YoLo v1

A Preprint

The main points: extremely fast, treating as regression problem(straight from image pixels to bounding box coordinates and class probabilities).

Other models at that time seems to not have direct access to whole image, while YOLO sees the entire image during training and test time so it implicitly encodes contextual information about classes as well as their appearance. The input is divided into $S \times S$ grid, If the center of an object falls into a grid cell, that grid cell is responsible for detecting that object(later described with detecting bboxes). For each grid cell B bboxes are predicted. Each bbox itself consists of (x; y; w; h; c) where x; y - coords of center; w;h - width and height respectively; c - confidence score which implemented as Intersection over Union with ground truth multiplied by probability that object exists. Each grid cell also predicts

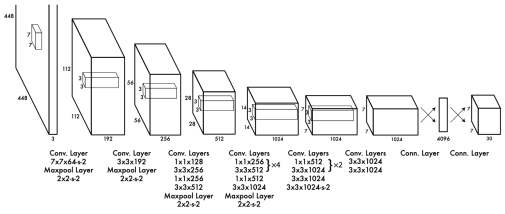
$$\Pr(\text{Class}_i|\text{Object})$$

At test time:

$$\Pr(\text{Class}_i \mid \text{Object}) \cdot \Pr(\text{Object}) \cdot \text{IoU}^{\text{truth}}_{\text{pred}} = \Pr(\text{Class}_i) \cdot \text{IoU}^{\text{truth}}_{\text{pred}}$$

Then all predictions are encoded as tensor with shape $S \times S \times (B \times 5 + C)$.

Overall structure:



All layers except last one followed by Leaky ReLU; for the last one the linear activation is used.

The whole loss function is defined as:

$$\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] + \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} (C_i - \hat{C}_i)^2 + \lambda_{\text{no}}$$

Using sum-squared error equally weights all errors. In object detection, many grid cells in an image don't contain any objects. If the confidence scores for these empty cells are pushed towards zero, it can overwhelm the gradients from the few cells that do contain objects. That is why λ_{coord} and λ_{noobj} are presented and set to 5 and 0.5 respectively.

However, there are signifact limitations such as only one class in grid cell and only 2 boxes.