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— Machine learning, Big Data, Visually Impaired, Natural Language, Image, Question and Answer

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 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

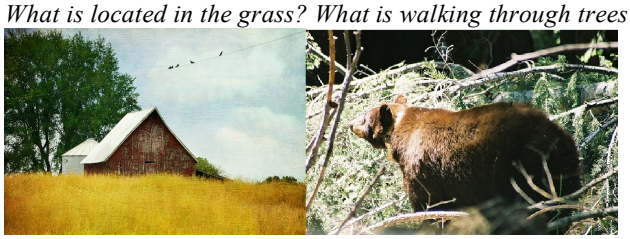


Figure 1: Examples of free-form, open-ended questions collected form MS COCO Dataset

II. RELATED WORKS

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. PROPOSED WORK

We have used Microsoft Common Objects in Context (MS COCO) dataset which has around 83k training images, we are focusing only on one context of images. We are specifically using Park images from this dataset in order to maintain similarity among our datasets while reducing the overall size of the datasets we are processing. Based on different keywords that represents park context such as bench, fountain, park etc. are used for data collection task.

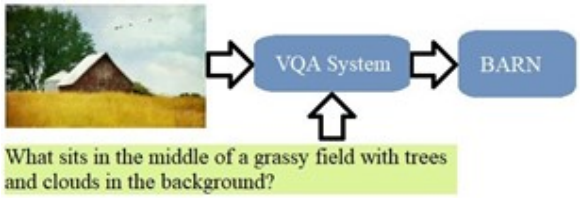


Figure 2: VQA Architecture Overview

IV. Dataset Analysis

Below is the analysis of VQA questions and answers in training dataset. The dataset contains 443,757 questions and 4,443,570 answers for a total of 83k images from MSCOCO dataset.

A. Questions

All the questions are open ended questions with corresponding Image ids and questions ids associated with it. The questions included Why, What and How kind of questions.


```
"image_id": "144", "question": "What are the animals doing?", "question_id": 144000
```


B. Answers

There are 10 answers associated with each question along with answer confidence as Yes and maybe.


```
"question_type": "what are the",
"multiple_choice_answer": "eating", "answers":
[{"answer": "drinking", "answer_confidence":
"yes", "answer_id": 1
```


Each of the 5 members of our team, has filtered out a subset of the VQA dataset using keywords related to parks and split that subset into training and validation parts. The question, answer, and image ID information is stored in json file during pre-processing.

V. System Architecture

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

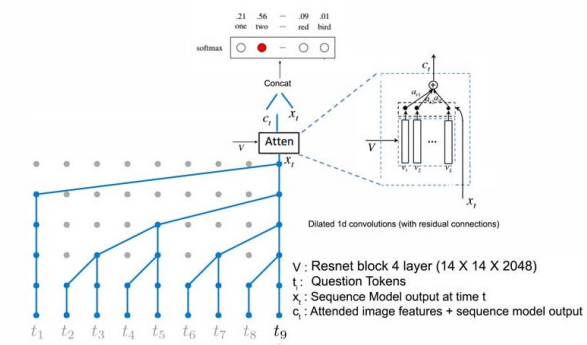


Figure 3: Overview of Visual Question Answering System with different components

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.

A. Models

We are using two pretrained models such as VGG16 and ResNet for feature extractions at image level. Our system extracts the features of the images in our 5 data-subsets using

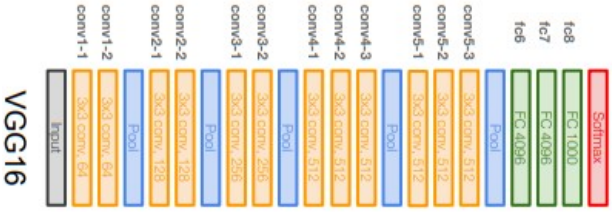


Figure 4: Architecture of VGG16

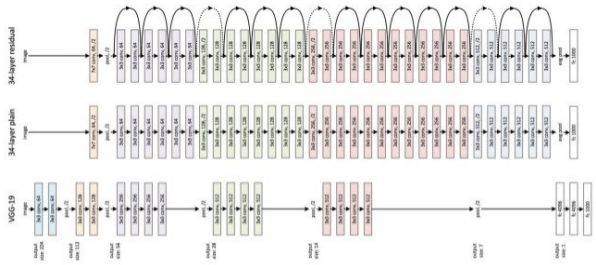


Figure 5: Architecture of ResNet

these two pretrained models, from the 5th pooling layer of VGG16, and from the 4th block of ResNet.

B. Attention

In order to improve the results of our VQA system we have included an attention model. This model generates two probabilistic attention maps per image using the encoded question and the image features. These attention maps are intended to focus the prediction model on relevant portions of the image as the answer is generated. After each training step the VQA training system generates a new model and, in this process, new attention maps are generated.

C. Implementation

The basic steps involved in training our VQA models are as follows:

- Pre-process data
  - Filter Images
  - Filter Questions and Answers
  - Embed Question and Answer vocabulary
- Extract Image Features
- Training

We have already discussed the data pre-processing and image feature extraction steps of this process. The training step is complex and involves integration of the extracted features, the question and answer embeddings, and the attention model.

First a model is initialized. Then the system trains for a number of epochs specified by the user. We have used different hyperparameters, including epoch, for each of our 5 datasets. Algorithm 1 shows a simplified pseudo-code of the algorithm used for training:

Algorithm1

```
model = VQA_Attention_model(model_opts)
model.build_model()
t = tf.train.optimizer(lrn_r).minimize(model.loss)
step = 0
for epoch in range(epochs):
    batch_no = 0
    while(batch_no*batch_size)<leng(data_train):
        q, a, img_f, img_id = get_batch()
        loss_val = sess.run(t, model.loss,
          feed_dict = {
            model.q = q
            model.img_f = img_f
            model.a = a
          }
        )
    if step % evaluate_every == 0:
        evaluate_model()
```

Training is defined to be minimizing the loss function (cross entropy) using the optimizer; for each dataset we have used the Adam Optimizer. Elsewhere in the program the training and validation sets are separated into batches, and once training begins, for each epoch we load every batch and train on that batch, updating the model. Within the initial model construction and during reconstruction during training the questions are encoded using ByteNet LSTM and the attention mechanism creates attention maps. After training a step our system checks if we have specified to evaluate at that step and if so the entire validation set is evaluated and training and validation accuracies are stored for that model.

VI. Results and Evaluation

Table 1 shows the Results obtained when trained on VQA model. While the accuracy scores shown in this table indicate that more data can lead to better results, it is not a guarantee.

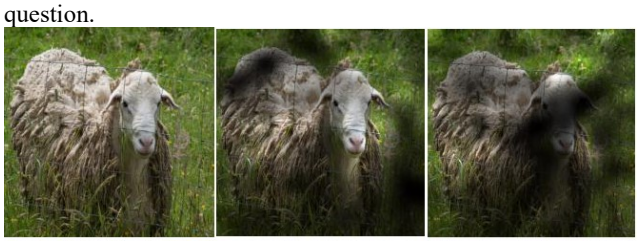
Dataset	Training Images	Validation Images	Validation Accuracy
I.	450	150	VGG16 0.378753 ResNet 0.392185
II.	300	100	VGG16 0.298762 ResNet 0.423127
III.	30	9	ResNet 0.201550
IV.	62	14	ResNet 0.228571
V.	1233	308	VGG16 0.366881 ResNet 0.364301

Table 1: Evaluation Results based on Dataset based on Park Images

A. Qualitative Results

We now visualize answers generated by our VQA model and how attention maps help us to focus on some specific areas of images. Below figure shows some examples we generated from model along with the answer and images with attention maps.

Our model pays attention to words in questions as “How Many”, color of image and then focuses attention on those parts of images which helps in answering the corresponding question.



Question: Is the sheep caged?
Predicted Answer: yes
Ground-truth Answer: yes

Question: How many horses are in picture?
Predicted Answer: 2
Ground-truth Answer: 2

Question: What is the color of the bench
Predicted Answer: brown
Ground truth answer: brown

Question: what jump up to catch the frisbee
Predicted Answer: Old
Ground-truth Answer: Dog

Question: How many slices have been taken?
Predicted Answer: ‘yes’
Ground-truth Answer: 1

Question: "Is someone pitching the ball to him?"

Predicted Answer: 2

Ground-truth Answer: no

Question: What color is the cat?

Predicted Answer: black

Ground-truth Answer: black and white

Figures 6-12: Visualization of image and question co-attention maps on the MSCOCO dataset. From left to right: original image and question pairs, word level co-attention maps and question level co-attention maps

These examples are taken from all models. While in some of these examples it appears that the attention model is focusing on relevant information—for example in the case of the dog catching the frisbee, or the slice of missing pizza--in other cases it is not at all clear what, if anything the model is actually focusing on. We have included a variety of examples including examples for which the answer is clearly wrong, for which the answer is clearly correct, and examples for which the predicted answer is close, but not entirely accurate; in the case of the cat the VQA model is able to identify that the cat is black, but does not give a full description of the cat's color. However hopefully these examples illustrate, for better or worse, the quality of the VQA model we have developed.

Additionally, we have used the Bleu machine translation metric to check the accuracy. The idea is that the Bleu metric will check for partial matching and that the accuracy should be better, however this is not what we saw in practice as shown in Table 2. Thus far we are still unsure of why what Bleu-1 score's accuracy is less than the string-matching accuracy, in principle it should be at least the same. Another mechanism by which we could likely improve the numeric representation of our calculated predation accuracy would be to compare the predicted answer with each of the ten ground truth answers in the VQA dataset rather than just the 'best' answer, in case our model predicts some answer which is not that best answer.

Dataset	Bleu-C	Bleu-1	Bleu-2	Bleu-3	Bleu-4
V.	4.16e-232	0.2273	5.10e-309	5.10e-309	5.10e-309

Table 2: Scores from Bleu machine translation metrics

VII. CONCLUSION

While the results we have achieved with our VQA system do not meet the state-of-the-art accuracy of VQA models, our results are encouraging considering the small size of the datasets. There is a myriad of applications for visual question answering technologies, from aiding the visually impaired in understanding a scene to generation of large-scale metadata about large image datasets which could be used for further machine learning applications. While the VQA system we have developed is not highly-accurate, it can be trained very quickly on a relatively small amount of data and yield impressive results.

We intend to expand on this research by improving the datasets, tuning hyperparameters, and possibly making changes to the attention model and/or loss function. Additionally, we intend to evaluate with traditional machine translation metrics as answers can be up to 5 words long and hence machine translation metric comparisons are suitable. Without any modifications to our system though, we have demonstrated the power of emerging machine learning algorithms which can be used for Visual Question Answering. Our visual question answering system, given an image and an open-ended natural language question about that image, can, roughly 1/3 of the time, correctly answer that question for images it was not trained on, an impressive thing for a computer to learn to do in the few hours of training that our models take.

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

APPENDIX

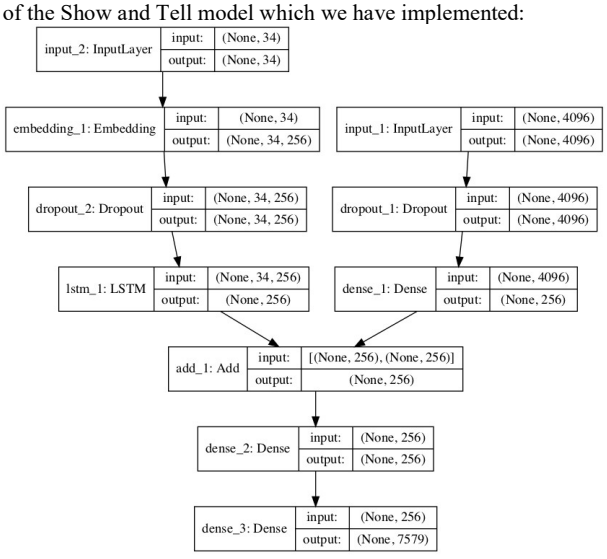
Datasets:

I. Priyanka Gaikwad
II. Sai Sri Narne
III. Sai Charan Kottapalli
IV. Sai Srinivas Vidiyala
V. Gregory Brown

Implementation of Show and Tell Model:

NB: During the construction of the VQA model for this project, there was a point at which we were considering changing to a caption generation model because of difficulties in getting the VQA model working. Thus, for three of our datasets Show and Tell style models were constructed. The information about the construction of those models is included below.

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 see 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
III	0.962306	0.947157	0.931470	0.907180
IV	0.979792	0.971377	0.959821	0.943118
V	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.