**CSE 576 Natural Language Processing – Phase 1**

**Visual Question Answering**

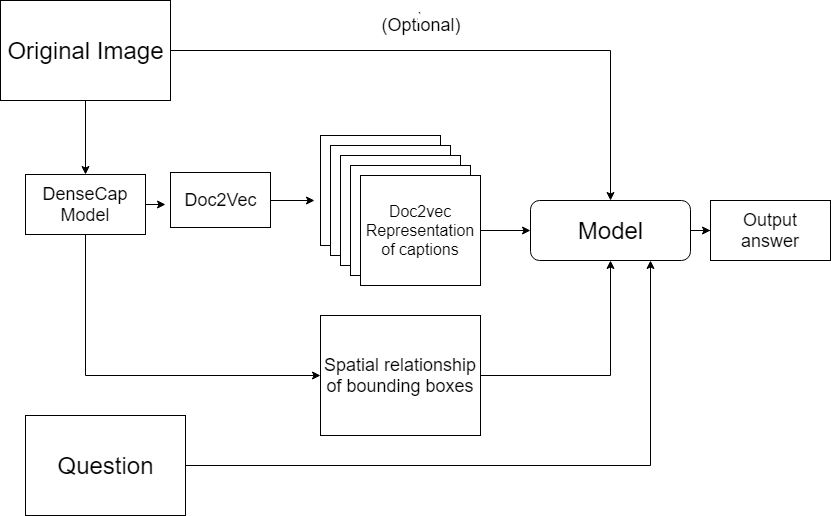
*Manjith Chakravarthy, Meredith Moore, Ravi Teja Pinnaka, Sai Veena Katta*

**Motivation:**

Object referring is important in visual question answering, as well as human robot interaction. We use semantically embedded references all the time, and inorder to be better able to interact with humans, the systems we use must understand how we communicate complex ideas like references.

**Problem Formulation:**

In this project we take the dense captions generated from the [] implementation on GitHub as the input for the network we are going to build using Convolutional Neural Network. As shown in the figure [num], we convert the captions generated by densecap into matrix representation using one of the document to vector tools available (doc2vec). With all the captions stacked together, the input becomes 3 dimensional. We also calculate the spatial relationship of the bounding boxes for which the captions are generated. We define the spatial relationship matrix as a 2D matrix with elements matrix[i][j] representing the distance between the bounding box i and bounding box j.

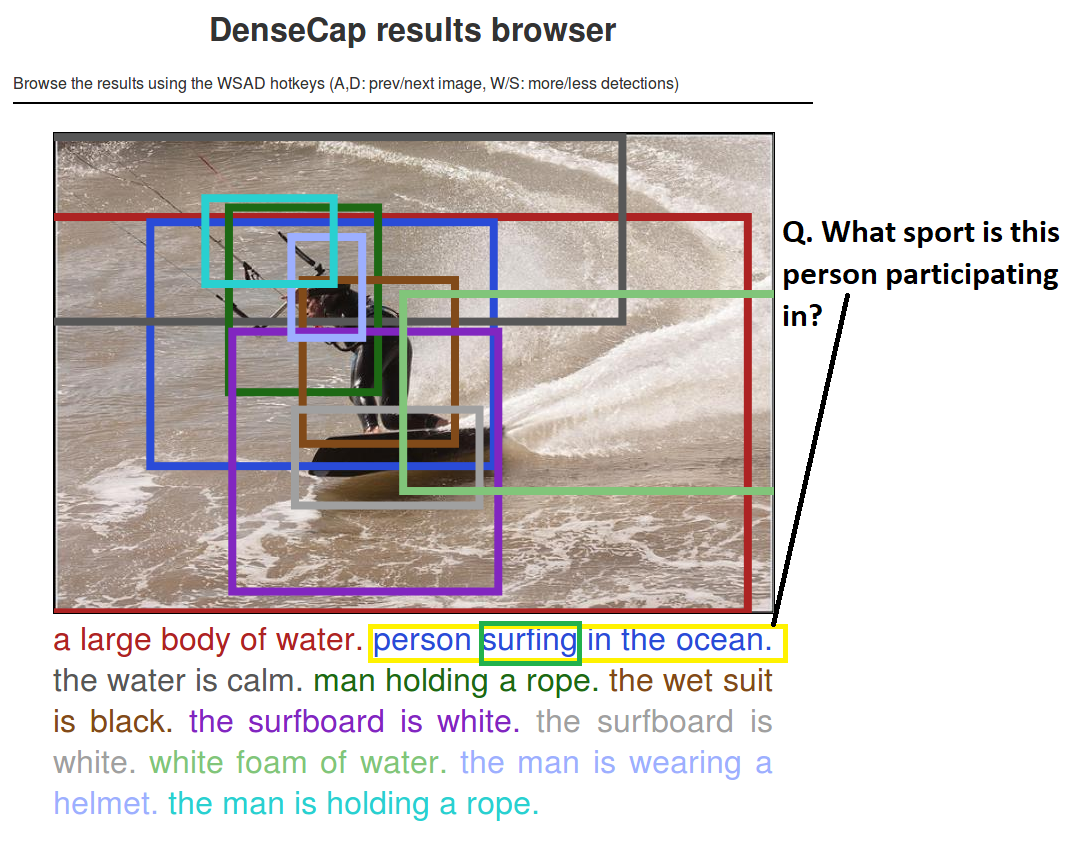
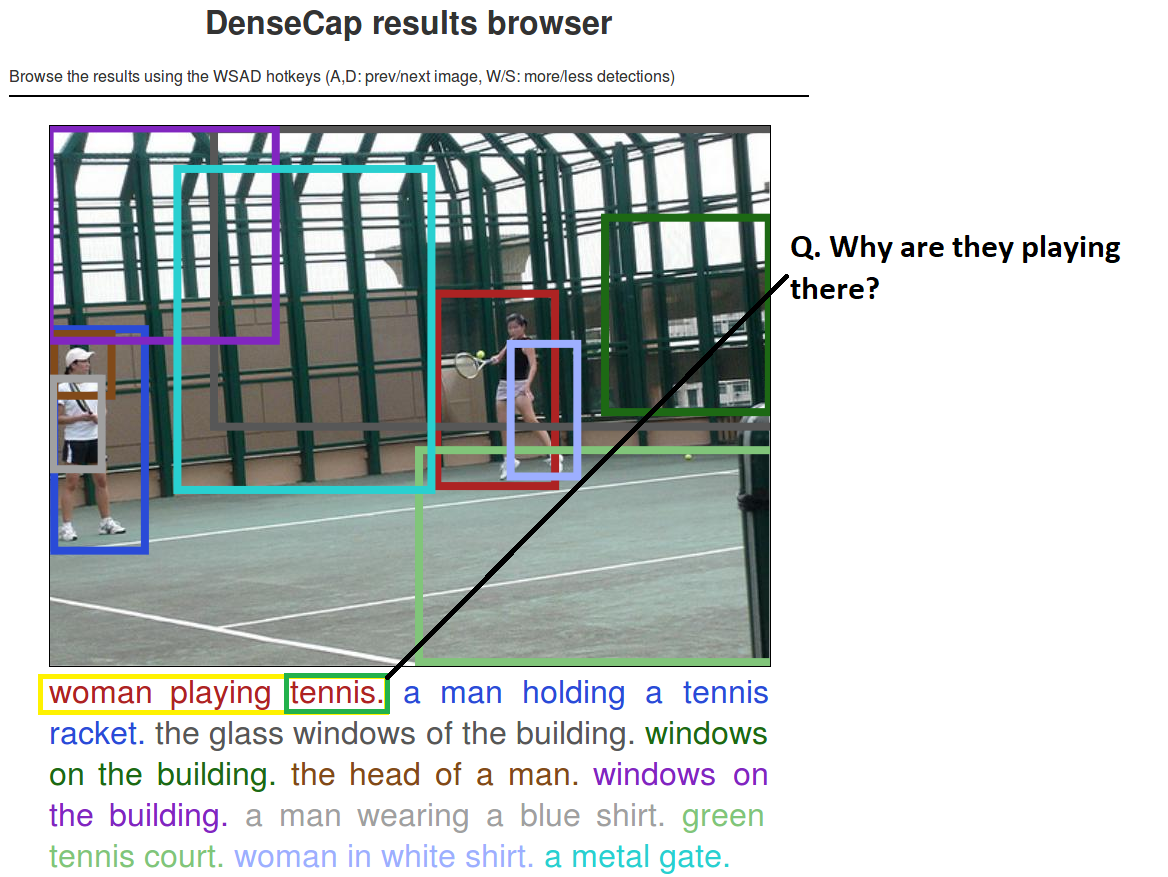
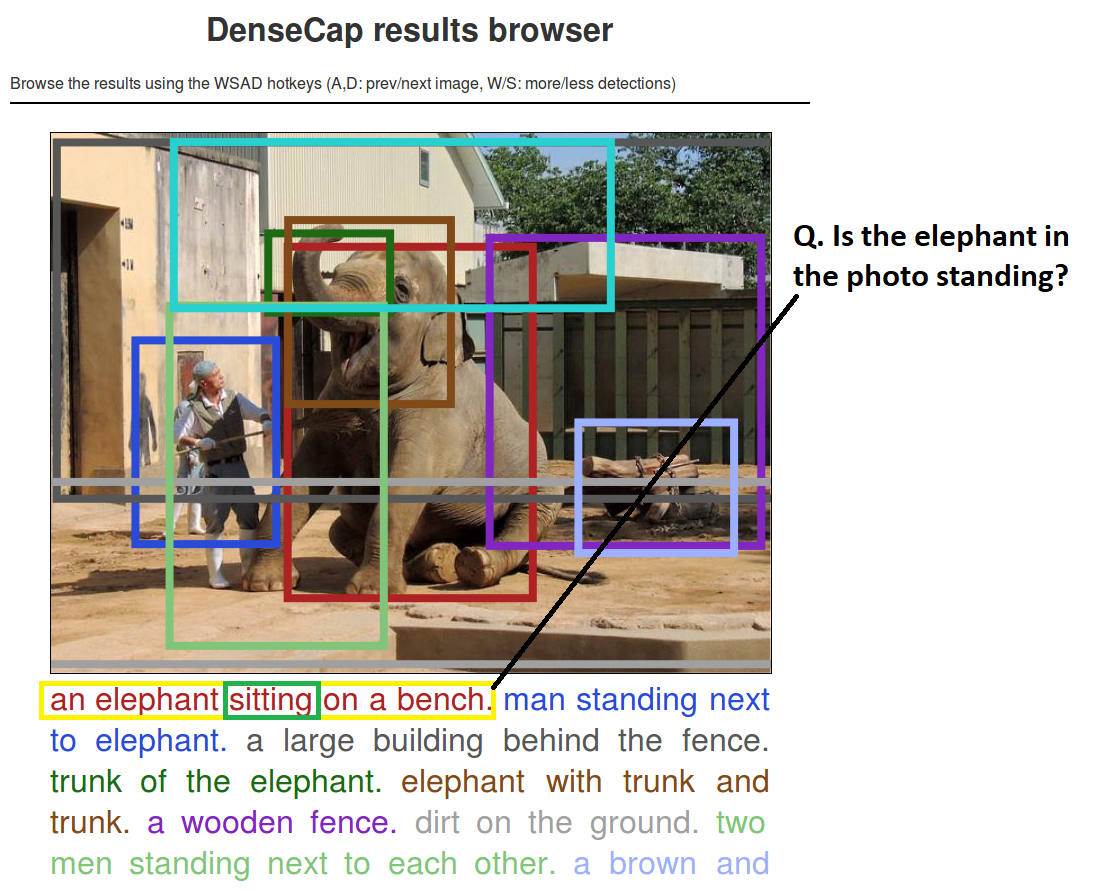
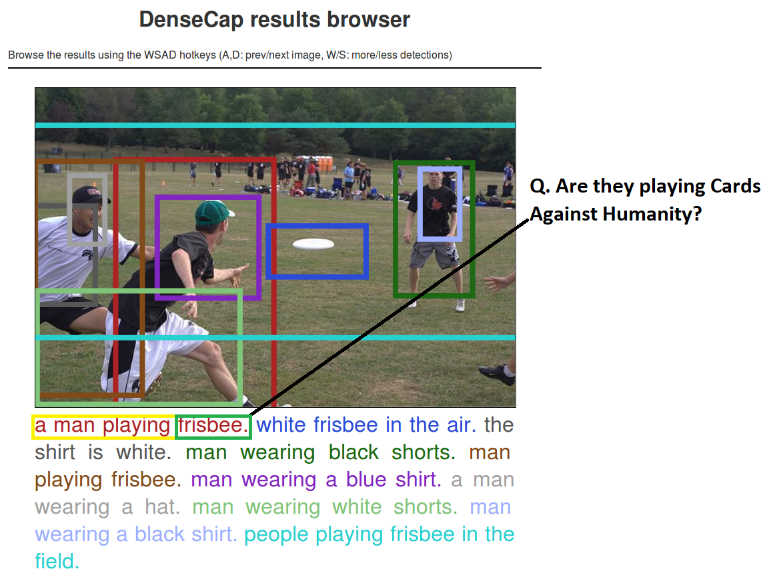
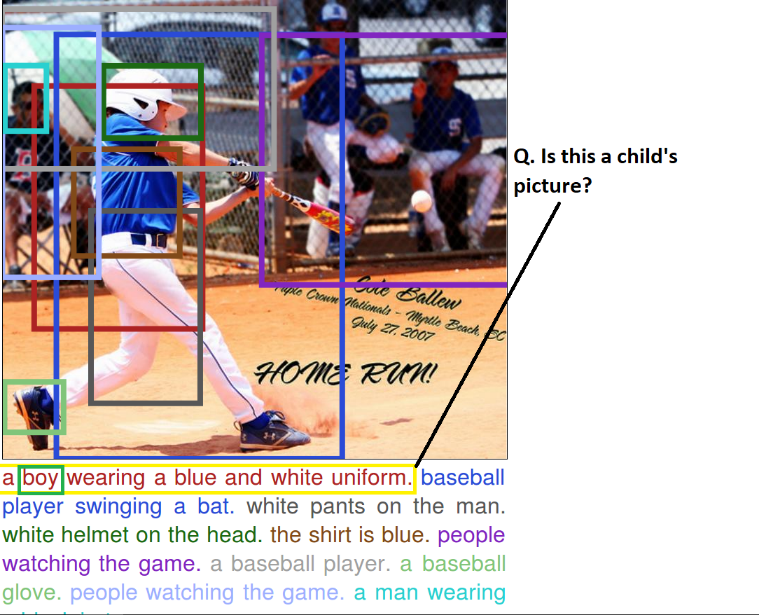
****

**Figure num**

The final output format is yet to be finalized (multiple choice or open ended).

**Examples:**

After setting up the dependencies and running the densecap code on a GPU for a few examples from VisualQA images (MSCOCO dataset), we found these to be interesting. From figures [num], [num], [num], [num] and [num], we found that for some of the questions the output can be predicted from just the captions.



**Training Dataset**

We plan to use the VQA dataset mentioned at [1]. The dataset consists of 123,287 training and validation images and 81,434 test images from the Microsoft COCO dataset. MS-COCO dataset is a rich collection of images with textual information on complex images with multiples objects in it. The dataset is a collection of open ended questions that will require a deeper understanding of computer vision, image analysis and knowledge to answer simple questions. The dataset consists of 265,016 images with at least 3 questions per image, 10 ground truth answers per question and 3 plausible answers. The questions are stored in JSON format with information on question, question type , multiple answers and answer accuracy metric. For splitting the dataset, we will use the MS-COCO train/test/val split strategy.

**Evaluation**

The evaluation of the dataset is similar to the process undertaken in the training the Machine Learning model. A k- fold validation model will be performed on the training dataset, training on k-1 parts and testing on the remaining part. This process is repeated for all the parts and the average accuracy is computed overall and returned as the accuracy of the model.

**Related work**

The expected accuracy with the standard word to vec model is limited. We would like to improve the accuracy of the model by changing the word to vector model to Long Short Term Memory(LSTM) units. LSTMs generally tend to give a better accuracy as they involve weighted memories of different words in the model.

If time permits, we would like to verify the accuracies of both the architectures- one with the word-to-vec and the other with LSTM.

**Timeline**

|  |  |  |
| --- | --- | --- |
| S.No | Details of the task | Expected Date of Completion |
| 1. | Deep Captioning | March 10th |
| 2. | Architecture Design for the model proposed | March 16th |
| 3. | Building the word-to-vec model | March 30th |
| 4. | Building a Convolution Neural Network Model | April 10th |
| 5. | Building a Feed Forward Multilayer network | April 20th |
| 6. | Documenting the results, report and presentation to the Professor | April 30th |

**Reference**

[1] VQA: Visual Question Answering, [www.visualqa.org](http://www.visualqa.org) - Aishwarya Agrawal∗, Jiasen Lu∗, Stanislaw Antol∗,Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh