**Mask R-CNN for detecting building footprints from satellite images**

Most conventional object detection locates the trained classes in the form of bounding boxes (most conventionally using xmin, ymin, xmax and ymax) apart from classification. However, for this particular problem statement, we require the algorithm to predict the edges of the class(building) in addition to its location. Mask R-CNN is an algorithm that is an extension of Faster R-CNN. It has an additional parallel branch that predicts an object mask (0 for pixels that do not belong to object and 1 for pixels that belong to the object) along with the existing branch for bounding box recognition.

In principle Mask R-CNN is an intuitive extension of Faster R-CNN, yet constructing the mask branch properly is critical for good results. The mask and class prediction are two independent processes i.e, it predicts a binary mask for each class independently, without competition among classes, and rely on the network’s RoI classification branch to predict the category.

The first stage of MRCNN is the Region Proposal Network(RPN), which proposes candidate object bounding boxes. The second stage extracts features from each candidate box and performs classification, bounding-box regression and a binary mask for each RoI. The features used by these stages are shared for faster inference.

During training of Mask R-CNN, we have an additional loss factor that has to be taken into account when calculating gradients for back propagation. In addition to the ‘bbox\_loss’ and ‘classification\_loss’ in Faster R-CNN, Mask R-CNN also computes the ‘mask\_loss’ to account for improper masks.

For the assignment, I have separated out the available images and corresponding annotations into training and cross validation dataset using a python script(attached as split\_train\_test.py). All images and polygonal annotations shared were placed in a single folder and then the script was used to split the dataset for training. I have chosen training-CV split ratio as 85%:15% before training is started.

The algorithm is trained using the given images and the masks generated from the polygon coordinates given in the text file. Since we are training for only one class, we give the same class ID for all polygons. Masks are generated from coordinates during run time as the training progresses.