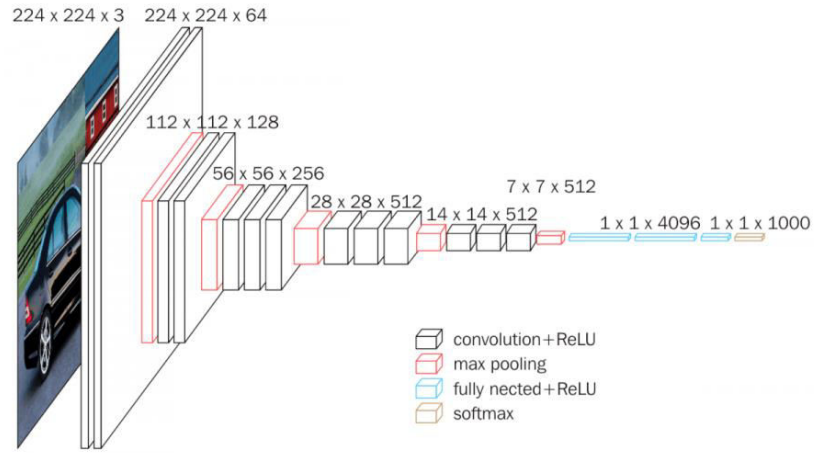
Fast RCNN:

Link: https://arxiv.org/pdf/1504.08083.pdf

Feature map:

Instead of building a feature map for every RoI like in RCNN, a convolution network is used to build a single feature map and only the features inside the suggested RoI is used for prediction. In this paper VGG-16 is used, replacing its last max pooling layer with the RoI pooling layer described below.



Region of Interest Proposal:
Selective search is used, but the number of regions is increased to 10,000 per image. Different scales are chosen so the area of the regions is 224^2.

RoI Pooling layer:
The RoI pooling layer uses max pooling to convert the features inside any valid region of interest into a small feature map with a fixed spatial extent of H × W, where H and W are layer hyper-parameters that are independent of any particular RoI. Each RoI is defined by a four-tuple (r, c, h, w) that specifies its top-left corner (r, c) and its height and width (h, w). The hXw RoI is divided into HxW sub-windows and the maximum value is taken from each sub-window. W and H need to be chosen to be compatible with the backbones first fully connected layer (7x7 in VGG-16).

Prediction:
After pooling the RoI, 2 fully connected are used and 2 networks are used to predict the class (softmax) and offsets to the RoI (L1 smooth loss).