Visual Question Answering

Team 1

Gregory Brown
Priyanka Gaikwad
Sai Sri Narne
Sai Charan Kottapalli
Sai Srinivas Vidiyala

Overview

Visual Question Answering is a research area about building a computer system to answer questions presented in an image and a natural language.

Visual Question Answering is a system that answers the questions given an image:

Following is example of image from existing COCO QA just to show a glimpse of how system answers the question asked:

COCO-QA: What does an intersection show on one side and two double-decker buses and a third vehicle,?
Ground Truth: Building

This gives a basic overview of Visual Question Answering System.

For our implementation we are going to consider single word answer that will keep the evaluation more easier.

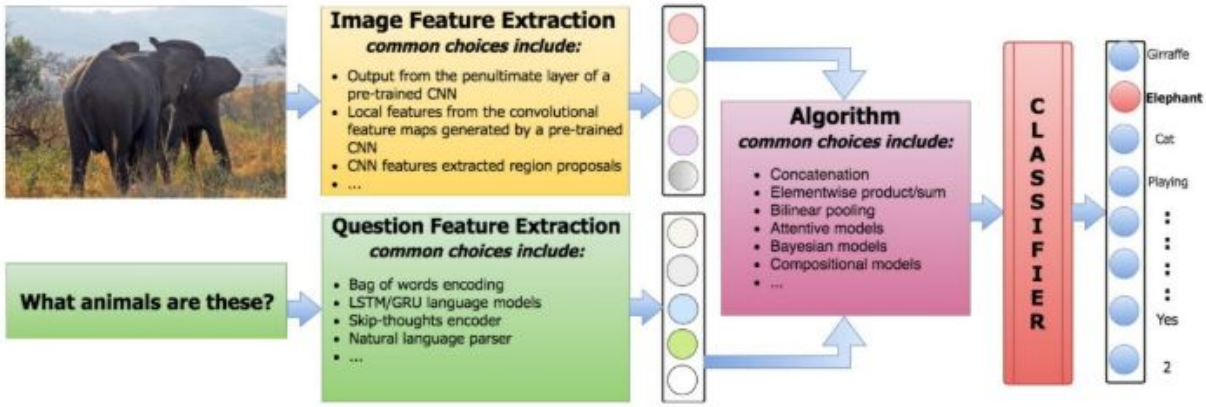
Context:

We are still undecided about what to go with as a context. We have considered 'Parks' and 'Store'. The MS COCO dataset includes good images for parks, though not 'Stores' (as do APR TC-12 and Google Conceptual Captions datasets), however we are still exploring the VQA dataset, and are uncertain of our ability to access the COCO QA dataset which we expect would have similar images of park settings as MS COCO. We will continue to consider and to analyse the datasets and then decide how we are going forward with it.

Implementation Approach:

The basic tasks that VQA performs are mentioned below:

1. Extract features from the question.
2. Extract features from the image.
3. Combine the features to generate an answer.



We are considering using the following datasets:

1. VQA - https://visualqa.org/
2. COCO QA - http://cocodataset.org/#home (COCO QA is separate from the standard MS COCO database and may require additional permission)
3. DAQUAR - https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/vision-and-language/visual-turing-challenge/

Related Applications:

Following is a list of reference papers which have implemented VQA:

1. **Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering**
2. **Adaptive Attention Fusion Network for Visual Question Answering**
3. **Fusing attention with visual question answering**
4. **A cascaded long short-term memory (LSTM) driven generic visual question answering (VQA)**
5. **Structured Semantic Representation for Visual Question Answering**
6. **VQA: Visual Question Answering**
7. **Increasing the Bandwidth of Crowdsourced Visual Question Answering to Better Support Blind Users**
8. **Object-Difference Attention: A Simple Relational Attention for Visual Question Answering**

Additionally below are some related projects:

https://github.com/paarthneekhara/neural-vqa-tensorflow

https://github.com/jazzsaxmafia/show_attend_and_tell.tensorflow