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

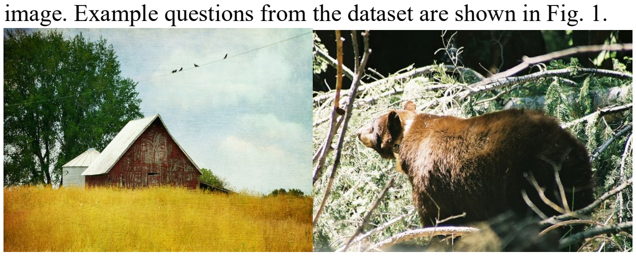
Abstract— **The task of Visual Question Answering is to produce a system which, given an image and a natural language question about the image, will provide an accurate natural language answer. There is more research going and a lot of investments by organizations in machine learning, big data, artificial 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.

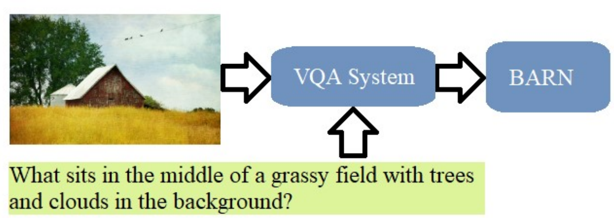


Figure 2: Overview of VQA system deals with image and related question

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

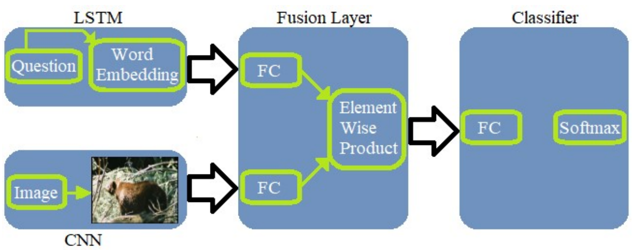
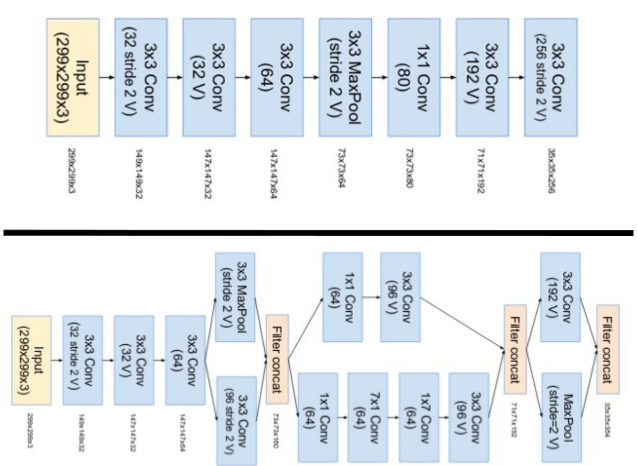


Figure 3: Overview of Visual Question Answering System with different components

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 the following results for dataset-1.

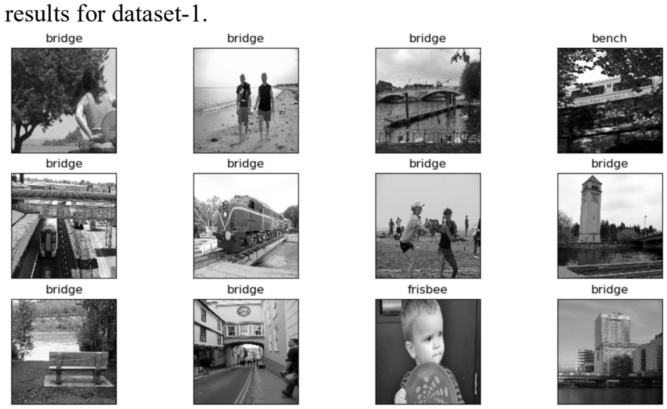


Figure 4: Overview of Classification Evaluation Results based on Bench, Bridge and Frisbee Keywords

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.

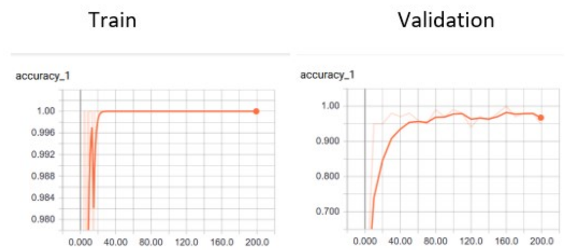


Figure 5: Training and Validation Accuracy in tensorboard graph

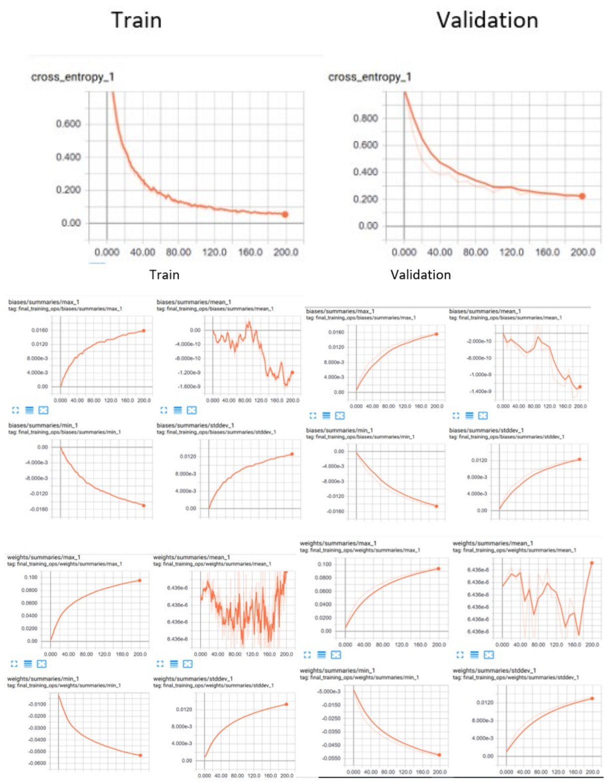


Figure 6: Training and Validation Loss in tensorboard graph

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.

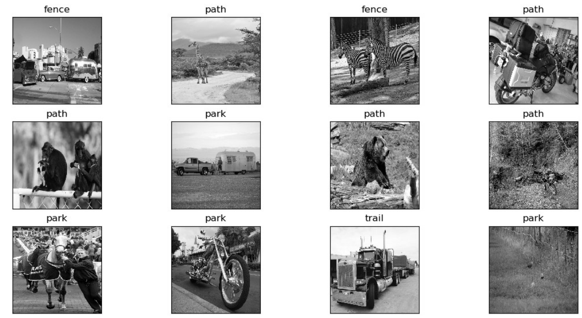


Figure 7: Overview of Classification Evaluation Results based on Fence, Trail, Path and Park Keywords

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%.

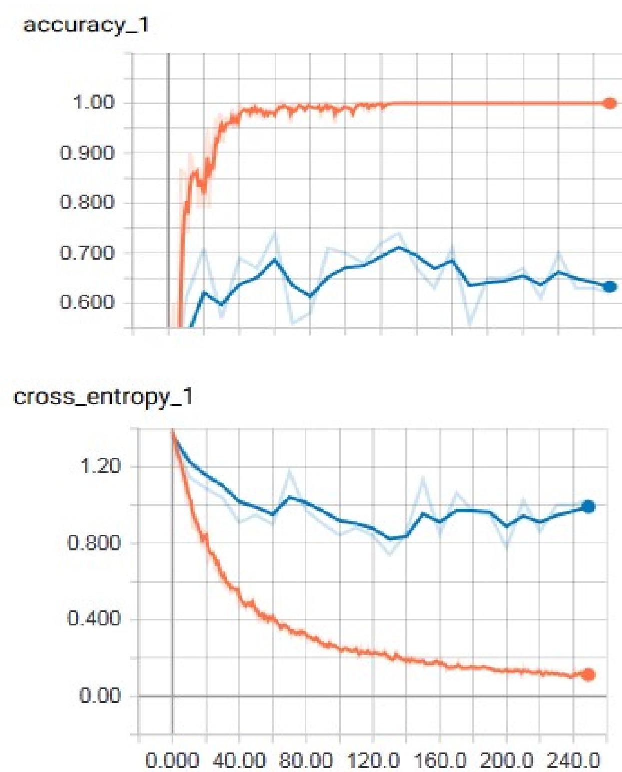


Figure 5: Training and Validation Accuracy and Cross Entropy Loss in Tensorboard

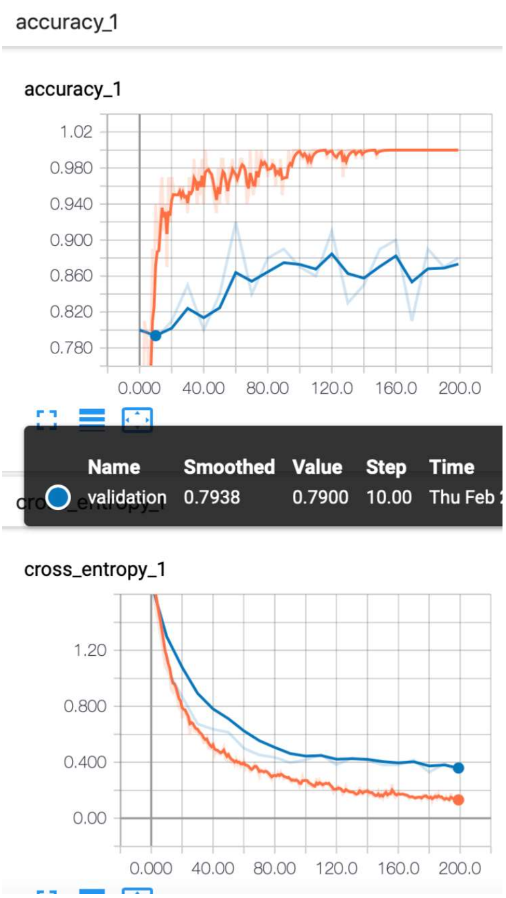
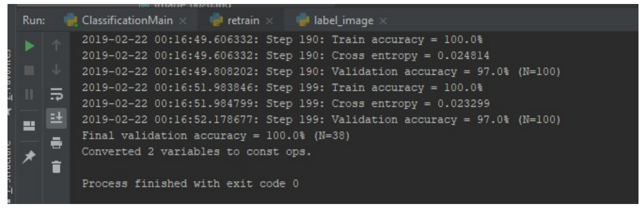


Figure 6: Training and Validation Accuracy and Cross Entropy Loss in Tensor board

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

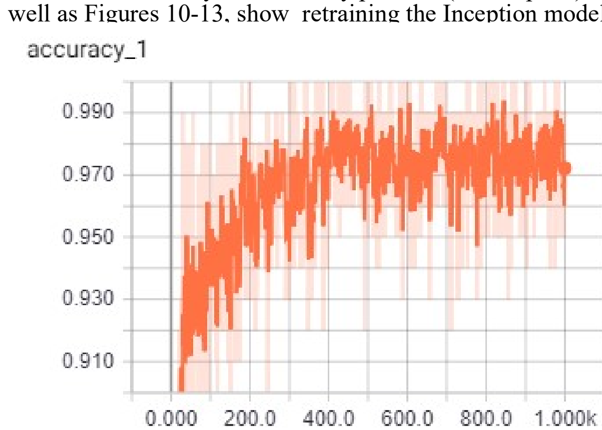


Figure 10: Training Accuracy Dataset 5

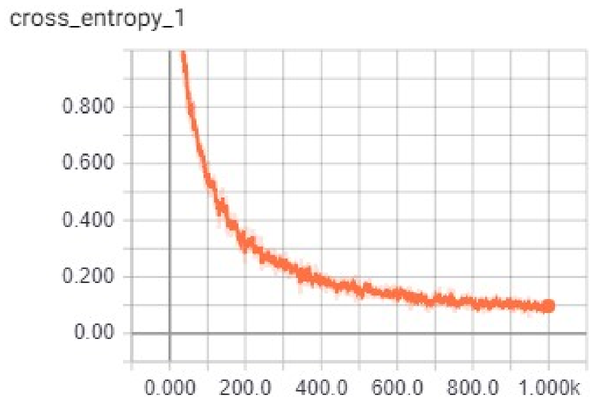


Figure 11: Training Cross Entropy Dataset 5

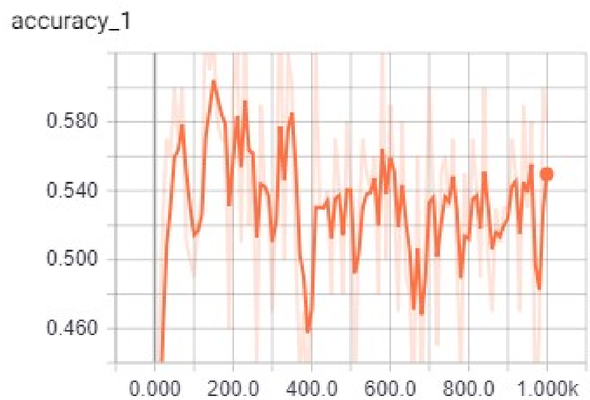


Figure 12: Validation Accuracy Dataset 5

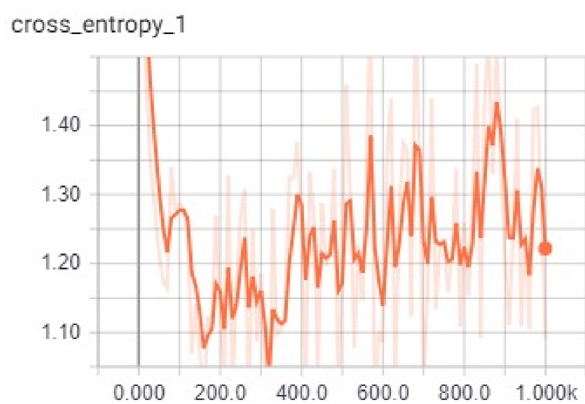


Figure 13: Validation Cross Entropy Dataset 5

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

VI. CONCLUSIONS

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

VII. 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)