VISUAL QUESTION ANSWERING

(Badri Narayan Patro, Vinay Kumar Verma, Atanu Samanta)
Instructor: Dr. Vinay P Namboodiri

Abstract

Visual Question and Answering (VQA) problems are one of the most attracting domain from multiple research disciplines. VQA takes the image and a natural question based on the image and gives the human like answers. Solving VQA problems requires techniques from both NLP(for generating semantics of the question and answers) as we as computer vision.

Introduction

Recently, researchers in the field of image caption generation have combined computer vision (CV) and natural language processing (NLP) and developed methods of jointly learning from image and text inputs to form high level description. Visual Question Answering (VQA) involves an extra layer of interaction between human and computes.

This work implements a generic end to end QA model using visual semantic embedding to connect convolutional neural network (CNN) and recurrent neural network (RNN)

In this work the problem is simplified by restricting the questions to have only single word answer so that the problem can be treated as a classification problem.

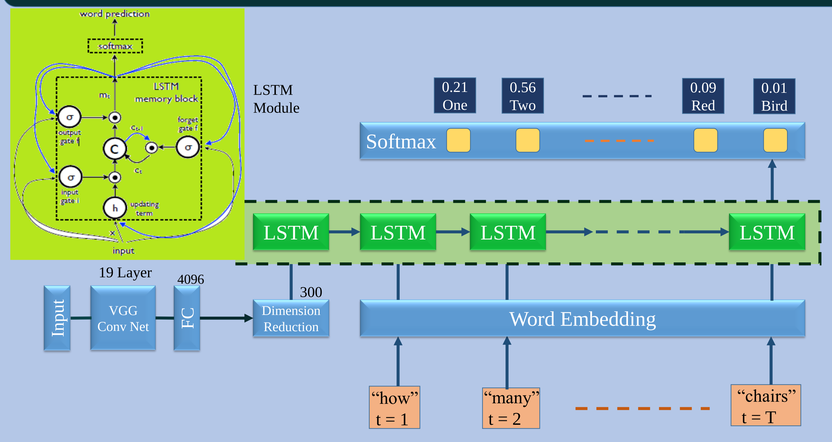
What is there in front of the sofa?
Ground truth: table

How many chairs are there?
Ground truth: one

Methodology

A model is built using Long Short Term Memory (LSTM), a form of recurrent neural network (RNN) with visual embedding using a trained CNN. Then the model is trained on COCO-QA dataset which contains images and QA pairs.

Model



- Last hidden layer of 19-layer Oxford VGG CNN trained on ImageNet 2014 Challenge is used as visual embedding. CNN was kept frozen during training.
- Linear transformation for dimension reduction of the image feature vector to match the dimension of the word embedding.
- Image is treated as the first word of the sentence.

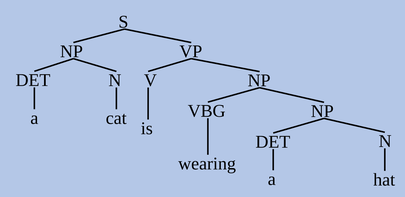
Question Answer Generation

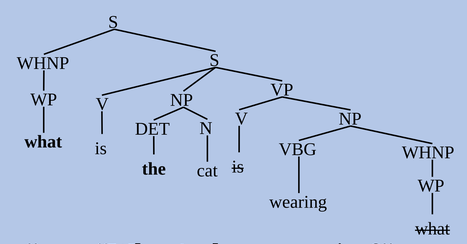
The following processing is performed on the MS-COCO image description dataset to automatically convert image description into QA form and create COCO-QA dataset.

1. Compound sentences to simple sentences
2. Indefinite determiners "a(n)" to definite determiners "the".
3. Traverse the sentence and identify potential answers and replace with "wh-" interrogative word such as "what".
4. Move the verb as well as "wh-" constituent to the front.

Question type:

Objective question: During traversal possible answer word is identified by finding noun using **WordNet** and **NLTK** software package.
Number question: During traversal, numbers are extracted from the image description.
Color question: Locate the color adjective and the noun to which the adjective is attached. Form a sentence "What is the color of the [object]" with the [object] replaced by actual noun.
Location question: Similar to object question generation except during traversal, algorithm will search for preposition "in"





Syntax Tree Example: "A cat is wearing a **hat**" => "**What** is **the** cat wearing?"

Results


Question:" What is the color of the tree?"Our Answere: Green, Gound truth : Green


Question: "how many Elephants are there?", Our Answer: 2 , Ground truth: 2


Question: "Which animal is this ?", Our Answer: Elephant, Ground truth : Cow


Question: "which fruit is this ?", Our Answer: Orange, Ground Truth : Orange

Discussion

In this project we used CNN for the image modeling and LSTM for the natural question and generating answers. We restrict our self to question based on the content, color & number etc. For most of the image in the restricted question domain we have very good result.
Further we would like to try object proposal to speedup the model. Here we used shallow LSTM network and VGG-CNN model. In future we would like to try deeper LSTM or neural Turing machine. Also we can use normalized CNN that promising better result.

References

1. M. Ren, R. Kiros, and R. S. Zemel, "Exploring Models and Data for Image Question Answering," CoRR, vol. abs/1505.02074, 2015
2. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," CoRR, vol. abs/1505.00468,2015.
3. H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Dataset and methods for multilingual question answering," CoRR, vol. abs/1505.05612, 2015.
4. L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," CoRR, vol. abs/1506.00333, 2015.