Generating Images Conditioning on VQA Based System

Krishnakanth Singh
cs15mtech11007@iith.ac.in

Varun Mishra
cs16mtech11018@iith.ac.in

Vinod Chhapariya
cs16mtech11019@iith.ac.in

Abstract

This project deals with the problem of generating images by conditioning on Visual Question Answering (VQA) System. We develop neural network-based models to answer open-ended questions that are grounded in images. Our model makes use of two popular neural network architecture: Convolutional Neural Nets (CNN), Long Short Term Memory Networks (LSTM) and Hierarchical Recurrent Encoder Decoder (HRED). We use state-of-the-art CNN features for encoding images, and word embeddings to encode the words. This project aims to construct an end to end differential model for visual dialogue. for given image, model will be able to generate questions and for asked given question it will provide the answer. Our model is extended to image generation using (question, answer) pair for each image. there is no state of the art model for visual dialogue system.

1 Introduction

In recent years, there has been a lot of progress in AI problems at the intersection of Natural Language Processing (NLP) and Computer Vision. One problem that has garnered a lot of attention recently are visual question answering However, the task is not well suited to track the progress of AI since image captions are nonspecific, and for event-centric images it does not work well. In VQA system, the input is an image and a question based on the image, and the output is one or more words that answer the question. Open-ended question answering requires one to solve several lower-level problems like fine-grained recognition, object detection, activity recognition, common-sense reasoning, and knowledge-based reasoning. Due to the specificity of the task, it can also be evaluated automatically by making end to end system, making it easier to track progress.

1.1 Motivation

To develop state of the art model for Visual Dialogue System by considering VQA system as a baseline model. Visually impaired person can use the system to understand the image by questioning the image. Proposed model is end to end differential model. generating image image based on question answer pair is new approach to the image generation problem.

1.2 Overview

Image captions and answers generated by VQA system and questions are generated from Visual Question Generation (VQG). Question Answer pair (q,a) is passed as a input to Deep Recurrent Attentive Writer (DRAW) model. DRAW model generates image which can be evaluated using observation.

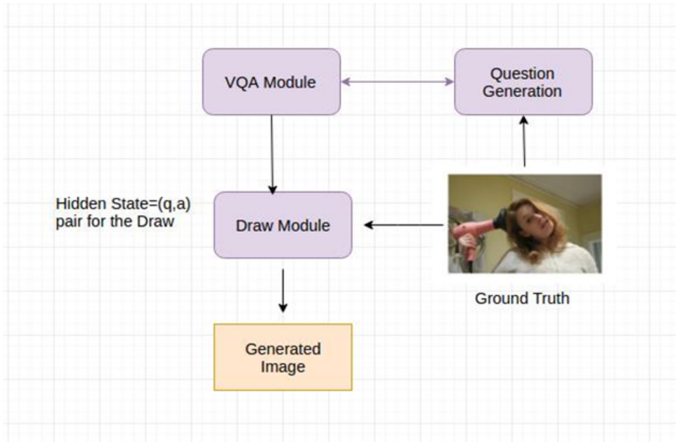

Figure 1: proposed model.

2 Related Work

2.1 VQA Model

for given image and an open-ended, natural language question about the image, the task is to provide an accurate natural language answer. dataset containing over 250K images, 760K questions, and around 10M answers. The wide variety of questions and answers in the dataset, as well as the diverse set of AI capabilities in computer vision, natural language processing, and commonsense reasoning required to answer these questions accurately.

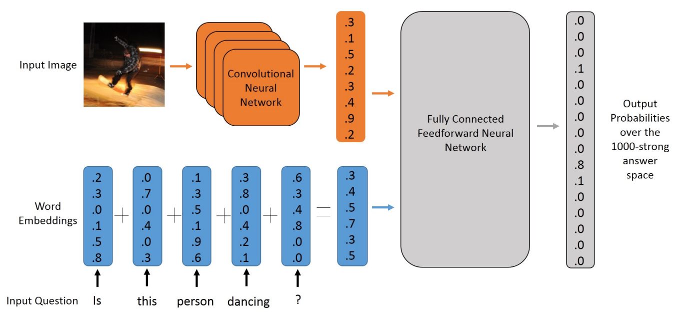

Figure 2: VQA model.

2.2 Visual Dialogue with Deep RL

This is first goal-driven training for visual question answering and dialog agents. Specifically, it is cooperative 'image guessing' game between two agents – Q-BOT and A-BOT– who communicate in natural language dialog so that Q-BOT can select an unseen image from a lineup of images. by using deep reinforcement learning (RL) to learn the policies of these agents end-to-end – from pixels to multi-agent multi-round dialog to game reward.

3 Model

Training a dialogue model end to end (generation of question + generation of answers) is not possible in a differentiable way if we do not take the Draw Module into account. We implemented 2 separate models:

1. Question Generation Conditioned on Image.
2. Answer Generation Conditioned on Image given Question generated in step1.

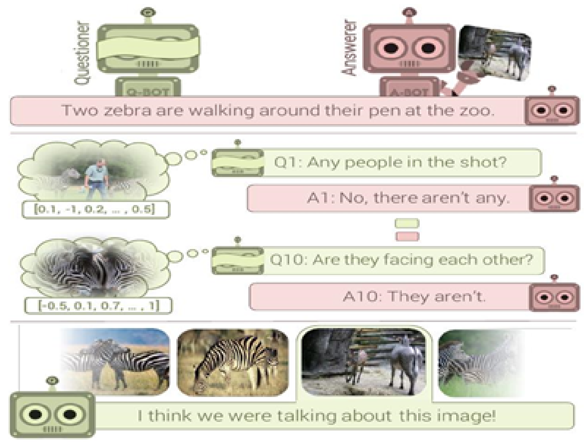

Figure 3: Visual dialogue with deep reinforcement learning.

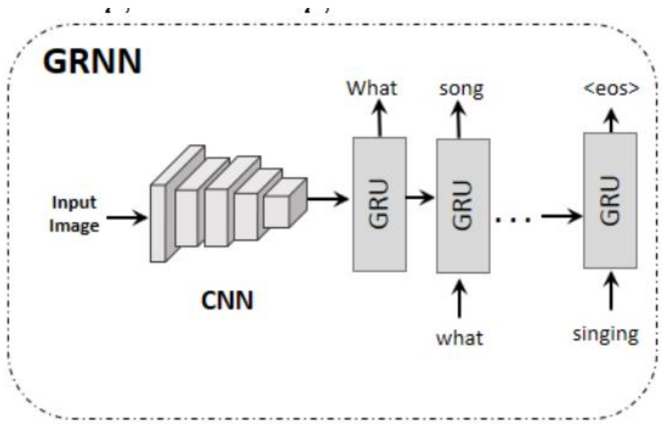

Figure 4: Visual Question Generation Module

To implement the Visual Question Generation (VQG) we followed the Moustafa et al[figure(4)] methods for generation of questions. We used pre-trained VGG-19 fc7 layer features to fees it to initial state of GRU. Using the questions from Vis-dial dataset to train the LSTM which is unrolled over 20 times frames, as the max length of question is 20.

Following is the graphical hierarchical sequence to sequence (s2s) model (HRED). Here we added another session level recurrent state to the s2s model for remembering the context of the dialogue. For current implementation, we used a basic s2s in our method instead of HRED. In this model initial state is conditioned on the input image features. To capture the maximum length wuestion from the data we unrolled th es2s 20 times.

3.1 Evaluation

We used the newer Vis-Dial dataset, where it consist of two mode of data. We used MCQ type.

- Given around 10 questions per 10 answers (pairs) that we have to choose for given the image.
- eg {img1,{1} {possible ans1,...,possible ans2}
- The dataset comprises of around 80k training images and 40k test images
- We trained on only 5k set of images using a pre-trained VGG net for feature extraction of images.

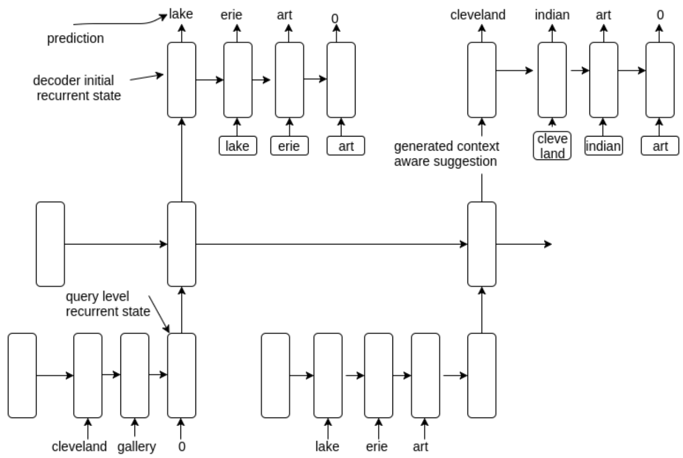

Figure 5: Visual Answer Generation Module

Question: does this appear to be a farmers market
Answer: it is close up, so hard to say
Question: can you see sale prices
Answer: no just fruits
Question: does the fruit ripe
Answer: yes, very nice
Question: can you see what the fruit is resting on in
Answer: no
Question: are the apples granny smith
Answer: i think there are no apples
Question: how many oranges do you see
Answer: at least 100
Question: do they look like navel oranges
Answer: like tangerine
Question: how many pears can you see
Answer: 0
Question: are they displayed in the same way you would I find them at the grocery store possibly
Answer: yes, next to each other

Figure 6: Good Results

- We used 3 layer multi RNN Cell With 500 dim hidden state size and 10 time step truncated backprop.
- Evaluation Metric: None(As we have generative model we could not find any suitable evaluation matrix).

3.2 Training

Both model, VQA and VQG modules trained separately on the vis-dial dataset. The VQG module is trained on the questions from the vis-dial dataset. We fed the output from VQG module to VAG module, where we choose the ground truth answer as the label for each step.

4 Results

Following are the some results that we were able to generate.

5 Future Work

1. Making the training end to end.
2. Evaluation metrics for the proposed work.
3. At present end to end training involves using a non differentiable learning mechanism -> Differentiable learning mechanism

Question: is there anyone in view
Answer: no
Question: what is the color of the tablet
Answer: brown
Question: how about the mechanical device
Answer: black
Question: is this inside a room
Answer: yes
Question: what is the color of the wall
Answer: no wall in view
Question: how about the floor
Answer: nope
Question: any flower vase in view
Answer: no
Question: what color is the keyboard
Answer: black
Question: any wires or cables
Answer: yes
Question: what is the color
Answer: black

Figure 7: Cases where our model failed

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. International Conference on Computer Vision (ICCV), 2015.

[2] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee and Dhruv Batra. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. ArXiv 2017

[3] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh and Dhruv Batra. Visual Dialog. CVPR, 2017