Sprint 2

Visual question answering

What we have done so far

-Set up SCC and github

-Datasets

VizWiz

COCO

-Start to re-implement open-source modules

VizWiz Challenge: Visual Question Answering Implementation in PyTorch

Dataset VizWiz

- 20,523 training image/question pairs
- 205,230 training answer/answer confidence pairs
- 4,319 validation image/question pairs
- 43,190 validation answer/answer confidence pairs
- 8,000 test image/question pairs

Dataset COCO

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

Object segmentation
Recognition in context
Superpixel stuff segmentation
330K images (>200K labeled)
1.5 million object instances
80 object categories
91 stuff categories
5 captions per image
250,000 people with keypoints

Module ResNet-152

Input Questions are tokenized, embedded and encoded with an LSTM. Image features and encoded questions are combined and used to compute multiple attention maps over image features. The attended image features and the encoded questions are concatenated and finally fed to a 2-layer classifier that outputs probabilities over the answers (classes).

- No Attention: 2048 feature vectors consisting of the activations of the penultimate layer of pre-trained ResNet-152.
- Attention: 2048x14x14 feature tensors consisting of the activations of the last pooling layer of the ResNet-152.

The model uses only the "Attention" features.

Next Sprint Goals

1. Continue research on models
2. Try to design our own module for VizWiz Dataset
3. Work around in SCC