# Lab 2: YOLO pre-report

CHI YEONG HEO[1],
[1]School of Computer Science and Engineering, Pusan National University, Busan 46241 Republic of Korea

## 1. Introduction

This report provides a summary of a paper in the field of computer vision: You Only Look Once: Unified, Real-Time Object Detection[5].

## 2. Overview of YOLO

You Only Look Once (YOLO) is a real-time object detection framework known for its speed and efficiency. Unlike traditional methods that treat detection as a classification task, YOLO frames it as a single-stage regression problem, processing the entire image in a single network pass. This design enables real-time performance, end-to-end training, and global reasoning about the image. However, YOLO faces challenges in accurately localizing small objects.

## 3. Limitations of Prior Methods

Earlier detection methods, such as deformable parts models (DPM) [1] and R-CNN [3], use multi-stage pipelines like **sliding windows** or **region proposals**, which are complex, slow, and difficult to optimize due to separate training requirements for each component.

## 4. Unified Detection

### 4.1. Prediction Mechanism

The YOLO framework formulates object detection as a single-stage regression problem, where the input image is divided into an $S \times S$ grid. Each grid cell is responsible for predicting object detections if the object's center falls within that cell. For each grid cell, the network outputs:

$B$ **bounding boxes**, where each bounding box prediction consists of five components: $x, y, w, h$, and a confidence score. The coordinates $(x, y)$ denote the center of the bounding box relative to the boundaries of the grid cell, while $w$ and $h$ represent the width and height, respectively, normalized to the dimensions of the entire image. The **confidence score** is formulated as $\Pr(\text{Object}) \times \text{IOU}_{\text{pred}}^{\text{truth}}$, where $\Pr(\text{Object})$ indicates the probability of an object being present in the grid cell, and $\text{IOU}_{\text{pred}}^{\text{truth}}$ represents the Intersection over Union (IoU) between the predicted bounding box and the ground-truth bounding box.

$C$ **conditional class probabilities**, denoted as $\Pr(\text{Class}_i|\text{Object})$, are also predicted for all possible object categories. These probabilities are generated once per grid cell, regardless of the number of bounding boxes $B$, and indicate the likelihood of each class conditioned on the presence of an object in that cell.

The overall detection confidence for each bounding box is computed as:

$$\Pr(\text{Class}_i|\text{Object}) \times \Pr(\text{Object}) \times \text{IOU}_{\text{pred}}^{\text{truth}} = \Pr(\text{Class}_i) \times \text{IOU}_{\text{pred}}^{\text{truth}} \tag{1}$$

which provides a class-specific confidence score that encodes both the probability of the predicted class being present in the bounding box and the accuracy of the bounding box in localizing the object.

The predictions from all grid cells are combined, and Non-Maximum Suppression (NMS) is applied to filter out duplicate boxes for the same object.

### 4.2. Multi-part Loss Function

The YOLO training process optimizes a multi-part loss function:

$$\begin{aligned}
&\lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[ (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
&+ \lambda_{\textbf{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[ \left( \sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left( \sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\
&+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left( C_i - \hat{C}_i \right)^2 \\
&+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left( C_i - \hat{C}_i \right)^2 \\
&+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left( p_i(c) - \hat{p}_i(c) \right)^2
\end{aligned} \tag{2}$$

where $\mathbb{1}_i^{\text{obj}}$ denotes if object appears in cell $i$ and $\mathbb{1}_{ij}^{\text{obj}}$ denotes that the $j$th bounding box predictor in cell $i$ is "responsible" for that prediction.

The loss function comprises three components: **Localization Loss**: Captures the error between predicted and actual bounding box coordinates. The use of square roots for width and height helps ensure that small deviations in larger boxes contribute less to the loss than in smaller boxes. **Confidence Loss**: Measures the difference between the predicted confidence and the actual confidence, where the true confidence is 1 if an object is present and 0 otherwise. **Classification Loss**: Assesses the accuracy of the predicted class probabilities for the detected object.

To balance the loss contributions, the weighting factors $\lambda_{\text{coord}} = 5$ and $\lambda_{\text{noobj}} = 0.5$ are applied. These factors increase the influence of localization errors and reduce the impact of confidence errors for bounding boxes that do not contain objects.

## 5. Advantages and Limitations

| Real-Time Detectors | Train | mAP | FPS |
|---|---|---|---|
| 100Hz DPM [7] | 2007 | 16.0 | 100 |
| 30Hz DPM [7] | 2007 | 26.1 | 30 |
| Fast YOLO | 2007+2012 | 52.7 | **155** |
| YOLO | 2007+2012 | **63.4** | 45 |
| Less Than Real-Time | | | |
| Fastest DPM [8] | 2007 | 30.4 | 15 |
| R-CNN Minus R [4] | 2007 | 53.5 | 6 |
| Fast R-CNN [2] | 2007+2012 | 70.0 | 0.5 |
| Faster R-CNN VGG-16[6] | 2007+2012 | 73.2 | 7 |
| Faster R-CNN ZF [6] | 2007+2012 | 62.1 | 18 |
| YOLO VGG-16 | 2007+2012 | 66.4 | 21 |

Table 1: **Real-Time Systems on PASCAL VOC 2007.** Comparing the performance and speed of fast detectors. Fast YOLO is the fastest detector on record for PASCAL VOC detection and is still twice as accurate as any other real-time detector. YOLO is 10 mAP more accurate than the fast version while still well above real-time in speed.

YOLO's single-pass architecture makes it **significantly faster** than other detection models, enabling real-time performance. By processing the entire image, YOLO utilizes **global context** for more accurate bounding box predictions and has a relatively **simple architecture** compared to multi-stage approaches.

However, YOLO **struggles with detecting small objects** due to the grid-based prediction. Its speed comes at the **expense of some accuracy**, and it may **face challenges in localizing objects near grid boundaries**.

[1] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[2] R Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[4] Karel Lenc and Andrea Vedaldi. R-cnn minus r. *arXiv preprint arXiv:1506.06981*, 2015.

[5] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149, 2016.

[7] Mohammad Amin Sadeghi and David Forsyth. 30hz object detection with dpm v5. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 65–79. Springer, 2014.

[8] Junjie Yan, Zhen Lei, Longyin Wen, and Stan Z Li. The fastest deformable part model for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2504, 2014.