Visual Question Answering

Gregory Brown, Sai Sri Narne, Priyanka Gaikwad, Sai Srinivas Vidiyala, Sai Charan Kottapalli
Department of Computer Science Electrical Engineering
University of Missouri-Kansas City
gbkhv@mail.umkc.edu,sn69k@mail.umkc.edu,pvg9gb@mail.umkc.edu, svcxc@mail.umkc.edu, skn4g@mail.umkc.edu

Abstract— The task of Visual Question Answering is to produce a system which, given an image and a natural language question about the image, will provide an accurate natural language answer. There is more research going and a lot of investments by organizations in ma- chine learning, big data, artificial intelligence is increasing to help visually impaired or physically disabled humans. Similarly, this paper discusses about developing a Visual Question Answering system on given image and natural language question about the image. The task by the system is to provide an accurate natural language answer. This helps a visually impaired personality to grab more information about the image based on the question posed.

Keywords— Deep Learning, Convolutional Neural Network, Recurrent Neural Network, Question and Answer Encoding

I. INTRODUCTION

As machine learning algorithms are emerging, image analysis is getting easier. But the question still being considered is how accurate it is analyzing and giving the answer text. Many of the researchers are coming up with different techniques to analyze what is in the image and read it out in natural language. In spite of this model being accurate, what if the system is not giving the natural language answer that a user need. Keeping this in mind, Visual Question Answering system was developed. This system takes the a natural language question along with an image and gives natural language answer more related to it.

In this paper, we attempt to produce an open ended Visual Question Answering (VQA) system. A VQA system uses as input an image and a free-form, open-ended, natural-language question about the image and returns a natural-language answer as the output. This task is applicable to scenarios encountered by visually impaired people to elicit information present in the image. Example questions from the dataset are shown in Fig. 1.

What is located in the grass What is walking through trees

Figure 1: Examples of free-form, open-ended questions collected from COCOQA dataset.

II. RELATED WORK

Several papers have begun studying the area of Visual Question Answering. VQA is recent ongoing search going in Artificial Intelligence System. The problem of Visual Question Answering (VQA) is a recent one which has been compared to a new kind of visual Turing test. The aim is to show progress of systems in solving even more challenging tasks as compared to traditional visual recognition tasks like object detection and segmentation. Stanislaw et. Al presented a large dataset containing 204,721 images from the MS COCO dataset and a newly created abstract scene dataset that contains 50,000 scenes. The MS COCO dataset has images depicting diverse and complex scenes that provide scenes of which compelling and diverse questions can be asked. Huijuan Xu et. Al proposed a novel multi-hop memory network using spatial attention for the VQA task. This allows one to visualize the spatial inference process used by the deep network. They have designed an attention architecture in the first hop which uses each word embedding to capture fine-grained alignment between the image and question.

Badri Patro et. Al adopted an exemplar-based approach to boost visual question answering (VQA) methods by providing what they call differential attention. They have evaluated two variants of differential attention - one where only attention is obtained and the other where both differential context and attention were obtained.

III. PROPSED WORK

In this paper, we adopt a network based on Convolutional Neural Network which is used for feature extractions and Recurrent Neural Network for question encoding task.

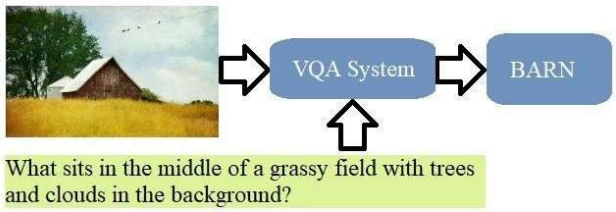


Figure 2: Overview of VQA system deals with image and related question

We have used COCOQA dataset which has around 123k images which are split between 78k for training and around 38k for testing.It has image ids which corresponds to images from MSCOCO dataset.In our paper,we are focusing only on one context of images.The context we selected for our project is Park images.Based on different keywords that represents park context such as bench,fountain,park etc are used for data collection task.

IV. SOFTWARE &SYSTEM ARCHITECTURE

Software:

In general filtering of the COCOQA dataset was performed using Scala and Spark. Retraining of Inception model was performed using Python 3.6.8 and Tensorflow
3.4.1. Visualizations of tensorflow training were created using Tensorboard. For each of 5 datasets these tasks were carried out on heterogenous machines.

System Architecture:

From our research of related projects, we have developed the project architecture depicted in Figure 3. Thusfar, we

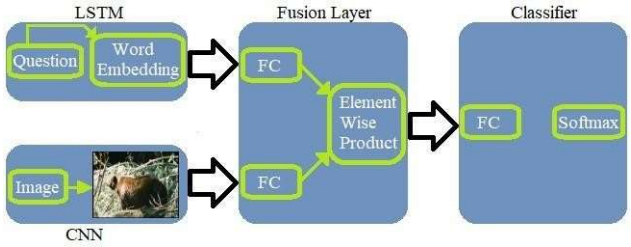
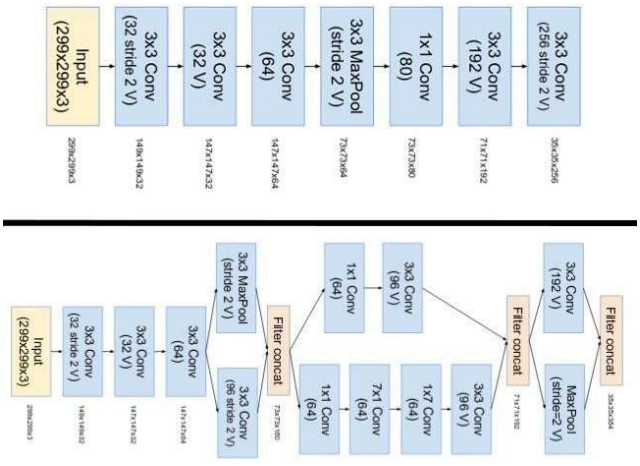


Figure 3: Overview of Visual Question Answering System with different components

have yet to work on the LSTM, the Fusion of features or our final Classifier. We have done feature extraction by retraining the Inception CNN, and to provide results on this have, in general, classified the images in our validation datasets into the keywords—relevant to parks—for which they were chosen. Below is the structure of inception v3



V. IMPLEMENTATION AND EVALUATION

No	Keywords	Train Images Count	Validation Images Count	Train Accuracy (Inception)	Validation Accuracy (Inception)
1	Bench	351	320	100	89
	Bridge	84	99		
	Frisbee	326	320		
2	Fence	100	30	100	52.2
	Trail	100	30		
	Path	100	25		
	Park	100	30		
3.	Roses	232	15	100	91.4
	Tulips	150	15		
	Sunflowers	175	15		
	Anchor	246	15		
	Accordion	173	15		
	Airplanes	193	15		
4	Butterfly	100	45	100	97
	Brain	100	40		
	Camera	100	30		
5	Court	132	34	99.0	66.3
	Flower	619	154		
	Leaf	209	53		
	Meadow	29	8		
	Stick	208	53		
	Walkway	62	16		

Table 1: Train and Validation Accuracy of Inception Model based on keywords related to Park Context

1. Evaluation Results (Based on dataset 1 results in table 1)

CNN and Inception models had results. CNN has thefollowing results for dataset-1.

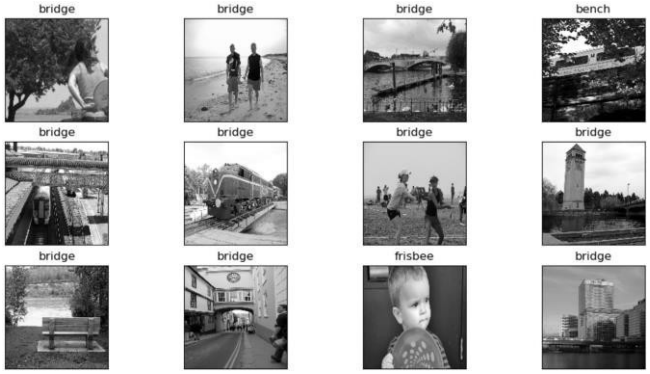


Figure 4: Overview of Classification Evaluation Results based on Bench, Bridge and Frisbee Keywords

There are few images that are misrepresented by this model, there is an image with bench, so this image should be under bench, but CNN model recognizes it as bridge, so this model not 100% accurate but the results are quite good.
While in Inception model the accuracy is best for dataset-1,the train accuracy is 100% and final validation accuracy is 89% with N=146. Below are the train and validation accuracy in graphical representation.

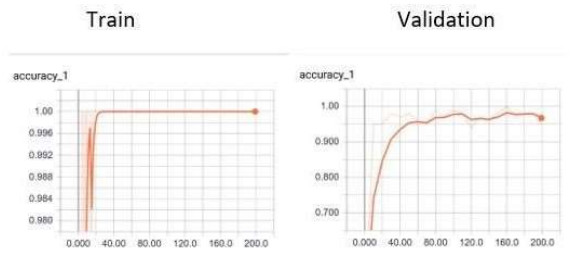


Figure 5: Training and Validation Accuracy in tensorboard graph

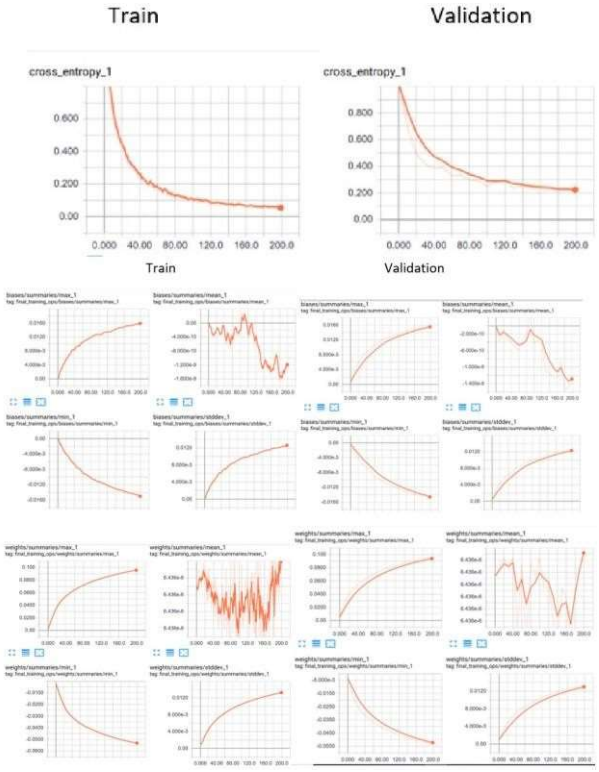


Figure 6: Training and Validation Loss in tensorboard graph

2. Evaluation Results (Based on dataset 2 results in table 1)

Classification CNN is trained with images from keywords such as Fence, Trail ,Path and Park. Hyperparameters are changed during training process as follows:
No of Epochs = 7
Learning Rate = 1e-4
Figure 7 represents the classification Model output in terms of classes identified based on given keywords.

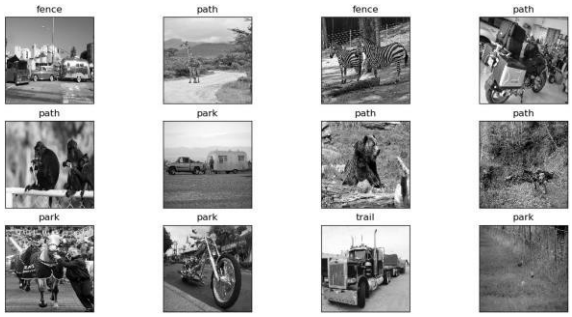


Figure 7: Overview of Classification Evaluation Results based on Fence, Trail, Path and Park Keywords

Results based on Inception Model:

The Validation Percentage for pretrained Inception Model is 52.2%. The reason for such a low Validation accuracy is due to non-similarity in images which are used. As lot of keywords doesn't truly represents the entire class the accuracy drops to 52.2%.

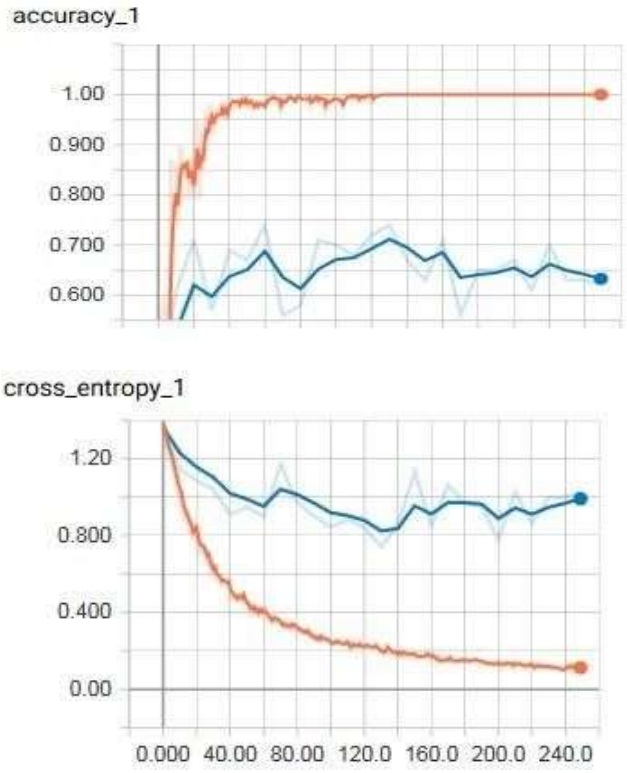


Figure 5: Training and Validation Accuracy and Cross Entropy Loss in Tensorboard

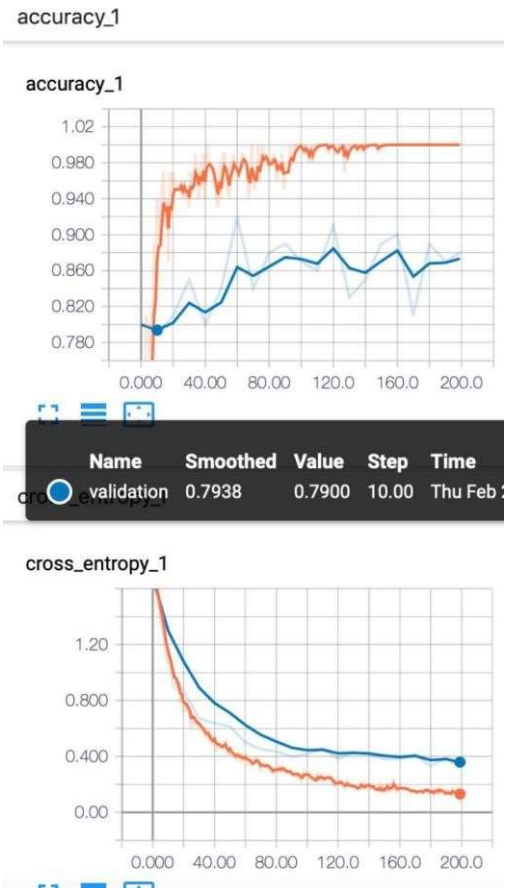
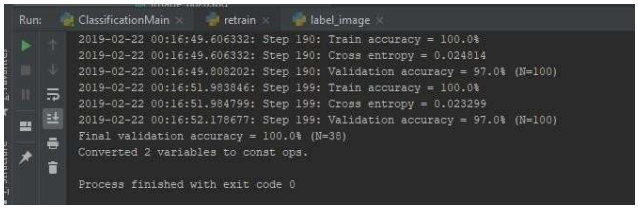


Figure 6: Training and Validation Accuracy and Cross Entropy Loss in Tensor board

Fig 8 Tensor board Results based on results data sets butterfly, brain, and camera

Fig 8

The validation percentage is 97 percent



The Validation is good because of more similar are used in my data set, While CNN haven't provided much percent as above inception model

5 Evaluation Results (Dataset 5)

Notably, our use for the retrained Inception model will not be classification but to provide features to be combined with features from the questions. The combined features will be used to train an additional classifier. This is useful to know because, while some datasets got good results in using the retrained inception model to do classification others did not. The best results achieved for dataset 5 were achieved using the following hyperparameters:

No of Epochs = 1000
Learning Rate = 0.01

Using these hyperparameters the retrained Inception model got 66.3% accuracy on the 190 total training samples (as can be seen in table 1). This means that the retrained Inception model is having difficulty distinguishing between the features of images which include the keywords 'Court', 'Flower', 'Leaf', 'Meadow', 'Stick', and 'Walkway'. Since these keywords are very similar it is unsurprising that the model has difficulty distinguishing between them. Based on these results it remains to be seen if the retrained Inception model can be used to provide a feature vector to our VQA system. Other strategies might be to train on the answers of the questions rather than the keywords used for filtering, or to train a new CNN on the images (either by keyword again, or by answer). One thing that retraining the Inception model using our data did show was the capabilities of the inception model. As the many results of my partners (shown prior) as well as Figures 10-13, show retraining the Inception model

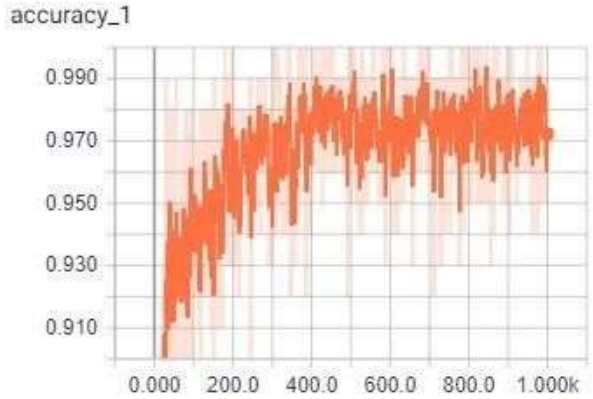


Figure 10: Training Accuracy Dataset 5

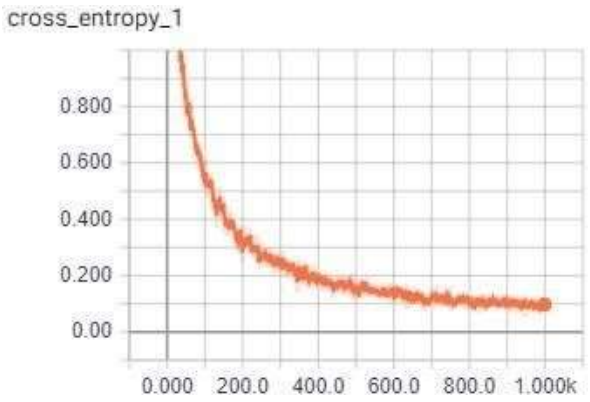


Figure 11: Training Cross Entropy Dataset 5

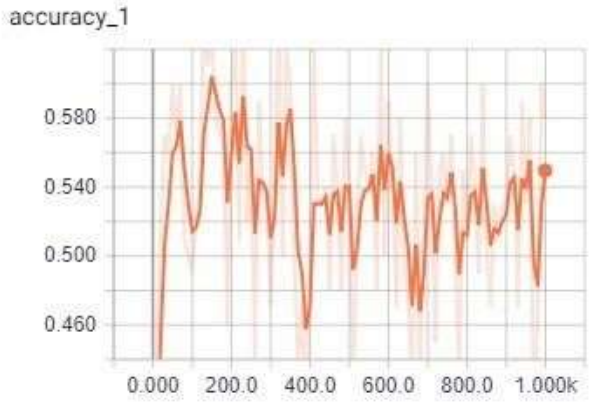


Figure 12: Validation Accuracy Dataset 5

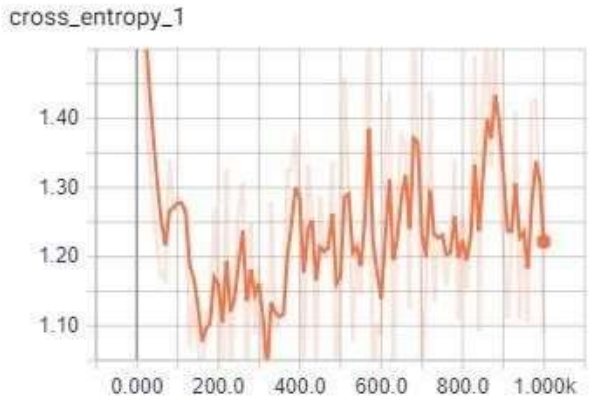


Figure 13: Validation Cross Entropy Dataset 5

can produce decent results in terms of classification, even with our datasets which do not lend themselves to classification in this way.

VI. IMPLEMENTATION

a. Implementation of Visual Question Answering:

As a part of Project Implementation, we have started working on VQA (Visual Question Answering), for which we refereeing to existing Model [6] for the same task:

System Architecture:

This is an attention-based model for VQA using a dilated convolutional neural network for modelling the question and a resnet for visual features. The text model is based on the recent convolutional architecture ByteNet. Stacked attention distributions over the images are then used to compute weighted image features, which are concatenated with the text features to predict the answer. Following is the rough diagram for the described model.

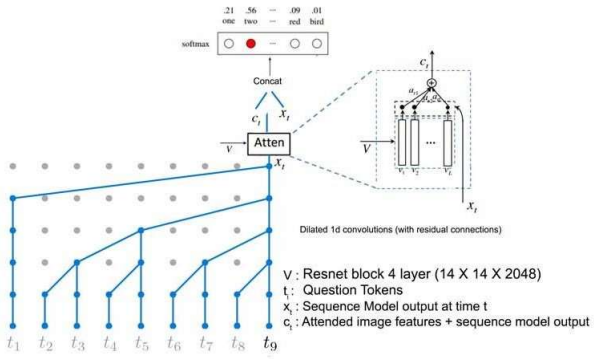


Figure 14: Overview of Visual Question Answering System with different components

Below is a brief description about existing VQA model:

- **Image Feature Extraction:**

The existing VQA model uses VGG16 and Resnet as part of Image Features Extraction with following options:

1. VGG16: Feature Extraction using **FC7 and POOL5**
   - VGG16 is a pretrained convolutional neural network for classification and detection
2. Resnet: Feature Extraction using **Block4**
   - Similar to VGG16 resnet is a pre-trained CNN

- **Question and Answer Encoding:**

The model uses Bytenet and LSTM as a part of text processing models.

Stacked attention distributions over the images are then used to compute weighted image features, which are concatenated with the text features to predict the answer.

Just to test existing Model, using some abstract scenes

Question: **"What color is the ladies pants?"**

Question: **"Is the woman on the couch sporting white hair?"**

The model is executed only for few images so training accuracy for the Model is 100% as it is predicting answers for the questions exact.

Validation Stage:

The model is validated only for 1 single image as of now as it will be train and validate on real images as a part of next implementation.

We are passing following image as an input image and question to the trained model and it is generating answers as follow:

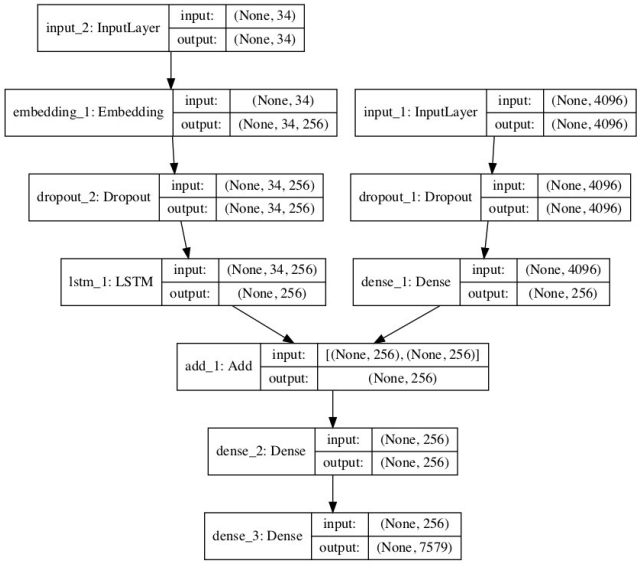
Question: **"What does dog want to chew on?"**

Answer: Exiting

As of now model is producing incorrect answers for the given image and question. So more data will be passed to model in future to fine tune it.

b. Implementation of Show and Tell Model:

For some members of our group, rather than working on VQA model during this time, we have worked on the Show and Tell caption generation model. While implementation of this model differs significantly from implementation of a VQA model, implementing the Show and Tell model have given us greater understanding of LSTM implementation which we can later use in our VQA model. The below diagram, taken from [7] gives an outline of the Show and Tell model which we have implemented:



In this diagram we can se that on the left-hand side shows the word embedding, culminating in an LSTM. The right-hand side shows the image embedding via CNN. The models resulting from the CNN and the LSTM are used for caption generation. Notably, the output shape of each layer on the LSTM side of the diagram will differ depending on our dataset, namely for datasets 3 and 4, 34 is replaced with 13, and for dataset 5 34 is replaced with 40, these numbers each corresponding to the maximum caption length.

Based on this Show and Tell model, we have performed example generation of captions for datasets 3, 4, and 5. The below table shows the Bleu metric results comparing the captions generated for test samples with the reference 'captions' for test samples.

Dataset	Bleu-1	Bleu-2	Bleu-3	Bleu-4
3	0.962306	0.947157	0.931470	0.907180
4	0.979792	0.971377	0.959821	0.943118
5	0.418192	0.255634	0.203081	0.129340

3.Dataset 3 was trained using 29 samples and validated using 29 samples. Below is an example of image an caption generated for dataset 3:

4. Dataset 4 was trained using 29 samples and validated using 29 samples. Below is an example of image and caption generated for dataset 4.:

Caption: 'a herd of sheep grazing on a lush green field'

5. In the case of Dataset 5 images associated with the keywords corresponding to dataset 5 form table 1. Ultimately 1222 <question, answer> pairs are used to train the model, and 320 <question, answer> pairs are used for testing. This, I believe, is different from what is done with datasets 3, and 4, each of which use the MSCOCO caption associated with the relevant images. For the same image used as an example for dataset 4 (img_id: 8881), this model generates the caption 'startseq what is the color of the meadow green endseq', which, ignoring the 'startseq' and 'endseq' as is done during the Bleu evaluation, this is identical to the reference <question, answer> pair for this image. As an additional example for img_id 74461 the reference <question, answer> pair is 'the park that has

what lined with benches walkway' and this show and tell model generates

the caption ‘startseq what is surrounded on the park walkway bench endseq’ which is not identical but very similar. However for every very similar or identical result, it seems that there is at least one result that is not even close. For example, for img_id 89005 the reference <question, answer> pair is ‘what are standing among

leaves and sticks birds’, however this model generates the question ‘startseq what is the color of the flowers green endseq’. The issues represented with the caption are the most common issues with using <question, answer> pairs as input for a caption generation system, namely that the caption generator may decide to ask a question about a different subject than the reference caption is about—here the question seems to be about the leaves misinterpreted as flowers—and the caption generation model may misinterpret some components of the image.

Clearly there is still some work to be done in tuning this model, however as ultimately this was practice in manual creation of LSTM it was successful. The captions generated even take a very similar grammatical form to the <question, answer> pairs provided for training.

VII. Conclusions

The Visual Question Answering is still ongoing researcharea in Artificial Intelligence System.

VIII. Datasets

Datasets:
1 – Sai Sri Narne
2 – Priyanka Gaikwad
3 – Sai Charan Kottapalli
4 – Sai Srinivas Vidiyala
5 – Greg Brown

References

[1] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying Visual Semantic Embeddings with Multimodal Neural Language Models. TACL, 2015

[2] Stanislaw Antol∗1 Aishwarya Agrawal∗1 Jiasen Lu Visual Question Answering

[3] Huijuan Xu, Kate Saenko Exploring Question Guided Spatial Attention for Visual Question Answering

[4] Badri Patro and Vinay P. Namboodiri Differential Visual Question Answering with Attention

[5] A Simple Guide to the Versions of the Inception Network, https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202, (For inception v3 diagram)

[6] https://github.com/paarthneekhara/convolutional-vqa

[7] https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/