Natural Language Description on Images

Team members: Hangjun Piao, Hongyi Wang, Xiaofei Liu, Zirui Tao

(1) The task to be addressed:

a. The team is interested in addressing the task of generating captions on images. The group has a four-phase plan. In the first phase, we would want to achieve simple object detection and segmentation. In the second phase, the program should be able to generate simple text description about the image. In the third phase, the program should be able to generate simple sentences regarding the picture which might be based on templates without contextual information. In the final phase the program would be able to generate a sentence without templates. Since all group members are undergrads without appropriate research experience, currently the group is expecting to achieve the first two phases in this class and if time permits the group would try to achieve the next two stages.

(2) The data set could be used:

a. The following are the datasets that have been commonly used in the researches relevant to the topic.

i. Flicker 8k, Flicker 30k, MS COCO

(3) From where you will get CPU cycles:

a. Start from personal computers.

b. Scale up on Amazon EC2 GPU enabled instances or HTCondor.

(4) Any existing software you plan to use:

a. Python Numpy and Scipy module.

b. Machine Learning Framework such as Google TensorFlow, or Caffe as have been used in cited papers.

c. A TensorFlow implementation of image-to-text model used by Vinyals et al. (2016).

    i. https://github.com/tensorflow/models/tree/master/im2txt

(5) The experimental methodology planned for evaluation:

a. BLEU: a metric that evaluates how accurate our generated caption is compared to the reference descriptions (Papineni et al., 2002).

b. More recently, a novel, human-based metric for image descriptions called CIDEr has been introduced and used by the organizers of the MS COCO Captioning challenge. In a nutshell, it measures consistency between n-gram occurrences in generated and reference sentences, where this consistency is weighted by n-gram saliency (TF score) and rarity (IDF-score) (Vedantam et al., 2015).

(6) Relevant researches:

a. https://arxiv.org/pdf/1609.06647.pdf (proposed model to generate image descriptions by Vinyals et al. (2016))

b. http://cs.stanford.edu/people/karpathy/cvpr2015.pdf (Another model that adopts the CNNs for image, bidirectional RNNs over sentences and a multimodal embedding alignment between these two (Karpathy & Li, p.1, 2015).)

c. https://arxiv.org/pdf/1504.00325.pdf (Description of the MS COCO Caption dataset and the evaluation server)

d. http://dl.acm.org/citation.cfm?id=1073135 (BLUE metric)

e. https://arxiv.org/abs/1411.5726 (CIDEr metric)