# General Notes

* Intersection over Union – this is a way of measuring the accuracy of bounding box prediction. Take a predicted bounding box (predBox) and a ground-truth bounding box (gtBox). The IoU is the overlap of those boxes over the sum of both areas, or
* Local translation invariance – the property of a network topology being robust to changes in position of a particular feature
* Feature map – the output of a particular filter convolving over an input layer. The nature filter weights is to learn to activate based on the presence of particular image features. Hence, a high positive output from any given filter encodes the presence and location of that filter’s related feature into a feature map in the next layer.
* Pooling – a kernel convolves a feature map and outputs the average or max value (generally) of each position to a single node of the next layer
  + 2x2 or 3x3 with stride of 2 is often used
  + Global pooling – instead of convolving a pooling kernel, just pool the entire feature map
  + Pooling is one of the tools one can use to downsample input in order to achieve LTI
  + Max pooling tends to work better than average pooling for image classification – why is that?
  + Stride can also help
* Softmax function – takes as input a vector z of K real numbers, normalizes into a probability distribution of K probabilities proportional to exponential of input
  + IE. Apply exponential function to each element from input vector z, then normalize via sum of all such exponentials.
  + The name softmax is a misnomer: it does not output a smooth approximation of the maximum, but rather a smooth approximation to argmax – output the index at which the function is maximum.
* Rectified Linear Unit (ReLU) -

# You Only Look Once: Unified, Real-Time Object Detection

* Abstract notes
  + <https://arxiv.org/abs/1506.02640>
  + Instead of repurposing a classifier for the purpose of detection, YOLO “frames detection as a regression problem to spatially separated bounding boxes and associated class probabilities.”
  + Base model processes 45 fps, smaller network processes 155 fps
  + Has localization errors but less likely to have false positives on baseline detection
* A classifier might evaluate an image at various locations and sizes, R-CNN proposes bounding boxes and then runs a classifier on those regions. R-CNN also uses post-processing to refine box sizes, reject duplicate detections, and rescore boxes based on other detected objects
* Demo – <http://pjreddie.com/yolo>
* Global reasoning – contextual information is implicit since the entire image is fed in. Whereas Fast R-CNN falsely categorizes background images as classifications because it lacks information in surrounding bounding boxes. “YOLO makes less than half the number of background errors”
* Generalizable representations of objects – I guess this means an apple in all states is still an apple.
* Struggles with accuracy, especially with small objects

How it works

* Divide input image into an SxS grid, cells predict whether they contain the center of any one particular object in the image. If no object is in the cell it should predict zero. Measured with Pr(Object) \* IOUofTruthPred (what). Each bounding box consists of the following predictions:
  + x, y – center of box relative to bounds of grid cell
  + w, h – width and height
  + confidence – IoU of predicted box over gt box
* each cell also predicts class probabilities. Cells only predict one class regardless of the number of bounding boxes it predicted

Potential reads:

* depthwise convolution
* Histogram of oriented gradients and object detection

# Dilated Convolution

* Basically the same as regular convolution except there is L-1 skipped pixels between every recorded pixel where L is the dilation factor. Hence, L=1 is just regular convolution.
* Often (always?) applied to dense prediction – pixelwise image classification
  + Image super-resolution, denoising, demosaicing, bottom-up saliency, keypoint detection
* Can replace pooling layers (WHY) so there is no loss of information due to downsampling.
* Larger receptive field without the added computational cost of a larger convolutional kernel.
* A combination of exponentially increasing diation factor kernels results in micro and macro scale context being drawn from an image with little associated overhead.