Aero2Astro

YOLO V1

Report

By

Anshul Singh

Research intern-Inspect

How is YOLO different?

YOLO is different from all these methods as it treats the problem of image detection as a regression problem rather than a classification problem and supports a single convolutional neural network to perform all the tasks.

Benefits:

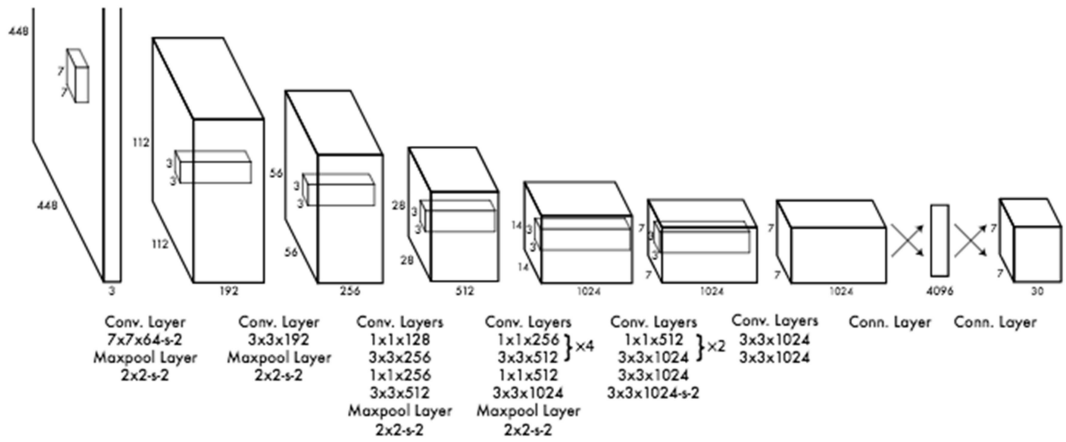
1. Speed
2. Less background mistake
3. Highly generalizable

Network Design

YOLO is implemented as a convolution neural network and has been evaluated on the PASCAL VOC detection dataset.

It consists of a total of 24 convolutional layers followed by 2 fully connected layers.

- First 20 convolutional layers followed by an average pooling layer and a fully connected layer is pre-trained on the ImageNet 1000-class classification dataset
- The pretraining for classification is performed on dataset with resolution 224 x 224
- The layers comprise of 1x1 reduction layers and 3x3 convolutional layers
- Last 4 convolutional layers followed by 2 fully connected layers are added to train the network for object detection
- Object detection requires more granular detail hence the resolution of the dataset is bumped to 448 x 448
- The final layer predicts the class probabilities and bounding boxes.



Loss Function

$$\lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

$$+ \lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2$$

$$+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2$$

$$+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2$$

Limitations

1. A cell can only detect one object
2. It finds it difficult to localize small objects or groups of small objects.

3. The model samples down the input image to an SxS grid where every grid cell is responsible for making bounding box predictions. Thus, due to the downsampling the model uses rather coarse features to predict the bounding boxes.