Visual Question Answering



Gera Versfeld • 05.30.2017

Overview

Challenge deadline

June 25, 2017

Recent progress

- Read papers and learn about previous work in the field
- Run and understand Yifan's code for the baseline model
- Implement image feature extraction for VQA-v2 using Inception-v3

Biggest challenge

Creating a new architecture that relies heavily on image features not just text

Project goals:

1. Reading papers about VQA and visual attention
2. Extracting image features using pre-trained deep neural network
3. Embedding questions to semantic vectors with pre-trained model
4. Implementing attention model for final VQA
5. Tuning hyper-parameters
6. Submitting to VQA challenge/ publishing

Progress - Papers about VQA

VQA: A Survey of Methods and Datasets [1]

- Joint embedding approaches - concatenation, elementwise product,dynamic memory networks, multimodal bilinear pooling
- Attention mechanisms - spatial attention map, convolutional feature map, question guided feature map

Revisiting VQA Baselines [2]

- Basis for Yifan's model
- Model takes question, image and answer as input
- Image features extracted using pre-trained ResNet
- Question/answer embedded by averaging word2vec representation of words

VQA Papers cont.

Making the V in VQA Matter [3]

- Are models actually using image to answer question or just getting good results by memorizing bias in question types?
- Create balanced data set to help prevent this issue
- Benchmarks/baseline for VQA v2

Multimodal Compact Bilinear Pooling for VQA [4]

- Winner of last year's challenge
- New joint embedding approach using compact bilinear pooling algorithms
- Uses FFT to make calculations simpler and more efficient

Next steps

Baseline

Finish implementing baseline model for VQA-v2

Train and fine-tune baseline

Compare results with previous work

Add attention model

Compare with baseline and other attention models

Explore new architectures

Design new model

Goals for the week

1. Finish implementing baseline
2. Train and test baseline on cluster
3. Read papers about visual attention

References

1. Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel. Vidual question answering: A Survey of methods and datasets. 2017
2. Allan Jabri, Ahmad Joulin, and Lauren's van der Maaten. Revisiting Visual Question Answering Baselines. 2016
3. Yash Goyal, Tejas Khot, Douglas Summers-stay, Dhruv Batra, Devi Parikh. Making the V in VQA Matter. 2017
4. Akira Fukui, Doing Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach. 2016