseball
*********************************
Ground truth Answer -> baseball
```

```
Question -> What game system are they playing?
*********************************
Stacked Answer -> wii
Tiled Answer -> wii
*********************************
Ground truth Answer -> wii
```

Figure 6: Examples of both models predicting the correct answer as per the given ground-truth



Question -> What color is the volleyball net?
****
Stacked Answer -> yellow
Tiled Answer -> blue
****
Ground truth Answer -> red

Question -> What kind of flooring does the room have?
****
Stacked Answer -> wood
Tiled Answer -> wood
****
Ground truth Answer -> carpet

Question -> What color is the girls sweater?
****
Stacked Answer -> brown
Tiled Answer -> blue
****
Ground truth Answer -> pink

Question -> Overcast or sunny?
****
Stacked Answer -> sunny
Tiled Answer -> sunny
****
Ground truth Answer -> overcast

Question -> How many boats are in the water?
****
Stacked Answer -> 1
Tiled Answer -> 1
****
Ground truth Answer -> 2

Question -> Is the person going uphill or downhill?
****
Stacked Answer -> man
Tiled Answer -> downhill
****
Ground truth Answer -> uphill

Figure 7: Examples of both models predicting the correct answer as per the given ground-truth

5 Conclusion

In this work we explore the problem of visual question answering using attention based methods to fuse information from the image and text domain. Although we are working on a subset of the questions with the 1000 most frequent answers, some questions like "Why" type questions seem too abstract and subjective for the proposed system to solve and often seem irrelevant to the given image. Better results could have been achieved by working of "what" type of questions that have stronger connection with the associated image. Both attention mechanisms have a latent attention map which we could not investigate. Both [1] & [10] have a latent attention map that gives strong clues to where the network focuses on to generate the answer. This can be an insightful investigation to work towards. This combined with the visual grounding can lead to better models. Another interesting direction of this work could be to explore more sophisticated attention mechanisms like transformers along with better language models trained on larger corpus of text data.

6 Supplementary Material

The code can be found in the following Github repository

The presentation video recording can be found here

References

[1] Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering

[2] A Focused Dynamic Attention Model for Visual Question Answering

[3] Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering

[4] VQA: Visual Question Answering, Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

[5] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A Visual Turing Test for Computer Vision Systems. In PNAS, 2014.

[6] J. Deng, A. C. Berg, and L. Fei-Fei. Hierarchical Semantic Indexing for Large Scale Image Retrieval

[7] G. Kulkarni, V. Premraj, S. L. Sagnik Dhar and, Y. Choi, A. C. Berg, and T. L. Berg. Baby Talk: Understanding and Generating Simple Image Descriptions. In CVPR, 2011

[8] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain Images with Multimodal Recurrent Neural Networks. CoRR, abs/1410.1090, 2014

[9] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and ZeroShot Recognition. In ICCV, December 2013.

[10] Zichao Yang , Xiaodong He , Jianfeng Gao , Li Deng , Alex Smola Stacked Attention Networks for Image Question Answering. In CVPR, 2016.