Visual Question Answering

Much work has been done in generating image captions in the recent years with the state of the art models capable of auto-generating captions given an image with at par human like accuracy. But captions lack to provide many meaningful informations about the relationship between different entities, instances and their attributes, which requires a more sophisticated intersection of Computer vision(CV) and natural language processing(NLP). Recently a new stream in vision community has come up which takes account of this challenge through what is called 'Visual question Answering' (VQA)[1]. VQA aims at providing accurate natural language answers to the natural language questions about the accompanying image.

Dataset:
The VQA [1] uses two datasets: MS-COCO [2], that has 204,721 images having 3 questions for each images and in total 7,984,119 answers for those questions(tain/val:123,287 and test:81434); and Abstract dataset having 50K scenes, with 3 questions each and in total 1,950,00 answers to those questions, as well as 5 captions each for images(train/test/validation: 20K/10K/20K). All the questions and answers are natural language type. There are two types of questions: open-ended free-form and multiple choice. Most of the answers are one two or three words. A detailed analysis is given in [1] regarding the distribution of the questions and their answers depending on the information content.

Methods:
For MS-COCO dataset if an answer is chosen randomly from the top 1K answers the accuracy is 0.12%, which is 29.72% when most popular answer is chosen(“yes”), this is because 38.37% of question for MS-COCO (i.e. real images) are yes/no type with 58.83% bias towards yes. Selecting most popular answer based on question type(like 'What is..', 'How many..', Is there...', etc) gives the accuracy of 36.18%; whereas nearest neighbor approach gives 40.61% on validation.

For baseline, top K=1000 most frequent answers are possible outputs which answers 82.67% of questions in train/val. Two models has been proposed: (i) a 2 layered(1000 units) multi-layer perception classifier and (ii) an LSTM model followed by softmax layer to generate the answer.

Features/setup for model (i):
Questions: Top 1000 words in the questions to create a 1000D bag of words(BOW) and top 10 first second and third words that start the questions for 30D BOW, then concatenated to create a 1030D input representation. Captions: Most popular 1000 words in the captions to create a 1000D BOW as input feature. Images: Last hidden layer of VGGNet [3] as 4096D feature.

Features/setup for model (ii):
Questions: The questions as one hot encoding of words are the input to the LSTM network with output in 1024D. Images: 4096D features (as above) after a transformation layer to 1024D to match the LSTM encoding of the questions. Before softmax layer the questions and image encodings are fused using element-wise multiplication.

For answers of open-ended question, the highest activation from possible K answers is selected, whereas, the highest activation among the options for a multiple choice question is it's answer.

Results: The best result comes for LSTM model when question and Image is given with an accuracy of 54.06% also the baseline for the VQA Challenge [4]. Other results through model (i) and regarding scenario where question, Image and caption has been used, for both real images and abstract scenes can be found in [1].

References:

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. arXiv:1505.00468, 2015.
[2] 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.
[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.
[4] Website: http://www.visualQA.org