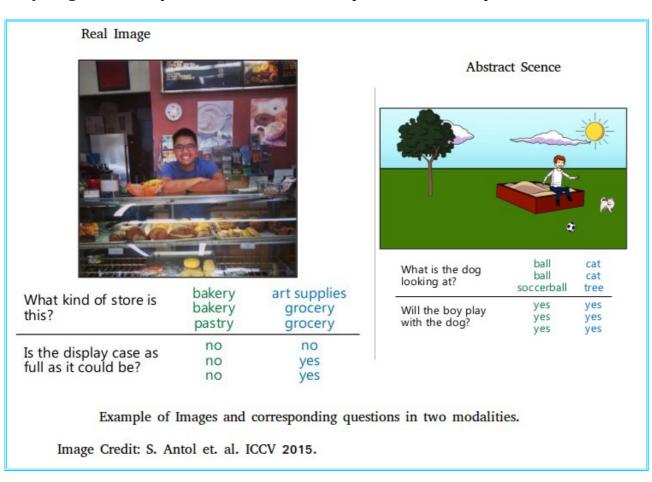
## Author: Abhishek Jha

## <u>Problem Statement</u>: Determining how many of the questions in the VQA dataset[1][2] are about common objects (e.g., bike, table, chair, person, toilet, etc.)?

VQA dataset which can be downloaded from VQA challenge webpage[1] consists of two datasets:

- 1.) **Real Images:** It has 123,287 training and validation images and 81,434 test images from the newly-released Microsoft Common Objects in Context (MS COCO)[10].
- 2.) **Abstract Scenes:** It consists of 50k scene of cliparts containing 20 "paperdolls", over 100 objects and 31 animals in various poses[6][7][8][9].

Since for the task to categorizing questions, I am will only discuss about the questions here and not other details like the answers and captions. Albeit, interested readers can find the more information regarding VQA dataset split and protocols adopted for framing questions and answers on [2]. For every image we have 3 questions, in two modalities: open-ended and multiple choice.



## Recipe:

- 1. First, download the training, validation and testing questions from the download page of [1], which are json files.
- 2. After extracting the questions, tokenize the sentence using nltk tokenizer[3].
- 3. Now use nltk POS tagger[3], and filter out all those words which are not proper noun and common noun. Since they are more likely to indicate which categories the question belong to
- 4. After this use WordNet lemmatizer [4] for converting the plurals to their singular forms. Still some of plurals will not be converted like 'men' will not be changed 'man' and so on. Hence we can use WordNet morphy [4] for this task.

Author: Abhishek Jha

- 5. Now the most important part: A category has to be a physical entity, an object. For example 'color' and 'show' are nouns but they are not common objects. They are abstract concepts. Therefore to filter out those wrods which no way be a common object we used the WordNet hierarchy[4].
- 6. The top most node of WordNet[4] hierarchy is 'Entity' which has 3 hyponyms: 'Abstraction', Physical Entity', and 'Things'. We are interested in only 'Physical Entity and that too its hyponym 'object' among its 6 hyponyms. We first took all the synsets(only noun synsets) of a word and looked up its hierarchy. If any of the sysnset has a hypernym 'object' then the word might be a category. This will further narrow down the number of common categories.
- 7. Finally, we use Sklearn[5] for creating the bag of words. In the code given I have taken top 200 or 100 most frequent words as the common categories.

Furthermore for representation you can make word clouds from <a href="http://worditout.com/">http://worditout.com/</a>

For Real Images, open-ended questions the wordcloud:

```
phone cow elephant vegetable right house vehicle kitchen giraffe wall window bird zebra guy proad bear dog road bear dog road bear dog road bear type man plate left water floor fire glass bus air top sport animal sign to bilding shape object background ball baby chair umbrella bench scene image computer orange beach motorcycle
```

For Abstract scene open-ended questions the wordcloud:

```
monkey pattern football sidewalk bottle glass
         football sidewalk bottle glass bucket front flower painting object
           stool fireplace curtain deer squirrel
      scene window couch person glide guy
                                              slide coat
  game bush chair picture floor shelf
   rack book
    doll shirt
                                      bird bench thing
   bar shirt man animal top animal top
leg log ball cat dog tv sofa rabbit tv sofa dogr side
                                                rug fish item
   air duck table
    pet table tree plant door side hair house
      foot pillow sun baby wall grill turtle apple pillow sun baby bike right
       apple pillow sun park hand blanket grass block skateboard ground skateboard park hand blanket grass block butterfly
                   scooter pant basket
bookshelf mushroom
watermelon
```

## **References:**

- [1] <a href="http://www.visualga.org">http://www.visualga.org</a>
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, "VQA: Visual Question Answering", ICCV 2015. <a href="mailto:arXiv:1505.00468v4">arXiv:1505.00468v4</a> [cs.CL]
- [3] <a href="http://nltk.org/">http://nltk.org/</a>

Author: Abhishek Jha

- [4] <a href="http://wordnet.princeton.edu/">http://wordnet.princeton.edu/</a>
- [5] http://scikit-learn.org/
- [6] S. Antol, C. L. Zitnick, and D. Parikh. Zero-Shot Learning via Visual Abstraction. In ECCV, 2014
- [7] C. L. Zitnick and D. Parikh. Bringing Semantics Into Focus Using Visual Abstraction. In CVPR, 2013.
- [8] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the Visual Interpretation of Sentences. In ICCV, 2013.
- [9] C. L. Zitnick, R. Vedantam, and D. Parikh. Adopting Abstract Images for Semantic Scene Understanding. PAMI, 2015.
- [10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In ECCV, 2014.