OverFeat Integrated Recognition, Localization and Detection using Convolutional Networks ax1402 iclr14

difference between localization and detection seems to be that the former only provides a location of the main object in the image which is used for classifying it while the latter is supposed to detect multiple objects of potentially different classes;
the latter is also penalized for false positives while the former is not

classification part seems to be using dense sliding window approach where a filter is applied exhaustively to each spatial location and scale and the feature maps thus obtained are pooled together and then passed to the classifier get the result

There is some idea of doing bottom-up computation so that the shared computation between consecutive sliding windows are only performed once but that part is not quite clear

The whole thing has evidently not been designed to be efficient

localization uses the same network and just adds a regression part with a couple of fully connected layers that produce confidence map for each spatial location and each class so that there are 1000 specialized class specific versions of this regression with each one producing the confidence map for the corresponding class;

greedy merge strategy is used to combine the predictions as well as bounding boxes – the algorithm seems to be one of those ad hoc annoying iterative ones;

Detection training is supposed to be similar to classification except that it is performed for each a spatial location though the description is not at all clear