YOLOv1的缺点

检测小目标效果差

Bounding box准确率低

每个cell只能预测一个分类

...

YOLOv2

	YOLO							YOLOv2
batch norm?		✓	✓	✓	✓	✓	✓	✓
hi-res classifier?			✓	✓	✓	✓	✓	✓
convolutional?				✓	✓	✓	✓	✓
anchor boxes?				✓	✓			
new network?					✓	✓	✓	✓
dimension priors?						✓	✓	✓
location prediction?						✓	✓	✓
passthrough?							✓	✓
multi-scale?								✓
hi-res detector?								✓
VOC2007 mAP	63.4	65.8	69.5	69.2	69.6	74.4	75.4	76.8	**78.6**

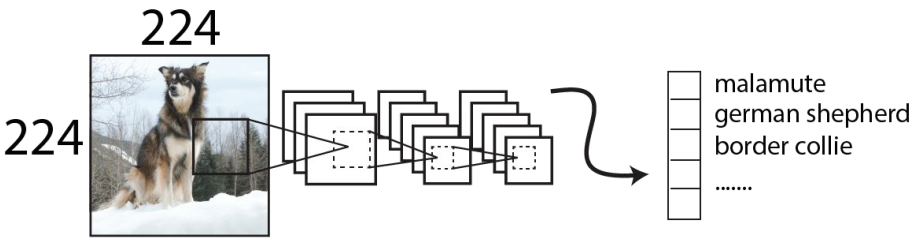
1. Batch Normalization

2. High resolution classifier

YOLO v1使用ImageNet的图像分类样本采用 224*224 作为输入，来训练CNN卷积层。然后在训练对象检测时，检测用的图像样本采用更高分辨率的 448*448 的图像作为输入。但这样切换对模型性能有一定影响。

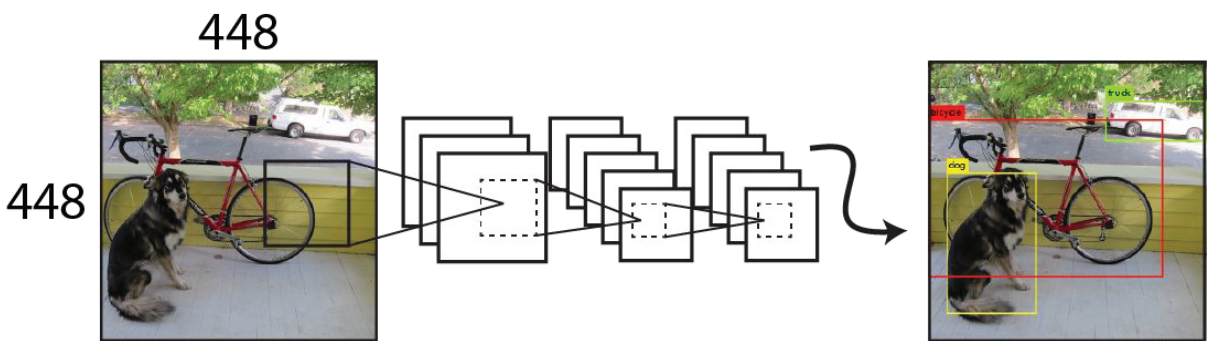
所以YOLO2在采用 224*224 图像进行分类模型预训练后，再采用 448*448 的高分辨率样本对分类模型进行微调（10个epoch），使网络特征逐渐适应 448*448 的分辨率。然后再使用 448*448 的检测样本进行训练，缓解了分辨率突然切换造成的影响。

Train on ImageNet



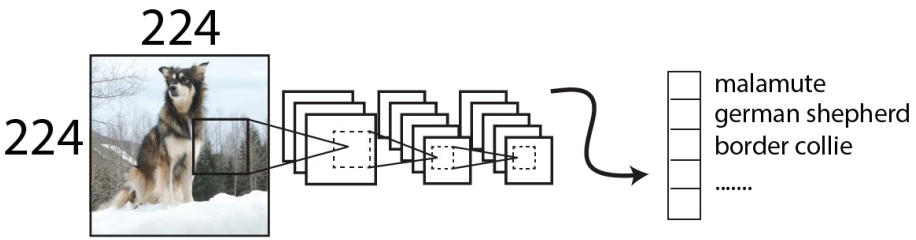
Resize network

Fine-tune on detection

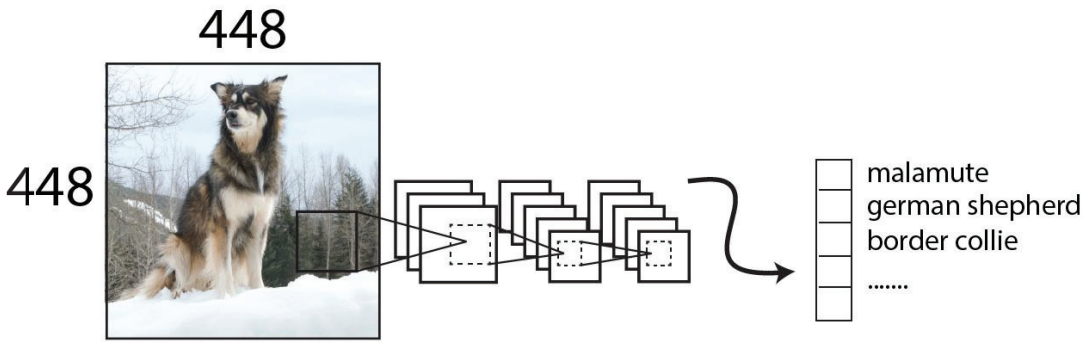


Fine-tune 448x448 Classifier: +3.5% mAP

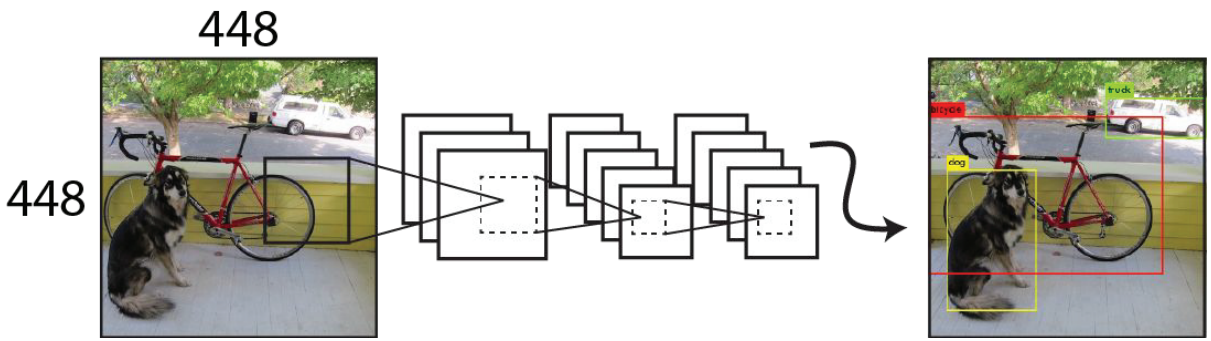
Train on ImageNet



Resize, fine-tune on ImageNet



Fine-tune on detection



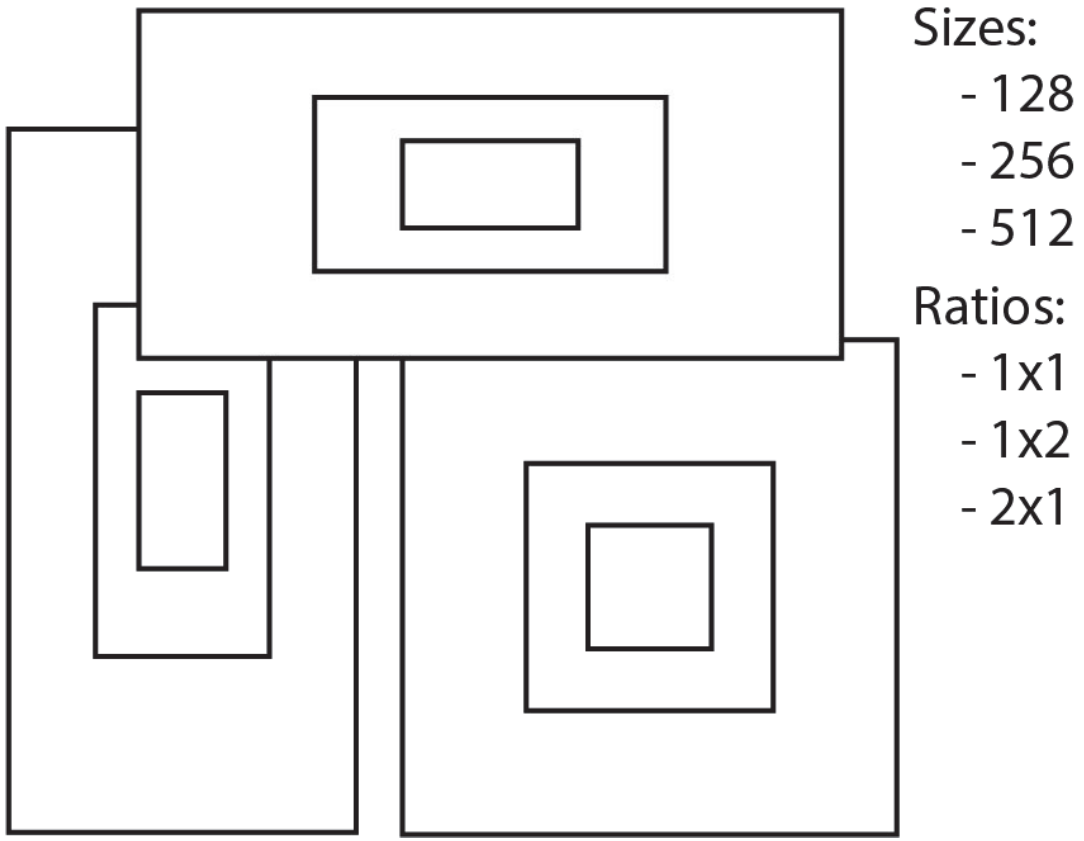
3. Anchor Boxes

借鉴Faster RCNN的做法，YOLO2也尝试采用先验框（anchor）。在每个grid预先设定一组不同大小和宽高比的边框，来覆盖整个图像的不同位置和多种尺度，这些先验框作为预定义的候选区在神经网络中将检测其中是否存在对象，以及微调边框的位置。

4. Dimension Priors

之前先验框都是手工设定的，YOLO2尝试统计出更符合样本中对象尺寸的先验框，这样就可以降低网络微调先验框到实际位置的难度。YOLO2的做法是对训练集中标注的边框进行聚类分析，以寻找尽可能匹配样本的边框尺寸。

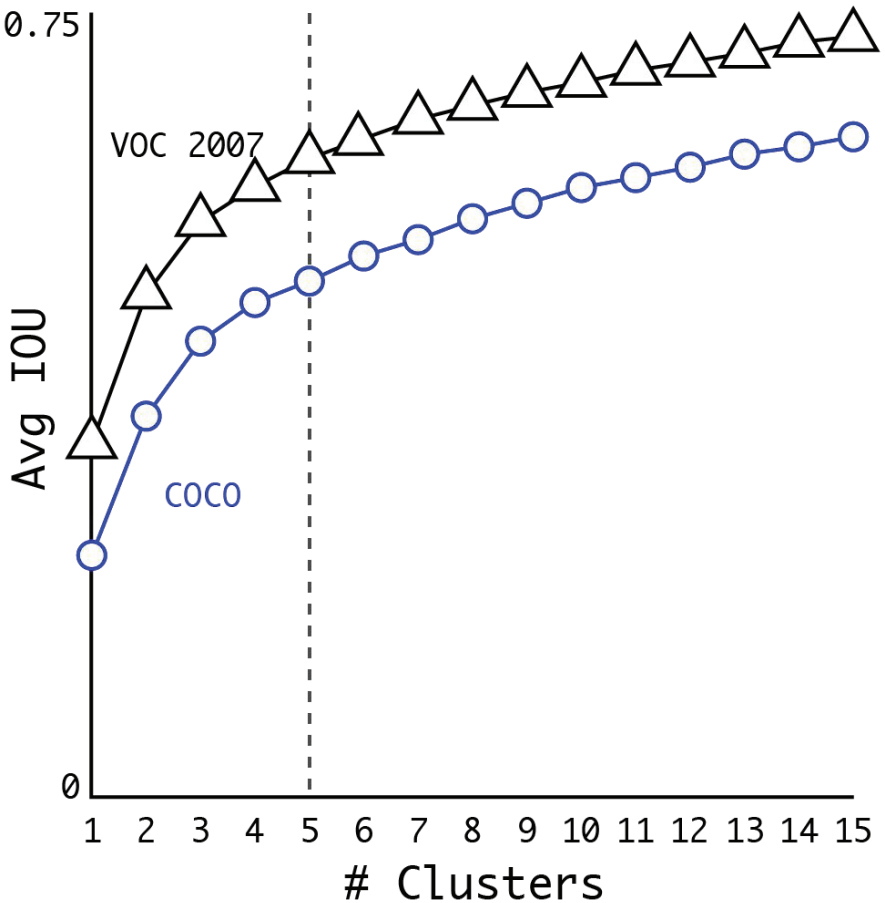
Anchor boxes use static initialization

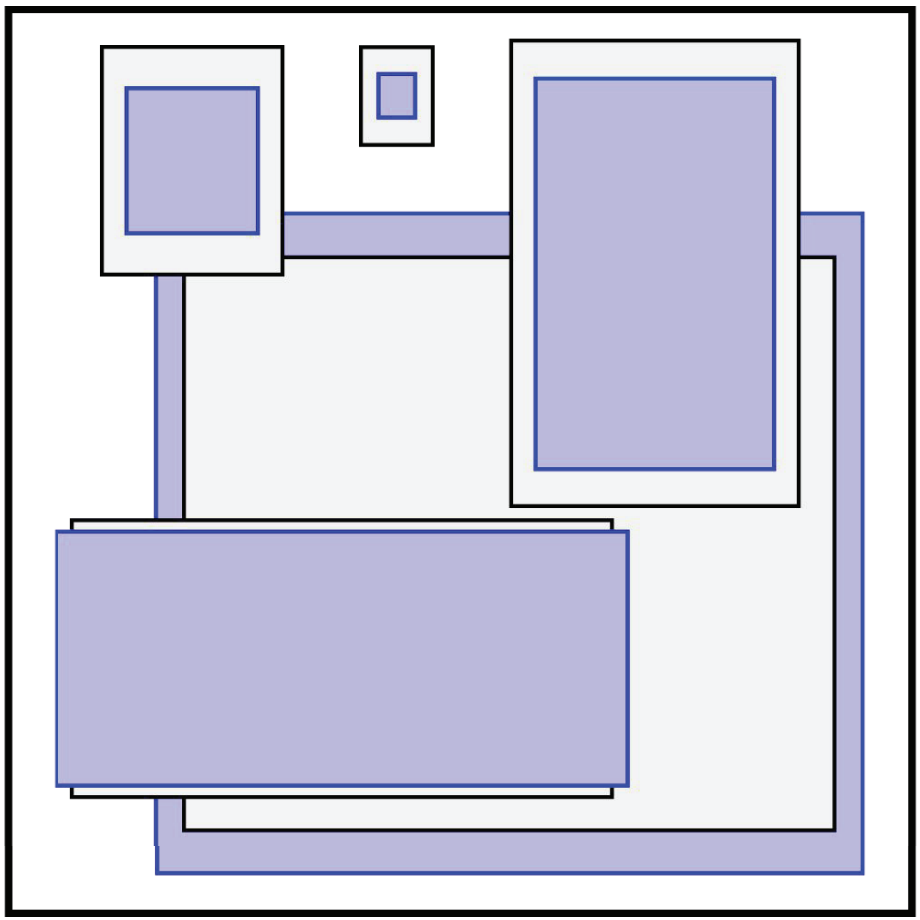


Sizes:
- 128
- 256
- 512

Ratios:
- 1x1
- 1x2
- 2x1

We use k-means to find better initializations





5. Network

Darknet-19

6. Location Prediction

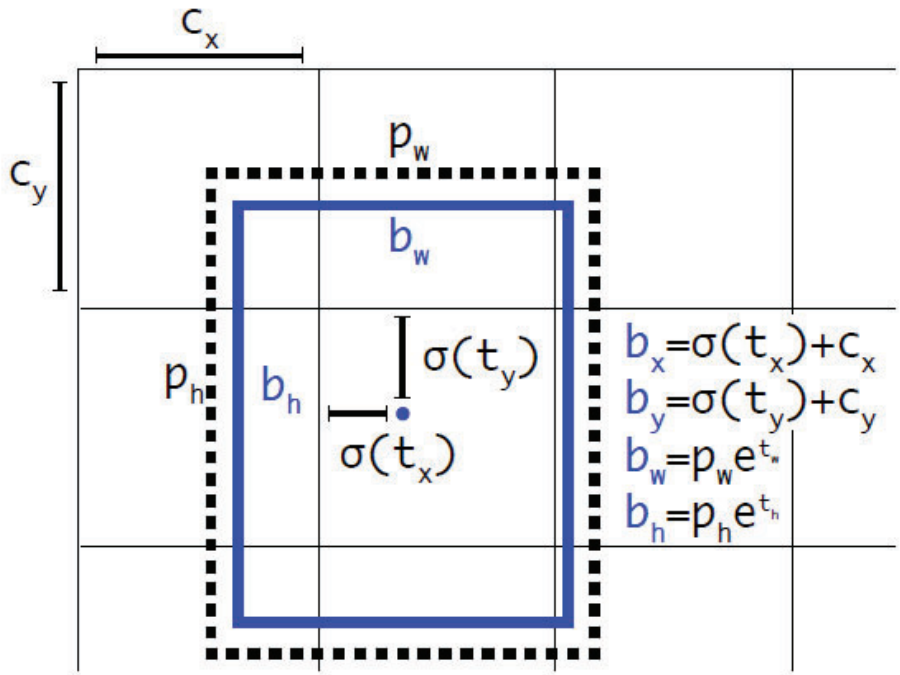


Figure 3: Bounding boxes with dimension priors and location prediction. We predict the width and height of the box as offsets from cluster centroids. We predict the center coordinates of the box relative to the location of filter application using a sigmoid function.

YOLOv3

1. Network

Darknet-19 ---→ Darknet-53

2. Class Prediction

softmax ---→ logistic

3. Predictions Across Scales

特征图	13*13			26*26			52*52		
感受野	大			中			小		
先验框	(116x90)	(156x198)	(373x326)	(30x61)	(62x45)	(59x119)	(10x13)	(16x30)	(33x23)

对于一个416*416的输入图像，在每个尺度的特征图的每个网格设置3个先验框，总共有
13*13*3 + 26*26*3 + 52*52*3 = 10647 个预测。
每一个预测是一个(4+1+80)=85维向量，这个85维向量包含边框坐标（4个数值），边框置信度（1个数值），对象类别的概率（对于COCO数据集，有80种对象）。