Visual Question Answering

Product Design in Electrical and Computer Engineering

Ajay Krishna Anand, Mingda Li, Saurabh Singh, Shangzhou Yin

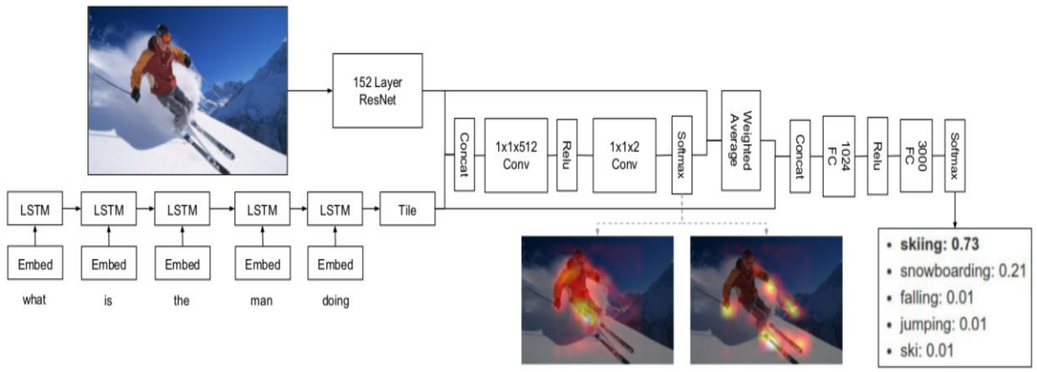
SCAN ME

Overview

A VQA system takes an image and a free-form natural language question as an input and produces a natural language answer based on the question asked as the output.

Our goal for this project was to design a system to answer a text-based question about an image to make it easier for people with visual impairments to interact with objects around them.

Architecture



Step1

A CNN based on ResNet is used to embed the image.

Step2

Input question is tokenized, embedded and fed to a multi-layer LSTM.

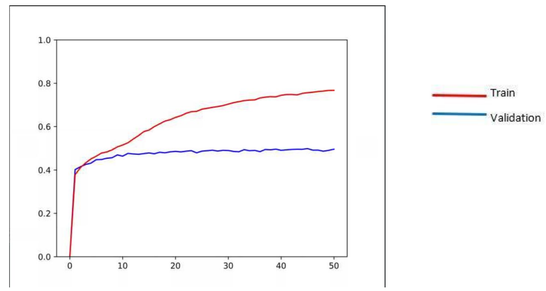
Step 3

Concatenated image features and final state of LSTMs are used to compute attention distribution over image features

Step 4

Image feature glimpses and LSTM state are fed to two fully connected layers to produce probabilities over answer classes.

Results



Training Accuracy: 0.54
Validation Accuracy: 0.46

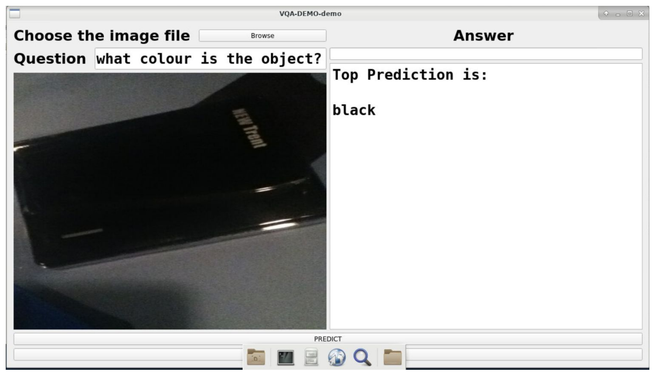
Answers	Percentage
Yes/No	5.40%
Unanswerable	32.30%
Other	61.00%
Number	1.20%

Dataset

Vizwiz Dataset

- 20523 training image
- 205230 training answer
- 4319 validation image
- 43,190 validation answer
- 8000 test image/question pairs

GUI Interface



Future Work

- Improve the ResNet152 accuracy by tuning hyper-parameters
- Implement the voice to text into GUI Interface.
- Existing VQA Datasets have less data for training and evaluation. Also, the datasets are biased.