Visual Question Answering using Deep Learning

Project proposal for CS671A: Natural Language Processing

Shubham Agrawal, Aahitagni Mukherjee

October 4, 2015

1 Motivation

Multimodal learning between images and language has gained attention of researchers over the past few years. Using recent deep learning techniques, specifically end-to-end trainable artificial neural networks, performance in tasks like automatic image captioning, bidirectional sentence and image retrieval have been significantly improved. Recently, as a further exploration of present artificial intelligence capabilities, the task of visual question answering [1] has been proposed.

Visual question answering generates a free-form answer to a free-form input question, based on an input image. This problem is more challenging, because

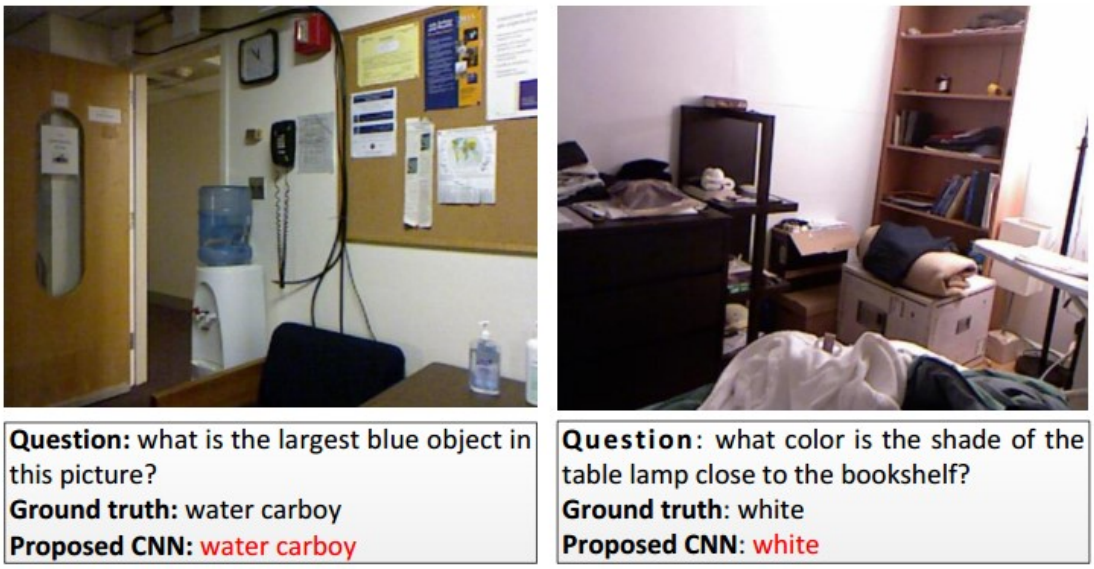
- The output is to be conditioned on both image and language inputs, unlike image captioning, where only conditioning on image is to be performed.
- A better representation of the image content is essential, because the task is not limited to a description of the objects in the image, but also their various attributes and relations between them. For example, if the question is, '*what colour is the shade of the table lamp close to the bookshelf*', we not only need to learn to identify the lamp, the bookshelf, and their colours, but also need to learn to model their relative positions.
- Interactions between the two modalities need to appropriately modelled.

2 Previous Work

Recently, several models based on neural networks have been proposed. Most of these use Recurrent Neural Network(RNN), Long Short-Term Memory(LSTM) or Convolutional Neural Networks(CNN). But the overall neural net architecture, i.e, the way in which the different RNN's and CNN's are combined, brings in the improvement in performance.

- In neural-based approach [2], image representation from a CNN is fed to each hidden layer of a single LSTM. The LSTM then models the concatenation of question and answer.

Figure 1: Examples of questions and answers [5]



- The mQA [3] approach contains four components: an LSTM to extract the question representation, a CNN to extract the visual representation, an LSTM for storing the linguistic context in an answer, and a fusing component to combine the information from the first three components and generate the answer.
- The VSE model(VIS+LSTM) [4] uses RNN's and Visual Semantic Embeddings. Here the image is treated as a single word, and the intermediate representation of the input thus obtained is used for classification into the correct class, which is the single word answer.
- The CNN approach [5] uses 3 CNN's - one to extract sentence representation, one for image representation, and the third is a multimodal layer to fuse the two.

3 Approach

There are a number of possibilities in terms of the choice of deep learning methods, and the way in which to combine them. Most of the previous work lacks an efficient way to model the complicated relationship between the inputs from the two modalities.

Our first objective is to implement the state-of-the-art CNN model [5] proposed by Lin Ma et al. Afterwards, we plan to modify the sentence representation using an LSTM. We intend to explore - how to model the relations between image representation from CNN and question representation from LSTM.

4 Dataset

We plan to train and test our models on the Toronto COCO-QA [4](Common Objects in Context - Question Answering) and DAQUAR [6](DAtaset for QUestion Answering on Real-world images) datasets. The Wu-Palmer Similarity(WUPS) measure will be used as the performance metric.

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[cs.AI]).

[2] M. Malinowski, M. Rohrbach, M. Fritz. *Ask Your Neurons: A Neural-based Approach to Answering Questions about Images*, (arXiv:1505.01121v3[cs.AI]).

[3] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, W. Xu. *Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering*, (arXiv:1505.05612[cs.AI]).

[4] M. Ren, R. Kiros, R. Zemel. *Exploring Models and Data for Image Question Answering*, (arXiv:1506.00333[cs.AI]).

[5] L. Ma, Z. Lu, H. Li. *Learning to Answer Questions From Image using Convolutional Neural Network*, (arXiv:1505.05612[cs.AI]).

[6] M. Malinowski, M. Fritz. *A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input*, (arXiv:1410.0210v4[cs.AI]).