Team Members - Daniel Griffin, Yudhister Satija

766 CV Multi-Scale CNN Object Detection Project Proposal

1) Briefly explain what problem you are trying to solve.

The reliable detection of objects in an image is a challenging, fundamental problem in computer vision. Many problems arise in object detection that can cause challenges such as multiple scales, deformation, motion blur, and occlusion. Further, practical issues such as computational resources available, and latency concerns also arise as limiting factors in being able to detect objects in certain scenarios. The standard pipeline for object detection includes image preprocessing, region of interest detection, object classification, and verification.

Deep convolutional neural networks (CNNs) have been seeing much interest and research in the past few years, and have cutting edge performance for many object detection tasks. However, many current methods have issues detecting objects at multiple scales. Many current methods use computationally expensive methods to try to handle scale variation of objects, and the question of how to improve recognition of objects at various scales is still and open problem.

2) Why is this problem important?

Object detection has been a fundamental problem in computer vision for many years. Object detection methods can form fundamental building blocks for more complex computer vision tasks for things such as object tracking. Object detection and recognition also has a place in larger, more complex vision or image based systems. For example, in autonomous driving, the detection of moving cars and pedestrians is necessary for the vehicle to be able to reason about traffic laws, and to be able to perform driving tasks without colliding with other objects. Object detection and recognition is also being currently used in image based search where images are submitted to a search engine, and similar images are returned.

The particular problem of reducing computational time and improving object detection at multiple scales is particularly important for CNNs. While CNNs have been shown to have cutting edge performance, the time to detect objects can be higher than other methods. In many contexts such as autonomous vehicles, object detection at multiple scales and classification within a time limit is necessary to be able to reason and act in an environment before it significantly changes.

3) What is the current state-of-the-art?

As stated in the first section, the current state of the art for object detection and recognition are deep convolutional neural networks. These CNNs mimic the neocortex of the brain, and have special localized structures for forming complex image representations. Significant improvements in accuracy and computational performance have been achieved through more intelligent preprocessing and region of interest generation. The first major improvement was the R-CNN network, which generated regions of interest, and then used these regions along with a CNN to create feature transformations which were then classified using an

SVM. The R-CNN samples object proposals at multiple scales, and then warps the proposals to a fixed size input for the CNN. The next improvement to this methodology was the Fast R-CNN, which used “region of interest pooling” along with a CNN applied to the full image to generate feature representations for classification. While faster, the region of interest bottleneck still existed. The next major improvement was the Faster R-CNN, which used a separate CNN network for performing the region of interest proposal task. However, Faster R-CNN’s region of interest proposal network generates region proposals by sliding fixed sized filters over variable sized convolutional feature maps. A fixed receptive field cannot cover multiple scales at which an object appears in natural scene images, and reduces detection capabilities when objects are small.

4) Are you planning on re-implementing an existing solution, or propose your own new approach?**

We initially plan on implementing the multi-scale deep CNN network proposed in http://www.cvlibs.net/projects/autonomous_vision_survey/literature/Cai2016ECCV.pdf. This network builds off of the previously described state-of-the-art methods described, and claims to improve the region of interest proposal performance for small images by using multiple scale output layers, and framing the learning problem as a multi-task optimization problem. The work also proposes feature based upsampling of images as an alternative to input based upsampling methods to improve memory and cpu usage. The notion of using deconvolution for upsampling in this case comes from previous work where deconvolution has been used for segmentation and edge detection problems.

5) If you are proposing your own approach, why do you think existing approaches cannot adequately solve this problem? Why do you think your solution will work better?

While we believe that implementing multiple (ROI proposal net and Detection net with custom structures, loss functions, etc…) multi-scale deep neural networks will be challenging enough to fill the time for our project this semester, we would like to specify some reach goals. If we have time, a possible extension to this work would focus on trying to use our baseline CNN for object tracking, utilizing the popular “tracking by detection” framework. In this scenario, we would use the object detection method as a baseline for detection, and build code around it for identifying multiple objects and estimating potential tracks over time. From here we would be able to evaluate the object detector’s capabilities in the face of object occlusion, disappearance, and reappearance in a video to determine the network’s feasibility for object tracking tasks, and ways in which the network or tracking system as a whole could be changed to improve the tracking task.

6) How will you evaluate the performance of your solution? What results and comparisons are you eventually planning to show? Include a time-line that you would like to follow.

We plan on implementing the CNN architecture, and using the KITTI data and benchmark set. We initially plan to create a network and show its performance without comparison to other frameworks. If we implement the initial framework, and have time to

develop comparison methods, we would likely work with the Faster R-CNN which is a precursor to this paper.

Timeline:

- Region Proposal CNN Implementation:
    - Week 2/19, Week 2/26, Week 3/5, Week 3/12
- Object Detection CNN Implementation:
    - Week 3/19, Week 3/26, Week 4/2
- Benchmarking and Interactive Display:
    - Week 4/9, Week 4/16
- Final Documentation and Presentation:
    - Week 4/23