BOSTON UNIVERSITY

Visual Question Answering

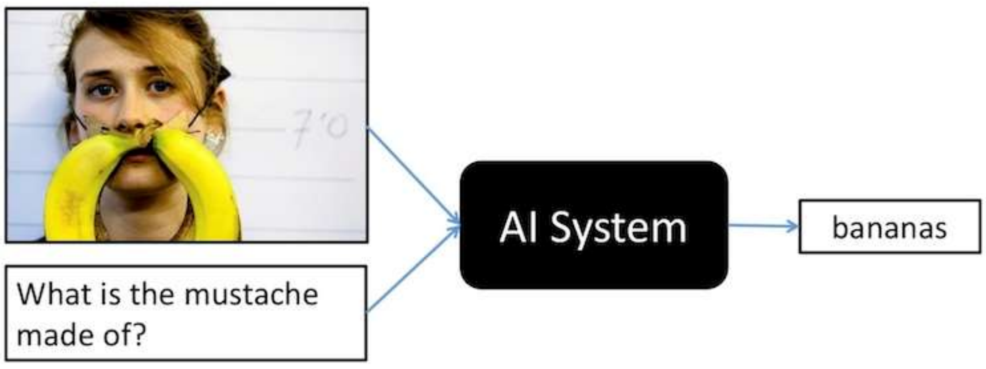
Qilong Wang, Yi Yang

Team#A2_16

Purpose

This project aims to use artificial intelligence to help blind and other visually impaired groups deal with some of the problems they encounter in life.

The VQA task is to provide an accurate natural language answer to the given question and picture.



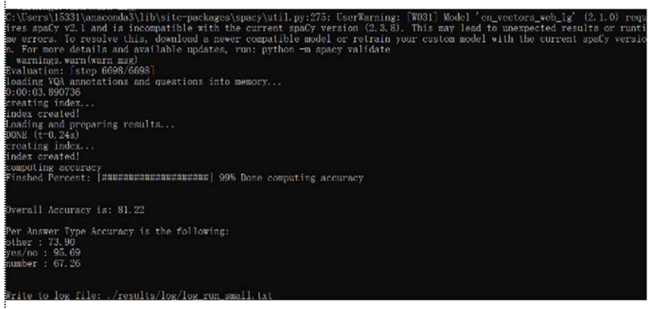
What is the mustache made of?

AI System

bananas

Result

The model performs very well on the VQA v2.0 validation set. The result is shown below.



```
C:\Users\15331\anaconda3\lib\site-packages\spacy\util.py:275: UserWarning: [W031] Model 'en_vectors_web_lg' (2.1.0) requ
ires spaCy v2.1 and is incompatible with the current spaCy version (2.3.8). This may lead to unexpected results or runti
me errors. To resolve this, download a newer compatible model or retrain your custom model with the current spaCy versio
n. For more details and available updates, run: python -m spacy validate
  warnings.warn(warn_msg)
Evaluation: [step 6698/6698]
loading VQA annotations and questions into memory...
0:00:03.890736
creating index...
index created!
Loading and preparing results...
DONE (t=0.24s)
creating index...
index created!
computing accuracy
Finshed Percent: [####################] 99% Done computing accuracy


Overall Accuracy is: 81.22

Per Answer Type Accuracy is the following:
other : 73.90
yes/no : 95.69
number : 67.26


Write to log file: ./results/log/log_run_small.txt
```

Question type	Accuracy
Other	73.90
Yes/no	95.69
Count	67.26
Overall	81.22

Demo

We also make a simple UI for the project which can output an answer according to the question and picture given by the user.

How many trains are in the picture
Pred: 2, Ans: 2
How many animals are there
Pred: 8, Ans: 8

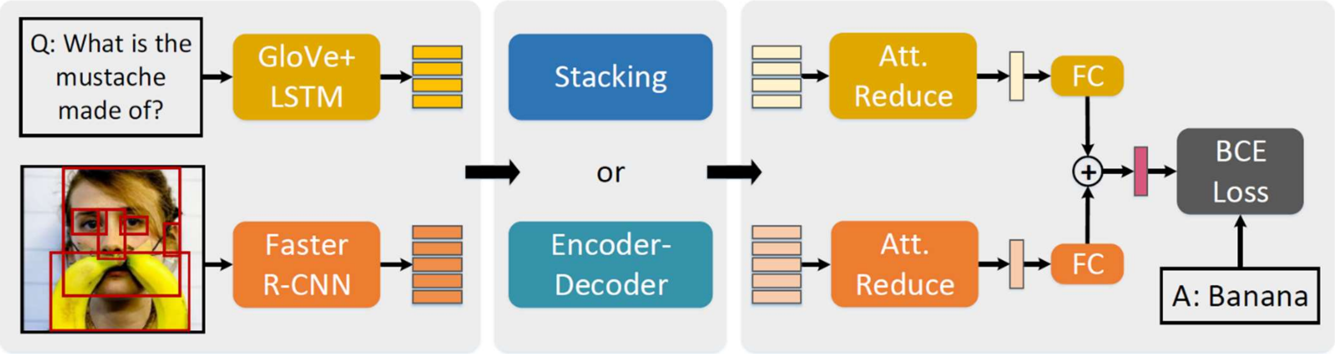
Model

The attention mechanism has shown very good results so far, and collaborative attention has also performed well in image and text representations.

BAN and DCN can solve the problem that co-attention cannot fully interact between multiple modalities.

Two general attention units: a self-attention (SA) unit for intramodal interaction and a guided-attention (GA) unit for modality interaction between states.

Connecting multiple module layers in series to form a MCAN network (Modular Co-Attention Network).



Q: What is the mustache made of?

GloVe+ LSTM

Faster R-CNN

Stacking

or

Encoder-Decoder

Att. Reduce

FC

Att. Reduce

FC

BCE Loss

A: Banana

Dataset

a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.

Future Work

Speech-to-text and text-to speech function: It would make the blind people use it conveniently.
Advice system: Give user some demands of picture so that the model can recognize pictures accurately.