VQA: Visual Question Answering

Pratik Antoni Patekar
Pursuing Masters in Computer Science
University of Texas at Arlington
Arlington, Texas - 76013
Email: pratik.patekar1123@gmail.com
Mobile: +1 682271274

***Abstract*— Visual Question Answering (VQA) is a Computer Vision task that requires a machine to answer questions about images. This can be a challenging task, as it requires the machine to understand the content of the image, as well as the meaning of the question.**

In this project, two VQA neural network models are developed, one without use of attention layer and another using attention layer. These models can answer questions about images with an accuracy of 49.84% and 28.02% respectively. We will explore these approaches to VQA and evaluate their effectiveness.

Keywords—Visual Question Answering (VQA), Neural network model and Computer Vision.

I. Introduction

Visual Question Answering (VQA) is a task that combines computer vision (CV) and natural language processing (NLP). Given an image and a natural language question about the image, a VQA system produces a natural language answer. VQA has many real-world applications, such as assisting visually impaired people, improving search engines and automating customer support. In simple words, VQA system takes image and a free-form, open-ended, natural-language question about the image as inputs and produces a natural-language answer as the output.

VQA is a challenging task, as it requires the system to understand both the content of the image and the meaning of the question. However, VQA is a rapidly developing field, and there have been significant advances in VQA accuracy in recent years. As VQA systems continue to improve, they will have a major impact on a variety of applications.

II. Related work

There is a lot of research and development going on related to Visual Question Answering. Following are a few that have been referred for this project work:

VQA: Visual Question Answering [1] was published in May 2015 and is one of the first papers that introduced the VQA task and dataset. The authors collected over 760,000 questions and 10 million answers on 200,000 images from the COCO dataset. They also proposed a baseline model that used a convolutional neural network (CNN) for image features and a long short-term memory (LSTM) network for question features, followed by a softmax classifier for answer prediction. The same datasset has been used for this project as well.

Later in 2016, ***Stacked Attention Networks for Image Question Answering*** [2] was published and it proposed a novel attention mechanism that allowed the model to focus on different regions of the image based on the question. The authors used multiple stacked attention layers to refine the image features before combining them with the question features. They showed that their model outperformed the baseline model on the VQA dataset.

Most existing VQA models relied too much on language priors and failed to capture the visual information in the image. The authors of "***Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering*** [3]**"** proposed a new balanced VQA dataset that reduced the language biases and increased the visual diversity. They also introduced a new model that used bottom-up and top-down attention to better align the image and question features.

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering [4] paper combined the bottom-up attention from object detection with the top-down attention from language modeling to improve the performance of both image captioning and VQA tasks. The authors used Faster R-CNN to detect salient regions in the image and generate object features, which were then weighted by an attention mechanism conditioned on the question or caption. They achieved state-of-the-art results on several benchmarks, including the VQA Challenge 2017.

A comprehensive survey of various datasets and models that have been used for VQA can be found in ***Survey of Visual Question Answering: Datasets and Techniques*** [5]. The authors compared different datasets based on their size, type, quality, difficulty, and diversity. They also reviewed different models based on their architecture, input, output, attention, reasoning, and evaluation.

III. Problem statement

The problem this project aims to address is the development of a Visual Question Answering (VQA) model that can accurately answer a wide range of questions about realworld images. The VQA task is challenging as it requires the model to understand both the visual content of an image and the meaning of the question posed in natural language. This requires the integration of techniques from multiple fields, including deep learning, natural language processing, and computer vision.

The problem this project aims to address is the development of a Visual Question Answering (VQA) model that can accurately answer a wide range of questions about realworld images. The VQA task is challenging as it requires the model to understand both the visual content of an image and the meaning of the question posed in natural language. This requires the integration of techniques from multiple fields, including deep learning, natural language processing, and computer vision.

The goal of this project is to develop a system that can answer questions about images in natural language. The system should be able to understand the meaning of the question, reason about the content of the image, and generate a correct and informative answer.

There are several challenges that need to be addressed to develop a successful VQA system. These challenges include:

1. Image understanding: The system needs to be able to understand the content of the image, including the objects, people, and their relationships.
2. Question understanding: The system needs to be able to understand the meaning of the question, including the entities, relations, and the type of answer that is being asked for.
3. Reasoning: The system needs to be able to reason about the content of the image and the question in order to generate a correct and informative answer.
4. Answer generation: The system needs to be able to generate a natural language answer that is relevant to the question and the image.

IV. Problem solution

This project uses the VQA v2.0 dataset (containing images from MS COCO image dataset and questions and answers with respect to those images) to build a neural network model that can answer questions that ask counts of objects in the image. The VQA task was reduced to VQA counting task to reduce the complexity of handling the huge dataset and to get better accuracy and results.

The explanation of the VQA counting model can be broken down into 3 main parts namely,

1. Image encoder and image model
2. LSTM (Long Short-Term Memory) model
3. Fusion model

Each of the above-mentioned parts are discussed in detail below.

V. Image encoder

This project uses VGG16 neural network model to encode all the images. Any given input image is first encoded using VGG16 model (using some pretrianed ImageNet weights).

VGG16 is a convolutional neural network (CNN) model that was developed by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group (VGG) at the University of Oxford. It was first introduced in their paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" in 2014. VGG16 is a 16-layer deep CNN model that is made up of a series of alternating convolutional and pooling layers. The convolutional layers extract features from the input image, while the pooling layers downsample the feature maps to reduce the amount of computation required. Figure 1 visualizes the VGG16 model used in this project.

The output of the VGG16 encoder is then given as input to the image model. The image model is nothing, but an input layer followed by a dense layer. Figure 2 shows the image model that follows the VGG16 image encoder.

VI. LSTM (Long Short-Term Memory) model

Long short-term memory (LSTM) is a type of recurrent neural network (RNN) that is commonly used to process sequential data, such as text and speech. LSTMs are able to learn long-term dependencies in data, which makes them well-suited for tasks such as machine translation, speech recognition, and text generation.

This project uses the LSTM model to convert and extract features from the question input which is in string format. As the questions are in string format, they cannot be directly fed to the LSTM model. This requires tokenization of every question string using an encoder. As already mentioned, this project uses VGG16 as an image encoder which is already available. Similarly, there are many readily available tokenizers and encoders such as BERT. But here we have built a tokenization method from scratch to have more control over what words will be tokenized as what and how the symbols in the question are taken care of.

For example, as per our encoding method, the word '*countertops*' is encoded as an integer '*1000*'. Any given question string is converted to an array of encoded array of length 18 (as 18 is the maximum length of the question string). This encoded array is then given as input to the LSTM model. The LSTM model is nothing but a combination of few embedding layers and LSTM layers followed by Dense layer.

Figure 3 shows the architecture of the LSTM model that is used for this project.

VII. Fusion model

The fusion model basically combines/ fuses the outputs of the image model and LSTM model together. The fusion model can learn to extract features from images and process sequential data. This makes the fusion model well-suited for a variety of tasks, such as image captioning, visual question answering, and machine translation.

This project uses the multiply method but there are several methods available for fusion such as:

1. Concatenation: This method simply concatenates the output of the image model and the output of the LSTM model. This results in a new feature vector that contains both the spatial information from the image and the temporal information from the LSTM model.
2. Addition: This method adds the output of the image model and the output of the LSTM model. This results in a new feature vector that contains both the spatial and temporal information from both models.
3. Multiply: This method multiplies the output of the image model and the output of the LSTM model. This results in a new feature vector that contains both the spatial and temporal information from both models, but it also weights the information from each model differently.
4. Attention: This method uses an attention mechanism to weight the information from each model differently. This allows the model to focus on the most relevant information from each model.

The architecture of the fusion model can be seen in Figure 4.

VIII. Model using attention layer

As mentioned in the introduction, we have also built another model with attention layer. The image model and the LSTM model used are same, the difference is just that the fusion method used is attention instead of multiply. The architecture of this model can be seen in Figure 5.

IX. Comparisons with other models

Both the models were trained on local machines and thus could not be trained to higher accuracies, but they showed a steady increase in accuracy while training on training dataset. Below is some information on the accuracies and training processes followed.

The first model which used multiply method for fusion acquired an accuracy of 49.84% in just 15 training epochs on local machine. The second model which used attention method for fusion acquired an accuracy of 28.02% in just 5 epochs. It is important to note that though the accuracy of these models is very low as compared to the state-of-art VQA models, these models were trained on relatively very small data due to shortage of time.

Comparison of accuracy of these models with the state-of-art VQA models is shown in Table 1:

TABLE I. Comparison with state-of-the-art models

Model name	Accuracy
BART-VQA	94.7%
ViLBERT	94.5%
DFAF-BERT	94.4%
Visual-BERT	94.3%
VilBERT	94.2%
VQA using multiply fusion	49.84%
VQA using attention fusion	28.02%

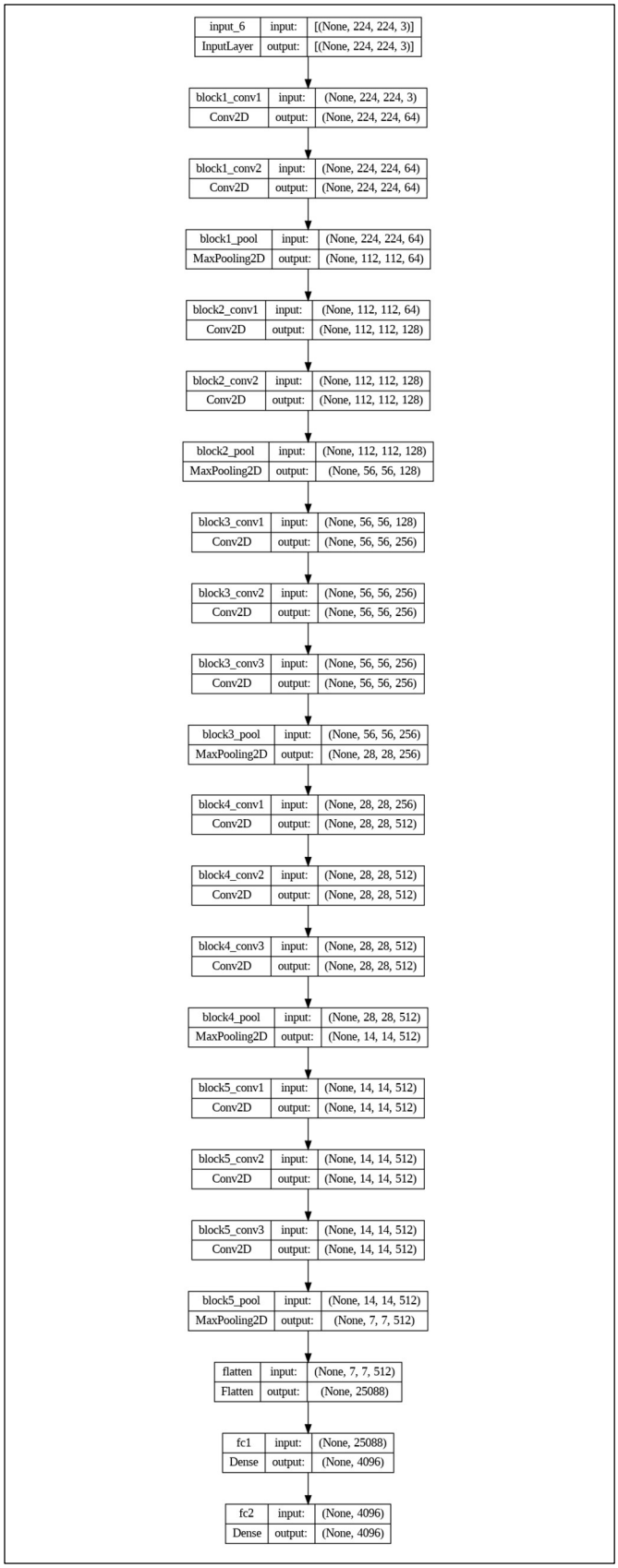

Fig. 1. VGG16 Image encoder (VGG16 is a 16-layer deep CNN model)

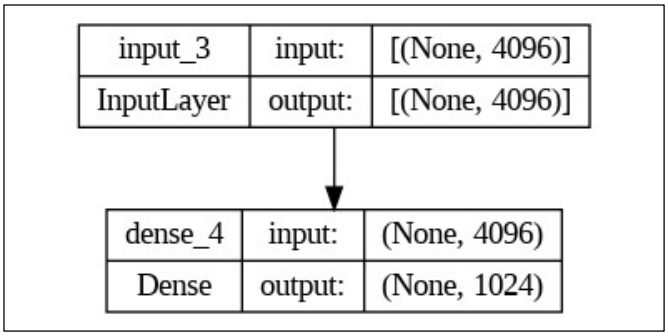

Fig. 2. Image model (This image model follows the VGG16 image encoder)

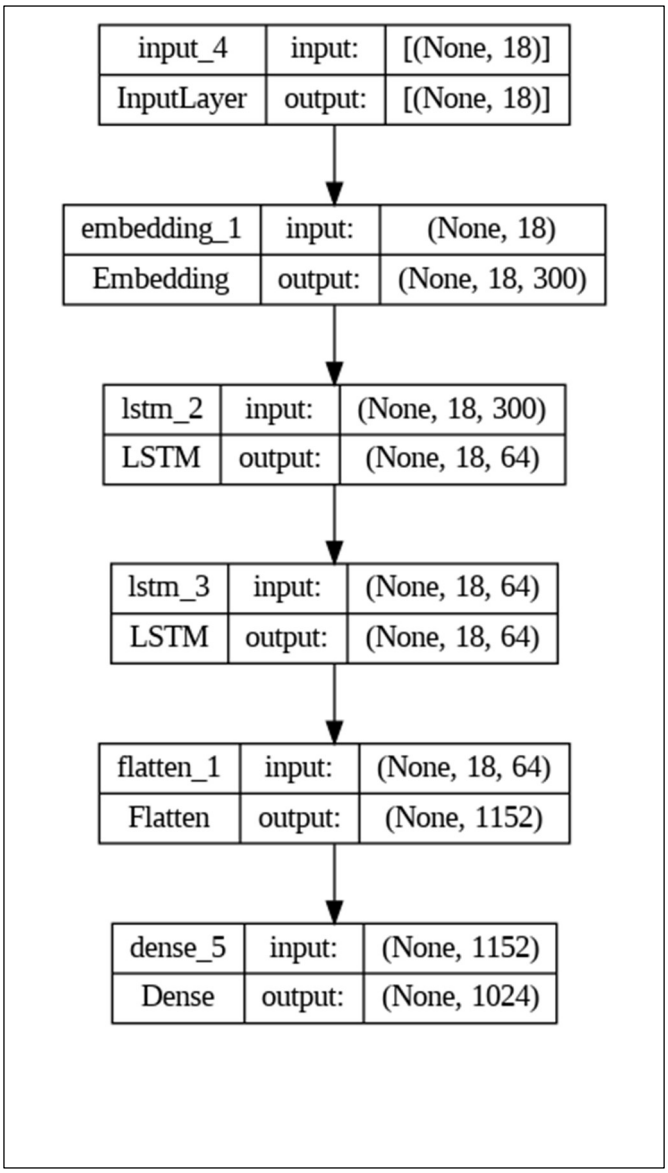

Fig. 3. LSTM model

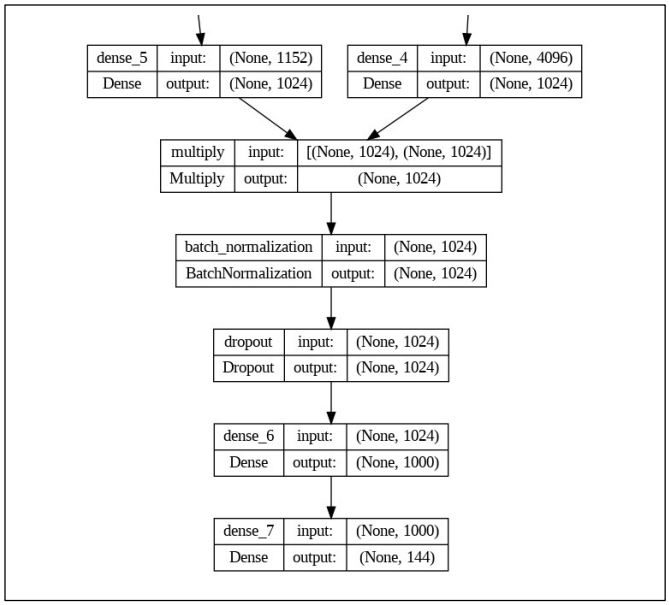

Fig. 4. Multiply based fusion model (The method used for fusion is multiplying output of image model and LSTM model).

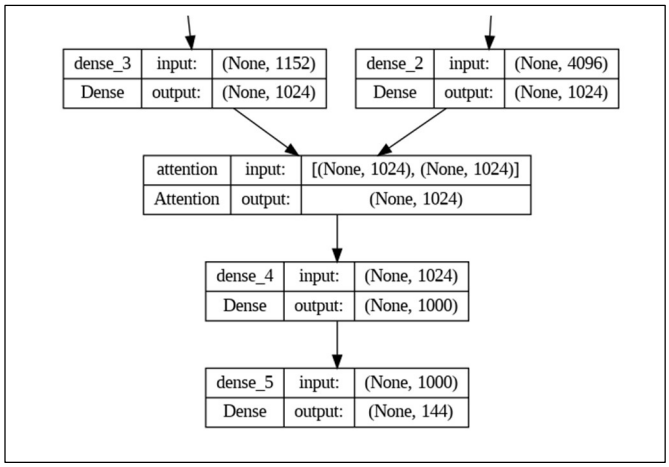

Fig. 5. Attention based fusion model

X. Results and Conclusions

Figures 6 to 9 show some results of the VQA model with multiply fusion method for some random images from training and validation dataset. The model basically gives the probabilties for all the possible answers in the answer set and the one with highest probabilty is given as the predicted answer. The code written also returns the top 4 answers and their respective probabilities.

Figures 10 and 11 show some results of the Attention based VQA model. The code written gives results same as above. It also gives a result image which has portions of input image highlighted. These highlighted portions are the parts of the image to which the model gives attention to. These is obtained from the output of the attention layer of the model.

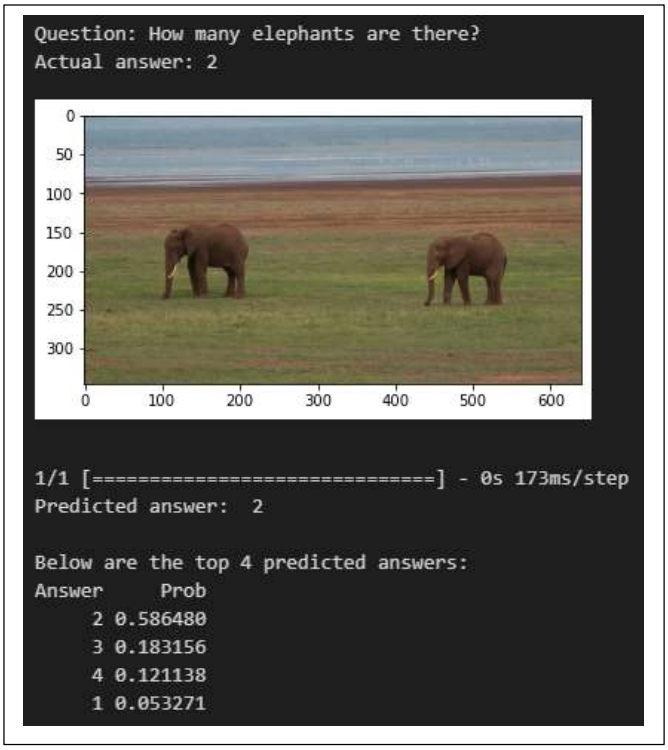

Fig. 6. Sample result 1 of model with multiply based fusion model for random training image from dataset.

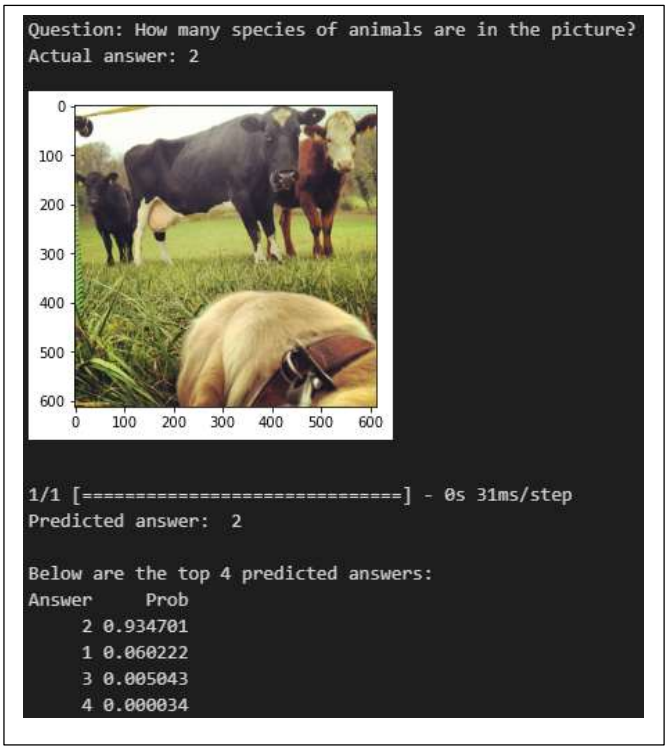

Fig. 7. Sample result 2 of model with multiply based fusion model for random training image from dataset.

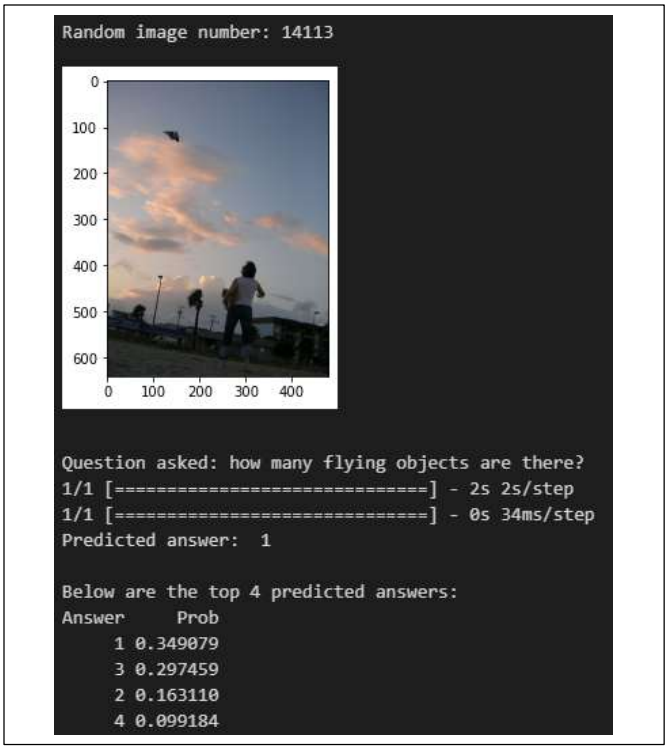

Fig. 8. Sample result 1 of model with multiply based fusion model for random image from validation dataset.

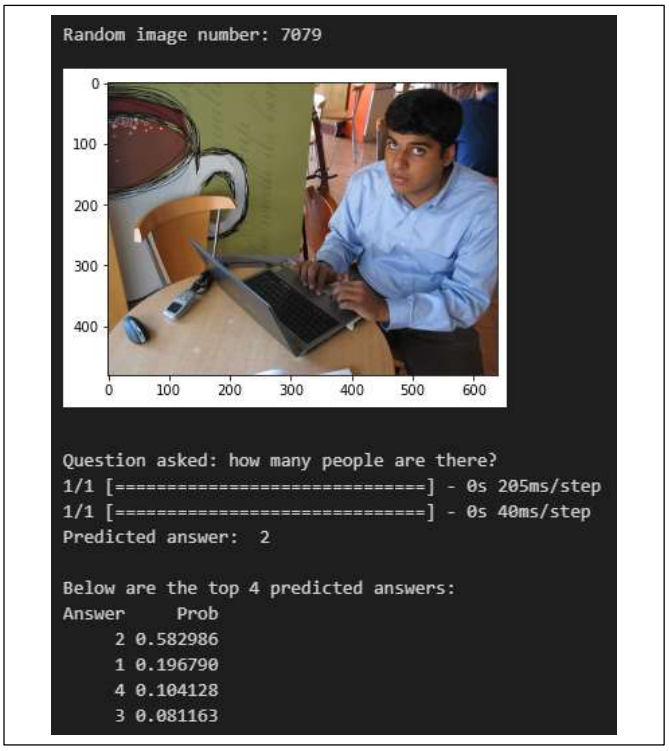

Fig. 9. Sample result 2 of model with multiply based fusion model for random image from validation dataset.

Fig. 10. Sample result 1 of model with attention based fusion model for random image from training dataset.

Fig. 11. Visulaization of what parts of image the model gives attention to while answering a question. (*Note: The attention map is not looking great as the model has not been trained to high accuracy*).

Conclusion: In this project, two models were built for the task of visual question answering (VQA). The first model was a baseline model, and the second model was a more advanced model that was built using a technique called attention. The two models were evaluated on the VQA v2.0 dataset, and the results showed that the second model, which used attention, was able to answer questions even though it was trained on less data and had lower accuracy compared to the baseline model.

The results of this project suggest that attention is a valuable technique for improving the performance of VQA models. Future work could explore the use of other techniques, such as reinforcement learning, to further improve the performance of VQA models.

REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, "VQA: Visual Question Answering," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 2425-2433, doi: 10.1109/ICCV.2015.279.

[2] Zichao Yang et al., "Stacked Attention Networks for Image Question Answering," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas NV USA , Jun-Jul 2016.

[3] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra and D. Parikh, "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6325-6334, doi: 10.1109/CVPR.2017.670.

[4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 6077-6086, doi: 10.1109/CVPR.2018.00636.

[5] K.-M Kim et al., "Survey of Visual Question Answering: Datasets and Techniques," Fifth IAPR International Conference on Computer Vision and Image Processing (CVIP), Mandi India , Dec 2020.

[6] K.-M. Kim et al., "Visual question answering: Datasets, algorithms, and future challenges," Computer Vision and Image Understanding, vol. 163, pp. 3-20, Nov. 2017.

[7] J.-H Kim et al., "Focal Visual-Text Attention for Visual Question Answering," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville TN USA , Jun-Jul 2021.

[8] D.-K Kim et al., "Learning to count objects in natural images for visual question answering," International Conference on Learning Representations (ICLR), Vancouver BC Canada , Apr-May 2018.

[9] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra and D. Parikh, "Yin and Yang: Balancing and Answering Binary Visual Questions," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 4956-4964, doi: 10.1109/CVPR.2016.542.

[10] A.-Lai et al., "Vanilla VQA," Allen Institute for AI Blog , Apr 2017 [Online]. Available: https://blog.allenai.org/vanilla-vqa-adcaaaa94336