VQA: Visual Question Answering

Anmol Popli and Sparsh Gupta
Agenda

Introduction
Dataset
Joint Embedding Based Model
Vanilla Attention Based Model
Bottom-Up and Top-Down Attention Based Model
Ideas from Multihop Attention Network
Results and Future Work
Agenda

Introduction

Dataset

Joint Embedding Based Model

Vanilla Attention Based Model

Bottom-Up and Top-Down Attention Based Model

Ideas from Multihop Attention Network

Results and Future Work
Visual Question Answering

Given an input image and a natural-language question about the image, the task is to provide a natural-language answer as output.

Some key areas of VQA application are:
-- Helping visually impaired users understand their surroundings
-- Helping intelligence analysts working on visual data
-- Efficient image retrieval for specific search queries
Visual Question Answering

Key Challenges
-- Multi-discipline problem - language understanding and vision understanding
-- Vision tasks: object detection, object recognition, counting, scene classification

Agenda

Introduction

Dataset

Joint Embedding Based Model

Vanilla Attention Based Model

Bottom-Up and Top-Down Attention Based Model

Ideas from Multihop Attention Network

Results and Future Work
Dataset

- VQA Dataset (v 2.0) for VQA 2019 challenge
- Our work focuses on **Balanced Real Images** portion of the entire dataset
- 444k questions based on 83k images
- 10 annotations per question

[www.visualqa.org](http://www.visualqa.org)
[https://visualqa.org/download.html](https://visualqa.org/download.html)
Preprocessing and Training Specifications

- Resized images to 224 x 224
- Chose only top 3000 most frequent answers
- Converted text to lowercase, removed punctuations, and treated named entities as UNK
- Max length of question text capped to 15 words
- Used GloVe embeddings to initialize word embeddings
- Used Cross Entropy Loss for the 3000-class classification problem
- Used Adam optimizer with scheduled learning rate decay
- L2 regularization of weights
Agenda

Introduction

Dataset

Joint Embedding Based Model

Vanilla Attention Based Model

Bottom-Up and Top-Down Attention Based Model

Ideas from Multihop Attention Network

Results and Future Work
Joint Embedding Based Model

Joint Embedding Based Model - Examples

- Is the TV on or off? On
- What side of the plate is the fork on? Left
- What flavor is the cake? Vanilla
- What is the man in the right wearing? Wetsuit
Agenda

Introduction

Dataset

Joint Embedding Based Model

Vanilla Attention Based Model

Bottom-Up and Top-Down Attention Based Model

Ideas from Multihop Attention Network

Results and Future Work
Vanilla Attention Based Model

Source: Kazemi et al. Show, ask, attend and answer: Strong Baseline for Visual Question Answering.
arxiv.org/abs/1704.03162
Vanilla Attention Based Model - Examples

is this man taking a selfie with his food? yes

how many slices of pizza is there? 1
Vanilla Attention Based Model - Examples

what kind of collar is the man's shirt? white

what kind of phone is the man talking on? none
Vanilla Attention Based Model - Examples

- What activity is about to take place? White
- Are the sinks facing up or down? No
Vanilla Attention Based Model - Attention Visualization

How many planes are in the air? 1
What kind of collar is the man’s shirt? white
Agenda

Introduction

Dataset

Joint Embedding Based Model

Vanilla Attention Based Model

Bottom-Up and Top-Down Attention Based Model

Ideas from Multihop Attention Network

Results and Future Work
Bottom-Up and Top-Down Attention Based Model

Bottom-Up and Top-Down Attention Based Model

- Faster-RCNN with ResNet-101 backend
- Regions of Interest (ROIs) obtained from Faster-RCNN
- Allows the model to compute attention for selected regions, and not the entire image
- Feature maps of each ROI are computed and are passed to the model
- Attention weight for each ROI feature map is computed
Agenda

Introduction

Dataset

Joint Embedding Based Model

Vanilla Attention Based Model

Bottom-Up and Top-Down Attention Based Model

Ideas from Multihop Attention Network

Results and Future Work
Multihop Attention Network

- Image features from Faster-RCNN for candidate ROIs
- Question features from LSTM
- Multiple layers of attention for attending to different regions of the image based on different words
- Aggregation - concatenation or addition
Agenda

Introduction

Dataset

Joint Embedding Based Model

Vanilla Attention Based Model

Bottom-Up and Top-Down Attention Based Model

Ideas from Multihop Attention Network

Results and Future Work
Results and Future Work

- Results obtained on validation data are at par with the results reported in respective papers
- Experimental model yet to be trained and tested
- Final test data predictions of experimental model to be submitted to challenge page
- Code will be made public shortly

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy on Validation Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Embedding Based Model - VGG 16</td>
<td>49.28 %</td>
</tr>
<tr>
<td>Joint Embedding Based Model - ResNet 152</td>
<td>51.76 %</td>
</tr>
<tr>
<td>Vanilla Attention Based Model - VGG 16</td>
<td>55.35%</td>
</tr>
<tr>
<td>Vanilla Attention Based Model - ResNet 152</td>
<td>57.13%</td>
</tr>
</tbody>
</table>
Q & A