Project Interior Progress Report

Satoshi Tsutsui
Manish Kramer

*we also put project description on the git repo.

We really had hard time for environmental setup. We wanted to use faster-RCNN, but we almost spend the first half of the project just setting up faster-RCNN.
https://github.com/rbgirshick/py-faster-rcnn

Other than that, what we have done is the following. Each task is from project description.

- To get familiar with data. (both)
- Preprocess data. (satoshi)
- Build a base line using [1,2,3]. (satoshi)
- To get familiar with (fast or faster) R-CNN to detect object with localization. (manish)
  –but still have problem for R-CNN… it can only detect 20 classes.

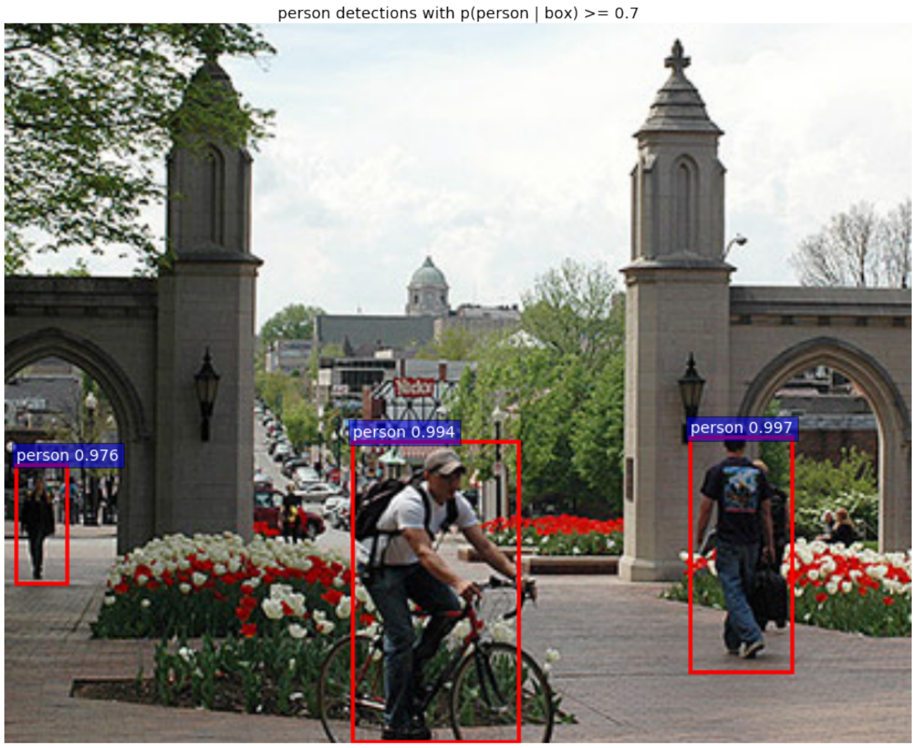
[1]Python implementation: https://github.com/apple2373/chainer_caption_generation
[2]Torch Implementation: https://github.com/karpathy/neuraltalk2
[3]Java implementation: http://nlp.stanford.edu/software/scenegraph-parser.shtml

What we have not done is the following.

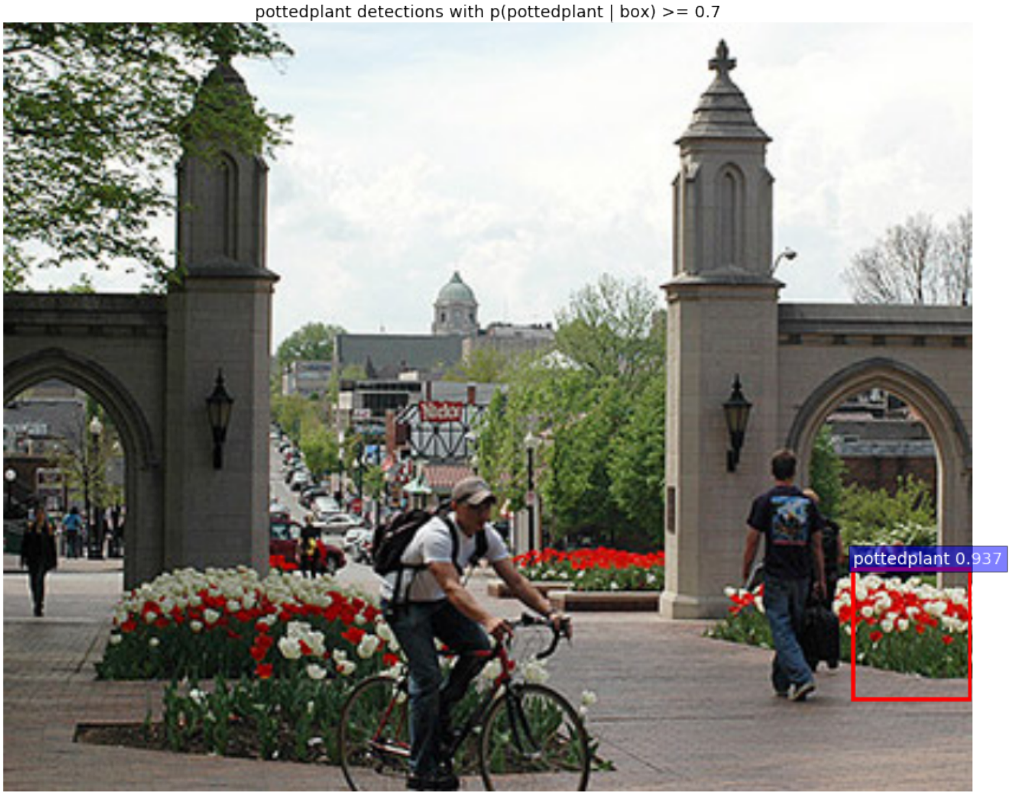
- Write evaluation program. Evaluate based on baseline. (manish)
- Implement [Neural Net 1] with a deep learning framework. Start training on GPU. (both)
- Train caption generation model with regional localization. (optional) (satoshi)
- Evaluate performance based on [Neural Net 1]. (manish)
- Modify network (satoshi)
- One more experiment with some network if possible. (satoshi)
- Start finishing up the whole system implementation with organized abovementioned modules. i.e. implement a method or class that takes image and output the graph. (manish)
- Finish up the whole system implementation with organized abovementioned modules. i.e. implement a method or class that takes image and output the graph. (both)
- Finish writing paper. (both)

We describe what we have done using the following example picture taken from:
http://www.indiana.edu/~stat/

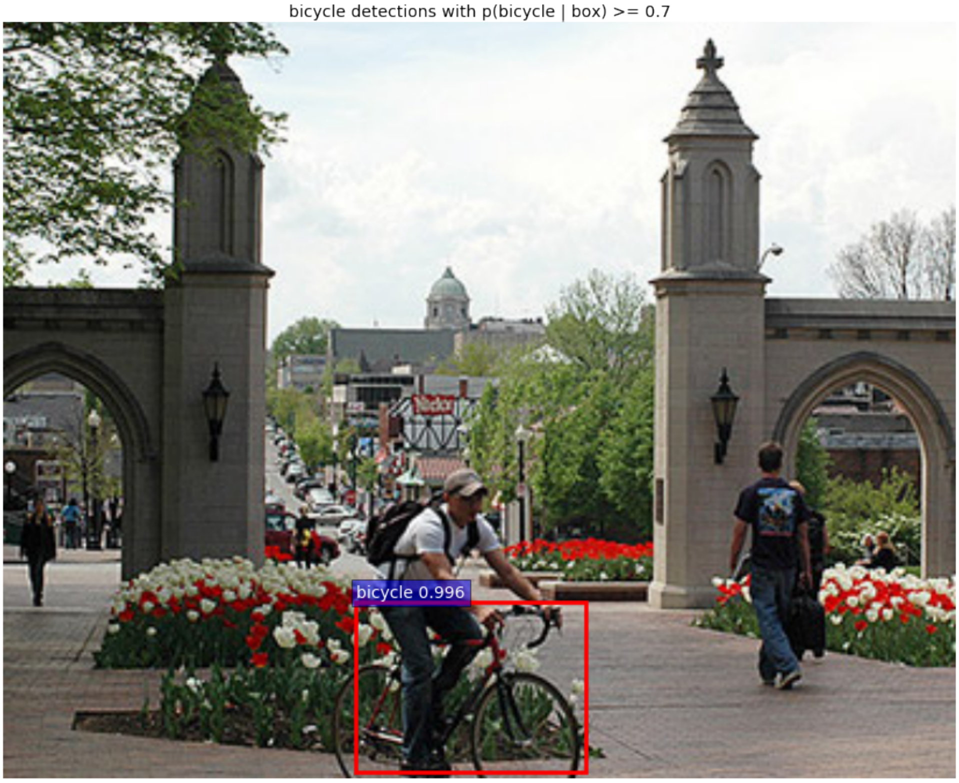
From R-CNN, we can currently detect:



pottedplant detections with p(pottedplant | box) >= 0.7



bicycle detections with p(bicycle | box) >= 0.7



The caption from the pic will be:

- a group of people riding bikes down a street 0.00274640011112
- a group of people riding on the backs of horses 0.00139282029438
- a man riding a bike down a street 0.000736576015526

The number is probability. Our current system does beam search with the beam size of 3. So the final caption will be “a group of people riding bikes down a street”.

Then this will be parsed to the caption parser to scene graph:

```
[[u'group-2', u'of', u'people-4'], [u'people-4', u'ride', u'bike-6']]
```