Visual Question Answering

Avni Mehta, Trinadha Raji Muppala, Harshil Patel, Santosh Jada
University of Missouri-Kansas City, USA
{amh8c, trm35c, hlp2k5, sjzkt}@mail.umkc.edu

***Abstract*—Visual Question Answering is a multi-disciplinary task and where the challenge is to predict an answer of a given question related to an image. We have selected a subset of MSCOCO dataset for the home interiors domain. These images are classified into four categories viz. bathroom, bedroom, kitchen, and living room. This paper compares the Image Classification performance of various shallow learning and deep learning methods for our dataset. Through experimental evaluation of various models, we show that the Inception model outperforms Spark and Clarifai models. A web application is built to enable the user to select an image and perform image classification using Spark, Clarifai and Inception model. The application is able to answer "Is there…", "Where is…" and "How many…" types of questions using Mask R-CNN and Yolo-Darknet models.**

***Index Terms*—visual question answering, bounding box, mask r-cnn, darknet, inception, tensorflow, clarifai, spark, performance, ms coco dataset, home interiors**

I. Introduction

THE motivation for undertaking VQA project was primarily our interest in doing a challenging project in multi-discipline Artificial Intelligence. In this present tech era, with AI research and computer vision on demand, we want to work on images to train the machine, see its analyzing capabilities and improve its accuracy.

The goal of this project is to implement the visual question answering system using open source neural network frameworks, Darknet [1] and Mask R-CNN [12]. The system would draw a bounding box in the image for the given "Where is…" type of question. The system is also able to answer "Is there…" and "How many…" type of questions using Mask R-CNN. The paper evaluates the performance of various image classification techniques including Spark (Shallow Learning), Clarifai, and Tensorflow Inception model (Deep Learning). The dataset used in the project is a subset of MS COCO dataset for home interiors domain.

A web application is built to enable the user to select an image and perform image classification using Spark, Clarifai and Inception model. The web application also allows the user to ask the system to draw a bounding box around a given object. Through this user interface, the user can upload any image and ask a question about the image. The system would predict the answer using Mask R-CNN model.

In this paper, we share our experiences on implementing various image classification methods. In Section II, we list projects with similar aims like VQA. In Section III, we outline our approach. In Section IV, we describe how to implement a visual question answering system on web as well as on a smartphone. In Section V, we present the dataset and performance evaluation results over various models. Finally, we conclude the paper in Section VI.

II. Related Work

VQA poses a rich set of challenges, many of which have been viewed as the holy grail of automatic image understanding and AI in general. However, it includes as building blocks several components that the CV, NLP, and KR communities have made significant progress on during the past few decades [2].

There have been several papers that are working on tasks of image tagging like [3, 4], image captioning [5, 6] and text-based Q&A [7, 8]. A team from MIT has worked on 'Simple Baseline for Visual Question Answering' [9]. The Show and Tell model also uses the COCO dataset for the same task [10].

However, there are no projects that are focusing on the interiors domain; this makes our application a one of a kind.

III. Approach

In this section, we outline the data sources, analytic algorithms, tasks, platforms, expected input/output and evaluation of different models.

A. Data Sources

We have collected a subset of home interior images from MS COCO dataset [11]. The dataset consists of images, questions and answers.

The images are classified into four categories viz. bathroom, bedroom, kitchen, and living room.

For shallow learning, we trained the models using a small dataset consisting of 50 training images per category and 5 test images per category. For deep learning, we trained the models on a dataset consisting of 150 training images per category and 20 test images per category.

B. Analytic Algorithms

The following algorithms were used for image classification:

- Shallow Learning – Random Forest, Decision Tree, Naïve Bayes using Spark API
- Deep Learning – Softmax, CNN, Inception model using Flask API
- Other Deep Learning (Fuzzy classification) - Clarifai and TensorFlow Lite. Our system can take the results from Clarifai model and classify the image into one of the four categories.

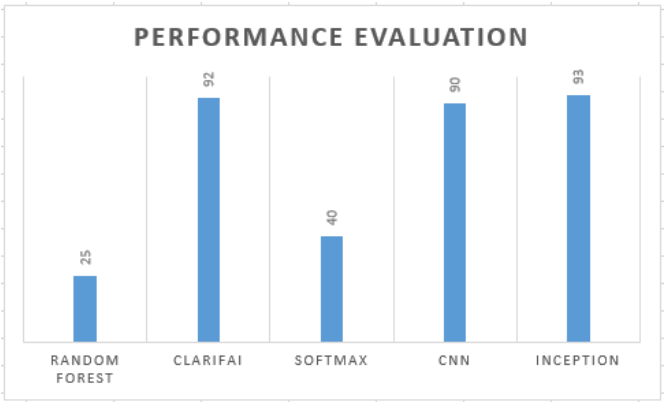

Fig. 1. Performance evaluation of various shallow learning and deep learning models on the home interiors dataset. Horizontal axis represents the model and the vertical axis represents the accuracy of the model.

C. *Tasks*

The following tasks were performed:

- Building Random Forest, Decision Tree, Naïve Bayes models for image classification
- Fuzzy image classification using Clarifai API
- Converting the project data to MNIST format for Softmax and CNN models
- Building Softmax, CNN and Inception models for image classification
- Building the darknet model for Bounding Box
- Building the VQA model using Mask R-CNN

D. *Platforms*

The following platforms were used:

- IntelliJ for shallow learning models
- PyCharm for deep learning models
- Android Studio for android application
- Webstorm for web application

E. *Expected Input / Output*

- Input: Image from a dataset of home interiors. Additional input of a question for the bounding box
- Output: Classification of image into one of the four categories. A bounding box for the given question

F. *Evaluation*

Fig. 1. shows the performance evaluation of the following models:

- Random Forest (Spark): 25% Accuracy
- Clarifai: 92% Accuracy
- Softmax: 40% Accuracy
- CNN: 90% Accuracy
- Inception: 93% Accuracy

IV. Implementation Details

In this section, we describe implementation details including software architecture diagram, activity diagram and sequence diagram.

A. *Software Architecture Diagram*

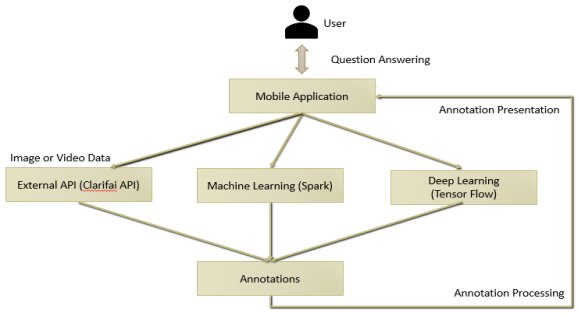

Fig. 2. Proposed architecture diagram

B. *Activity Diagram*

Fig. 3. Activity diagram for the proposed model

C. *Sequence Diagram*

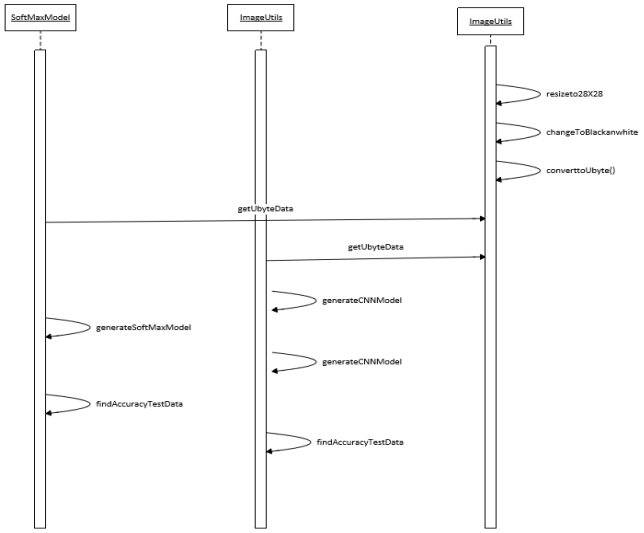

Fig. 4. Sequence diagram for the proposed model

D. *Feature Specification*

The following list includes all the features of our application:

i. Image Classification using Random Forest model using Spark API
ii. Image Classification using Clarifai API
iii. Image Classification Inception model using Flask API
iv. Visual Question Answering via a bounding box
v. Visual Question Answering via Mask R-CNN
vi. Object Detection using Tensorflow Lite
vii. A web application and an android application

E. *Existing Application/Services Used*

The following existing services were used for our application:

i. Flask API
ii. Clarifai API
iii. Inception model
iv. Darknet for bounding box
v. Mask R-CNN for VQA

V. Results

In this section, we present the dataset and performance evaluation results over various models.

A. Dataset

We have collected a subset of home interior images from MS COCO dataset [11]. The dataset consists of images, questions and answers. The images are classified into four categories viz. bathroom, bedroom, kitchen, and living room.

For shallow learning, we trained the models using a small dataset consisting of 50 training images per category and 5 test images per category. For deep learning, we trained the models on a dataset consisting of 150 training images per category and 20 test images per category.

B. Evaluation

The following table lists the performance for all models in the experimentation:

TABLE I
PERFORMANCE EVALUATION FOR SHALLOW AND DEEP LEARNING MODELS

Model	Accuracy
Random Forest	25%
Decision Tree	15%
Naïve Bayes	15%
Clarifai	92%
Softmax	40%
CNN (Adam)	90%
CNN (Adagrad)	43%
Inception	93%

Performance evaluation of various shallow learning and deep learning models on the home interiors dataset.

The following figures describe Zenhub boards and burndown charts:

i. Zenhub Board

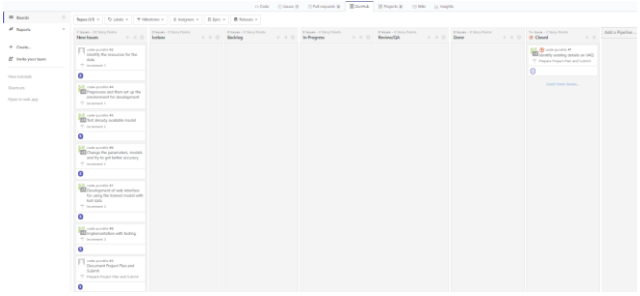

Fig. 5. Zenhub Board

ii. Milestones

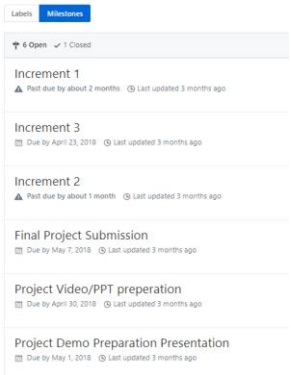

Fig. 6. Zenhub milestones

iii. Burndown Chart

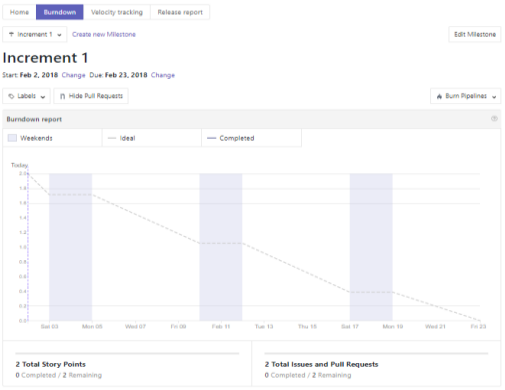

Fig. 7. Zenhub burndown chart

The following are the screenshots of our application:

i. Web Application:

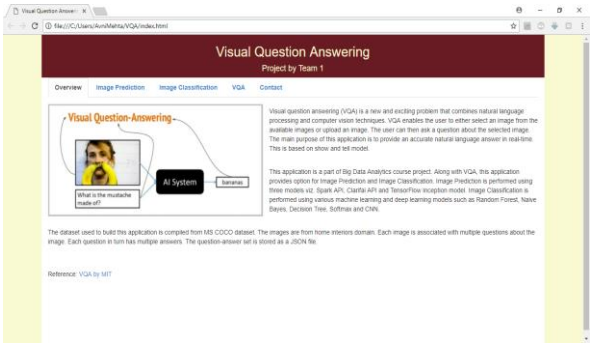

Fig. 8. Web Application - Overview

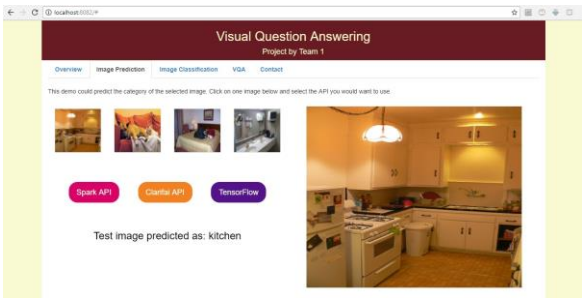

Fig. 9. Web Application – Image Prediction

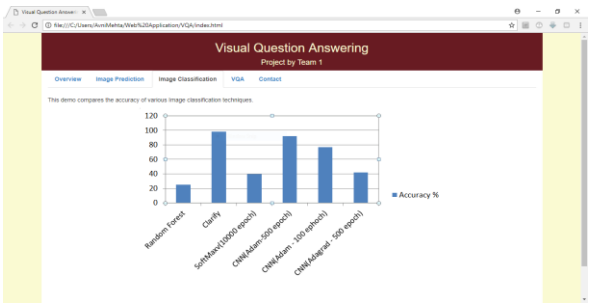

Fig. 10. Web Application – Evaluation

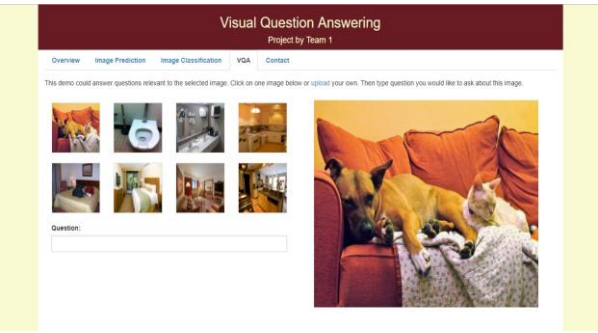

Fig. 11. Web Application – VQA

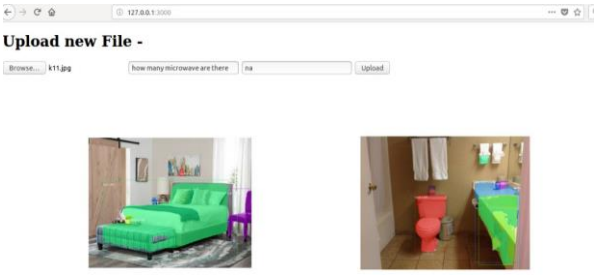

Fig. 12. Visual Question Answering using Mask R-CNN

Fig. 13. Bounding box for the question "Where is the dog?"

VI. Conclusion

In this paper we have presented experimental results on various methods for Image Classification. VQA is achieved by the means of segmentation and bounding box around the given object. VQA is achieved for the following types of questions: "Where is…?" via. bounding box using Darknet model, "How many…?" using Mask R-CNN model, "Is there…?" using Mask R-CNN model.

Implementing a pure android application for our VQA model was a challenge for us. Another challenge was to train the existing VQA model developed by MIT.

As future work, we plan to build a VQA system that predicts the answer to other types of question ("What is…", "Does the…") related to the image.

ACKNOLEDGEMENT

The work has been completed under the guidance of Dr. Yugi Lee, Mayanka Chandra Shekar, and TAs (Megha Nagabhushan, Sidrah Junaid) of CS5542 Big Data Analytics and Apps, University of Missouri - Kansas City), Spring 2018

REFERENCES

[1] Joseph Redmon. Darknet: Open Source Neural Networks in C. http://pjreddie.com/darknet/ 2016
[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. arXiv preprint arXiv:1505.00468, 2015.
[3] J. Deng, A. C. Berg, and L. Fei-Fei. Hierarchical Semantic Indexing for Large Scale Image Retrieval. In CVPR, 2011
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classifi- cation with Deep Convolutional Neural Networks. In NIPS, 2012
[5] 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.
[6] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences for Images. In ECCV, 2010
[7] A. Fader, L. Zettlemoyer, and O. Etzioni. Open Question Answering over Curated and Extracted Knowledge Bases. In International Conference on Knowledge Discovery and Data Mining, 2014
[8] A. Fader, L. Zettlemoyer, and O. Etzioni. Paraphrase-Driven Learning for Open Question Answering. In ACL, 2013
[9] B. Zhou, Y. Tian, S. Suhkbaatar, A. Szlam, R. Fergus. Simple Baseline for Visual Question Answering. arXiv:1512.02167
[10] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. Show and Tell. arXiv:1609.06647, 2016.
[11] Tsung-Yi Lin et. al. Microsoft COCO: Common Objects in Context. arXiv:1405.0312v3, 2015
[12] https://research.fb.com/wp-content/uploads/2017/08/maskrcnn.pdf