EECS 545 - Project Proposal

Sagnik Dam, Pranav Ramarao, Varsheeth Talluri
{sagnik, pranavr, varsh}@umich.edu

November 3, 2017

Problem Statement and Motivation

The project will attempt to build upon work in the field of bidirectional automatic image captioning which combines the problem of object recognition in images with expressing their relationships in a natural language. (by "bidirectional", we mean the ability to both caption an image and generate an image given a caption)

This could have great impact, for instance by helping visually impaired people better understand the content of images on the web. It could also have an impact on upcoming technologies such as the Google glass where the visually impaired can get a live feed of their surroundings.

Methodology

- Our base implementation will be Chen and Zitnick's paper [3] which has used a vanilla recurrent neural network implementation.
- We want to see the benefits of using Gated Orthogonal Recurrent Units (GORU) [2] instead of the vanilla RNN used in the base paper. The GORU paper primarily looked at question answering tasks and we feel that it would work well with the image captioning problem as well.
- We plan to use LSTMs similar to that outlined in the MSCOCO image captioning challenge [5]. We feel that the ability of LSTMs to learn longer sequences can impact the accuracy of the model and has been included as one of the areas to be explored by the authors of the base paper.
- There has been success in employing *attention*[1] in machine translation and object recognition. Hence, we would like to use them so that the models can attend to salient parts of an image while generating its caption.

Data Sets

- PASCAL 1k [6]
- Flickr 8k [7]
- Flickr 30k [8]
- Microsoft COCO [9]

Work Distribution

Initial implementation of the base paper [3] will be equally distributed and has to be done from scratch. Further work on extensions will be distrubuted as follows :

- GORU - Sagnik, Varsheeth
- LSTM - Varsheeth, Pranav
- Attention networks - Pranav, Sagnik

References

[1] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
`https://arxiv.org/pdf/1502.03044.pdf`

[2] Gated Orthogonal Recurrent Units: On Learning to Forget
`https://arxiv.org/pdf/1706.02761.pdf`

[3] Minds Eye: A Recurrent Visual Representation for Image Caption Generation
`https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Chen_Minds_Eye_A_2015_CVPR_paper.pdf`

[4] Deep Visual-Semantic Alignments for Generating Image Descriptions
`https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Karpathy_Deep_Visual-Sem Alignments_2015_CVPR_paper.pdf`

[5] Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge
`http://bengio.abracadoudou.com/publications/pdf/vinyals_2016_pami.pdf`

Dataset Links

[6] PASCAL 1k
`http://vision.cs.uiuc.edu/pascal-sentences/`

[7] Flickr 8k
`http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description/KCCA.html`

[8] Flickr 30k
`http://shannon.cs.illinois.edu/DenotationGraph/`

[9] Microsoft COCO
`http://mscoco.org/home/`

Examples

Examples from base paper[3] showing system generated (red) and human labeled (black) captions.

A group of people standing on top of a snow covered slope.

A group of people riding skis on top of a ski slope.

A person standing on a beach next to a surfboard in the ocean.

A man in a wetsuit with a surfboard standing on a beach.