Object Detection Reviews

Ao Zhang

July 3, 2021

1 Backbones

- **Xception**
- **VGG16**
- **ResNet50**
- **InceptionV3**
- **MobileNet**
- **DenseNet**
- **NASNet**

2 Neck Layers

- **Feature Pyramid Networks (FPN)**

3 Detection Head

3.1 One-Stage Detector

- **YOLO**: The final feature maps shape of YOLO Detection Head should be $(B, W, H, K \times (5 + num\_classes))$, where K means number of anchor boxes; 5 including box info $[x, y, w, h]$ and an "objectness". The box decoding system is shown as below. The anchor boxes of YOLO are directly found by K-means Clustering on the entire dataset with a certain error rate threshold.

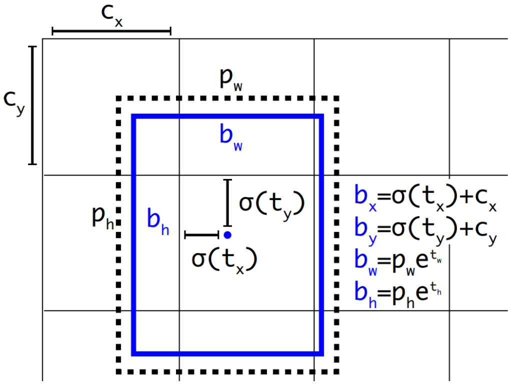
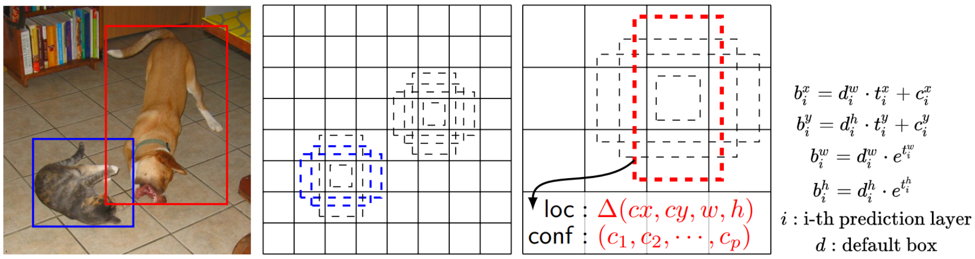


Figure 1: YOLO

- **SSD**: The final feature maps shape of SSD Detection Head should be $(B, m, W, H, K \times (4 + num\_classes))$, where m means number of output feature layers for the decoding; 4 consist $[x, y, w, h]$. The box decoding system is shown as below. The anchor boxes of



loc : $\Delta(cx, cy, w, h)$
conf : $(c_1, c_2, \cdots, c_p)$
$b_i^x = d_i^w \cdot t_i^x + c_i^x$
$b_i^y = d_i^h \cdot t_i^y + c_i^y$
$b_i^w = d_i^w \cdot e^{t_i^w}$
$b_i^h = d_i^h \cdot e^{t_i^h}$
i : i-th prediction layer
d : default box

(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Figure 2: SSD

SSD is a little different. It first defines box aspect ratios with K-means Clustering. Then define the boxes with scales s and aspect ratios a_r.

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 1}(k - 1), \quad k \in [1, m] \tag{1}$$

$$w_k^a = s_k\sqrt{a_r}, \quad h_k^a = s_k/\sqrt{a_r} \tag{2}$$

$$a_R \in \{1, 2, 3, \frac{1}{2}, \frac{1}{3}\} \tag{3}$$

Where, m is the total number of output feature layers. For the default box, it is defined with scale $s'_k = \sqrt{s_k s_{k+1}}$. On the other hand, since SSD doesn't have "objectness". The **hard negative mining** is required for training. Hard negative mining is to select the negative samples with highest prediction probilities to conduct backpropagation. In SSD itself, the ratio of negative samples to positive samples is ensured as 3 : 1.

3.2 Two-Stage Detector

- **Faster R-CNN: Region Proposal Networks + Classification Head**; It is an anchor-based method.

4 Metrics

- **Precision**:

$$\text{precision} = \frac{TP}{TP + FP} \tag{4}$$

- **Recall**:

$$\text{recall} = \frac{TP}{TP + FN} \tag{5}$$

- **Average Precision (AP)**: the area of precision-recall curves.
- **F1 Score**:

$$F1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = \frac{TP}{TP + \frac{FN+FP}{2}} \tag{6}$$

- **Receiver Operating Characteristics (ROC) Curves**:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{TN + FP} \tag{8}$$

- **Confusion Matrix**: assume a classification task, the confusion matrix is computed on the validation set with each entry of the matrix indicating the count for each class predicted by the classifier. Example shown as below.

$$C = \begin{bmatrix} 21 & 2 & 7 \\ 10 & 11 & 9 \\ 3 & 1 & 26 \end{bmatrix} \tag{9}$$

  Where, the perfect classifier would have results like,

$$C = \begin{bmatrix} 30 & 0 & 0 \\ 0 & 30 & 0 \\ 0 & 0 & 30 \end{bmatrix} \tag{10}$$

5 Others

5.1 Data Augmentation

- **Point Operations**: Change brightness, saturation, hue, gamma, correction, histogram equalization, etc.
- **Geometric Operations**: Shift, flip, rotate or shear.
- **Adding Noise**: Gaussian noise, etc.

5.2 Loss Functions

5.3 Regularization

5.4 Normalization

5.5 Skip-Connections