Deep Vision Tool

by

Ashish Rana
Sagar Shivani
Shaunak Dixit
Yuvraj Verma
CPG-14

1.Project Overview

This work aims to Construction of a Deep Vision Tool which consists of 4 modules: - Visual Question Answering Tool, Image Caption Tool, Depth Analysis Tool and a Natural Language Processing Chatbot. All the modules will be embedded into an offline GUI application.

The VQA Tool will address the problem of image-based question-answering (QA) with new models and datasets. Neural network architectures with memory and attention mechanisms exhibit certain reasoning capabilities required for question answering. One such architecture is VGG Convolution Neural Network architecture, through which an image vectors can be extracted. A LSTM model or an dynamic neural network along with GLoVE Vectors of Words may be used for answering any question about the image. We propose to use a large dataset that contains 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 are effective at eliciting compelling and diverse questions.

The Image Captioning Tool will generate caption by combining objects and their relationship in an image. A Deep learning tool like Convolution Neural network along with Recurrent Neural Network model is proposed to be used to create image captions for a particular image. In this work MSCOCO image captioning dataset is used. MSCOCO is a dataset developed by Microsoft with the goal of achieving the state-of-the-art in object recognition and captioning task. This dataset contains collection of day-to-day activity with their related captions. It contains around 2.5 million objects in 328K images. Dataset is created by using crowdsourcing by thousands of humans.

The Depth Classification module will calculate the distances between the intermediate images through conversion of an RGB image into an RGBD image by image processing through neural networks. Existing datasets will be converted into RGBD dataset and neural network will be trained on them. Last but not the Least, the NLP chatbot will be constructed using sequence to sequence models and will act as an interaction between the user and software.

2. Need Analysis

Our project is inspired by the increasing needs of intelligent AI in every daily-use appliance and also needs of physically challenged people. Some uses have been stated below:

1. Image Features Extraction

2. Educational Software for Children

3. Use in military for exploring the maps

4. Can be developed for blind people for visualization of the environment Our project has a great potential and can serve as an introductory step towards great applications.

A wide range of visual question answering has been developed and some are also used by common people. Google re-captcha is one of the VQA tool where end user has to identify objects in the given image. Visual question answering combines computer vision techniques and natural language processing. We need various datasets and models for VQA. Firstly, Objects in the Image gets compared with the datasets and then some models are applied on it which are classified into four types: non-deep learning models, deep learning models without attention, deep learning models with attention, and other models which do not fit into the first three. Then the performance is calculated and best performance gives the output.

In today's life VQA is an exciting and latest technology which will ease the job of human beings. There are many educational games for kids which let the kid answer some simple question like identifying the colour of some ball or kite which is same from last century.

Every country wants to expand its military with intelligence and strength. There are huge maps and sometimes it becomes too tedious and chaotic to search the maps. This VQA tool will help to get the desired details in no time which will save a lot of precious time of the military.

There are many speech-to-text and text-to-speech applications which can be used by the blinds to operate a machine but what if a blind wants to look at the details of the image. This VQA tool will provide the complete details of an image with an interactive chatbot to answer users query.There will be different answers for same question for similar images, the objects will be identified from the questions and then answered by the help of machine learning.

3.Literature Survey

1) [10] VQA: Visual Question Answering the task of free-form and open-ended Visual Question Answering (VQA) is by giving an image and a natural language question about the image, the task is to provide an accurate natural language answer **Pros:** In this paper proposed combining an LSTM for the question with a CNN for the image to generate an answer a similar model is evaluated in this paper. **Cons:** Accuracy 54.06%

2) [1] Dynamic Memory Networks for Visual and Textual Question Answering We have proposed new modules for the DMN framework to achieve strong results without supervision of supporting facts. These improvements include the input fusion layer to allow interactions between input facts and a novel attention based GRU that allows for logical reasoning over ordered inputs. **Pros:** architecture, the dynamic memory network(DMN). **Cons:** Accuracy 60.4%

3) [9] Generative Adversarial Text to Image Synthesis In this work we are interested in translating text in the form of single-sentence human-written descriptions directly into image pixels. **Pros:** develop a novel deep architecture and GAN formulation to effectively bridge these advances in text and image modelling, translating visual concepts from characters to pixels. **Cons:** On Close inspection it is clear that the generated scenes are not usually coherent;

4) [11] Estimated Depth Map Helps Image Classification Therefore, we present a way of transferring domain knowledge on depth estimation to a separate image classification task over a disjoint set of train, and test data. **Pros:** 2-layer feed-forward neural network yielded a performance increase of 4%, when comparing a NN trained on the RGB dataset to the NN trained on the RGBD dataset. **Cons:** We get 56% and 52% validation accuracy with RGBD and RGB dataset respectively.

4. Objectives

- Higher accuracy and better visualizations for VQA.
- Real Time Image Context Recognition of the image.
- Depth analysis and classification of the image.
- Sequence to Sequence model chatbot for demonstrative purpose.
- Open source deployment of such module and integrated tool.

5. Methodology

The project is divided into four major modules handling each case related to image analysis in project separately with seperated graphical user interface. After this a combined desktop application will be there doing all the work at once involving depth analysis, visual question answering and activity classification. An extra module of NLP chatbot is there as the mechanism used in this chatbot is underlying logic behind the activity classification or caption generating module as for this reason it used as demonstrative module in this project.

5.1 Visual Question Answering Module

Aim of this module is to provide an independent GUI which will be answer objective questions related to uploaded image for analysis. Deep Learning model Convolution Neural Network will be used in this module along with MS-COCO dataset for training purposes. Already such image analysis models exists, our aim is to find a model that gives better accuracy plus a better visualization of data with tensorflow library. With mechanism like pooling, rectifications, filtering and stacking of neural layers multiple times this can be achieved but looking for an optimum solution is an answer. GUI coding will be done on tkinter library in python. Here in the following figure explaining working of CNN is explained.

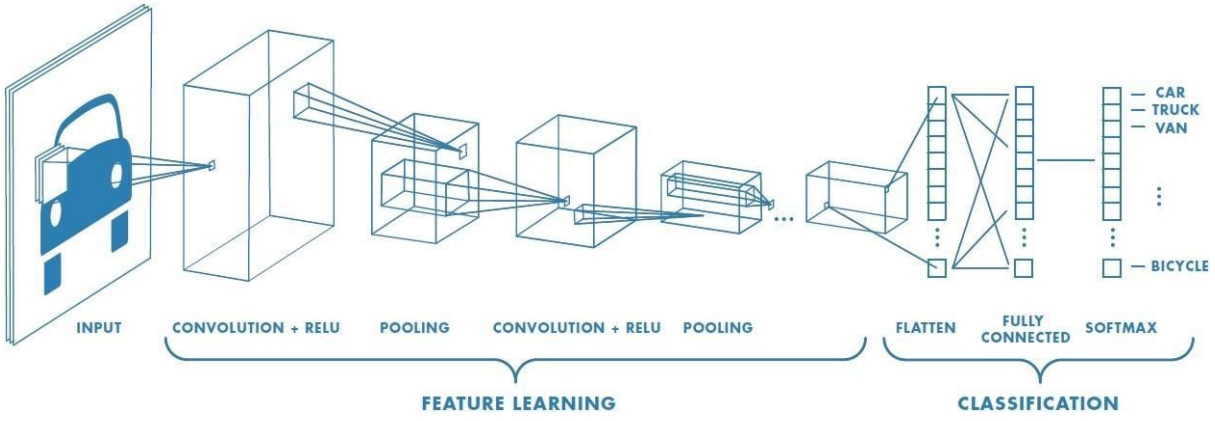


Fig 6.1.1: Basic working of CNN model.

5.2 Depth Analysis Module

Aim of this module to provide an independent GUI which will be to calculate depth or distance of objects in the following image for analysis basically it will be used for depth classification. The important part in this module is the dataset used which is RGBD dataset which involves extra channel for depth. This extra channel for depth will be used to CNN that was explained earlier in previous module. This extra channel of depth will give higher as compared simple RGB dataset for depth classification as more information for depth is encoded in it. Along with increased accuracy providing strong visualizations through this tool with tensorflow library. GUI coding will be done on Tkinter library in python. Current implementation are primitive in visualizations this GUI will definitely be aiming for an improvement.

5.3 NLP Chatbot module

This module is demonstrative module for highlighting the underlying mechanism of working of image captioning. This chatbot will be of end-to-end system type i.e. focusing on simple given model for obtaining desired results. For this module either Cornell Movie dataset or Reddit dataset will be used for encoding and decoding of questions and responses respectively. Seq-to-seq model will be used in which Recurrent Neural Networks will be used as building block for encoding and decoding the data. The decoding part will also be containing the LSTM(Long Short Term Memory) model for coherent responses. Here with this our module our aim will be provide basic demonstrative working structure of activity classification module in GUI form. The underling figure highlights working of seq-to-seq model.

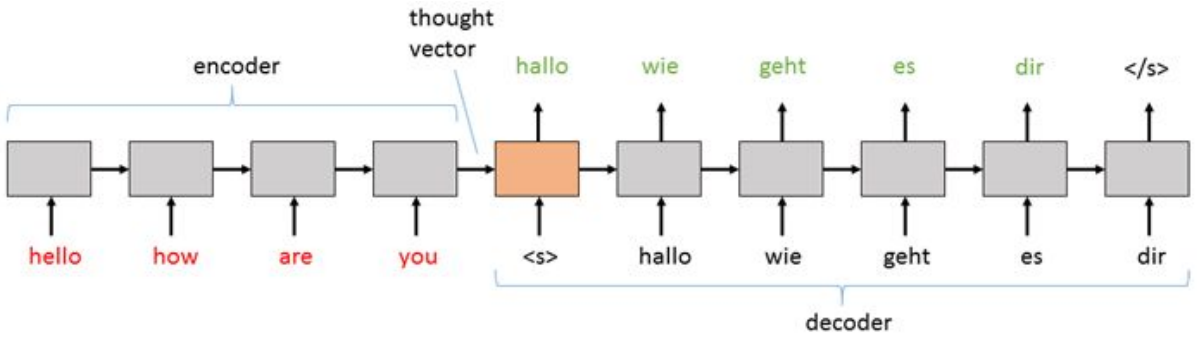


Fig 6.3.1: Basic encoding and decoding mechanism in sequence to sequence model.

5.4 Activity Classification Module

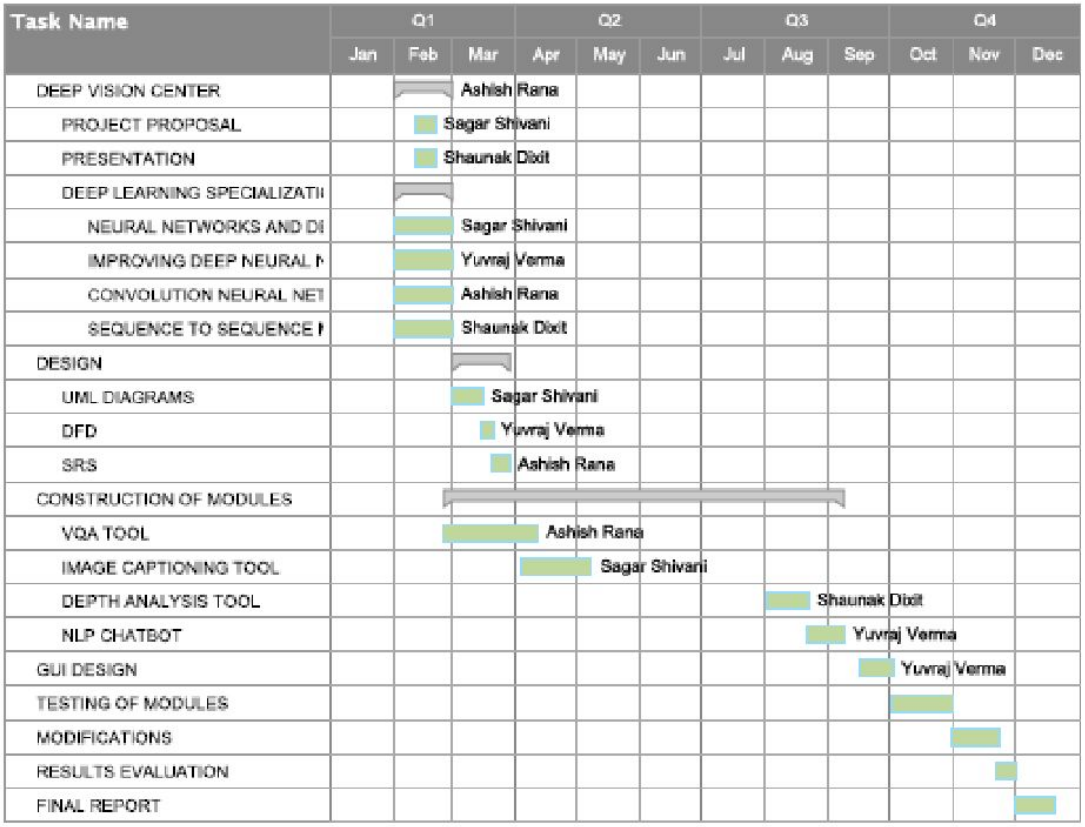
This GUI module will taking its basic functioning from the above mentioned module but now instead of questions being encoded image has to be encoded. Input as an image is provided to Inception-v3 model. At the end of Inception-v3 model, single fully connected layer is added. This layer will transform output of Inception-v3 model into word embedding vector. We input this word embedding vector into series of LSTM cells. LSTM cell provides ability to store and retrieve sequential information through time. The dataset used will be FLICKR dataset in this project for training the model. The tkinter GUI will be used for coding the GUI in python. The main aim of GUI will be providing captions and detecting activities along with it providing the visualizations for understanding.

5.5 Main Deep Vision Module

This module contains the three submodule namely VQA, depth classification and activity classification which will providing functionality cumulatively to this main module. This main module will providing visualizations for the results of the given image for analysis.

6. Work Plan

Deep Vision Center



7.1 Project Outcomes

Our project will be able to perform the following tasks:

1. Will be able to perform the depth analysis on an object in an image.
2. Will be able to fetch the context from an image.
3. Will be able to provide answers from the image as per questions.
4. Will be very helpful for the blind people as well as children
5. Will be able to enhance or alter any image according to user needs.
6. Will be helpful in predicting the next frame of a video.

7.2 Individual Roles

	Deep Learning	NLP Module	Documentation and Diagrams	Open Source	Testing and Optimizations
Ashish Rana 101690011	✓	✓	✓	✓	
Sagar Shivani 101512043	✓	✓	✓		✓
Shaunak Dixit 101562009	✓		✓	✓	✓
Yuvraj Verma 101512062		✓	✓	✓	✓

8. Course Subjects

Machine Learning (5th semester) And Deep learning(Convnets) (online)

Latex and Software engineering (5th semester)

Python, Django and TensorFlow(online)

9. Conclusion

This tool provides an efficient and easy way of extracting the much-needed features from images as modern day needs demand. With the concepts of deep learning the we will be able to automate tasks which were much harder to do earlier on manually like deciding captions, depth classification of terrains, etc. Along with it strong visualizations from libraries like TensorFlow will help in in depth analysis, object context analysis with vast variety of pluggable and deployment options available. Tool like this carries vast varieties of applications from forensics, defence, research, social media to interactive children education. With latest and efficient technologies this tool can solve conventional problems in image processing with proper analysis of spatial data with help of different neural networks incorporated in tool.

10. References

[1] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. arXiv preprint arXiv:1603.01417, 2016.

[2] Donahue, J., Hendricks, L.A., Guadarrama, S., et al.: 'Long-term recurrent convolutional networks for visual recognition and description'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015

[3] GloVe: Global Vectors for Word Representation. https://nlp.stanford.edu/projects/glove/.

[4] https://github.com/machrisaa/tensorflow-vgg?files=1

[5] Karpathy, A., and Fei-Fei, L.: 'Deep visual-semantic alignments for generating image descriptions'. The IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.. Szegedy, W. Liu, Y.

[7] Mansimov, E., Parisotto, E., Ba, J. L., and Salakhutdinov, R. Generating images from captions with attention. ICLR, 2016.

[8]M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Man´e, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Vi´egas, O. Vinyals, P. Warden, M.Wattenberg, M.Wicke, Y. Yu, and X. Zheng. Tensor-Flow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[9] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran Bernt Schiele, Honglak Lee. Generative Adversarial Text to Image Synthesis arXiv: 1605.05396v2 [cs.NE] 5 Jun 2016

[10] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Virginia Tech Microsoft Research, VQA: Visual Question Answering IEEE Explore, Year 2015

[11] Yihui He, Xi'an Jiaotong University, Xi'an, China Estimated Depth Map Helps Image Classification, arXiv:1709.07077v1 [cs.CV] 20 Sep 2017