


联系我们



请扫描二维码联系客服  
✉ webmaster@csdn.net  
☎ 400-660-0108  
💬 QQ客服    💬 客服论坛

关于   招聘   广告服务   网站地图

©2018 CSDN版权所有 京ICP证09002463号

👉 百度提供支持

经营性网站备案信息



网络110报警服务


中国互联网举报中心

北京互联网违法和不良信息举报中心

GitChat   TinyMind   论坛   问答   商城   ...

🔍

 写博客   

 RSS订阅

阅读数：3314

摘要：

我们介绍YOLO9000，一个最先进的，实时目标检测系统，可以检测超过9000个目标类别。首先，我们提出对YOLO检测方法的各种改进方法，包括新颖的和从以前的工作中得出的。改进的模型YOLOv2在如PASCAL VOC和COCO标准检测任务是最先进的。使用一种新颖的多尺度训练方法，相同的YOLOv2模型可以运行在不同的大小的图片上，提供速度和精度之间的轻松权衡。在67 FPS时，YOLOv2在VOC 2007上获得76.8 mAP。在40 FPS时，YOLOv2获得78.6 mAP，性能优于最先进的方法，例如使用ResNet的faster RCNN和SSD，同时运行速度明显更快。最后，我们提出了一种联合训练目标检测和分类的方法。使用这种方法，我们在COCO检测数据集和ImageNet分类数据集上同时训练YOLO9000。我们的联合训练方法允许YOLO9000预测没有标记检测数据的目标类的检测。我们在ImageNet检测数据集上验证我们的方法。YOLO9000在ImageNet检测验证集上获得19.7 mAP，尽管只有200个类中的44类检测数据。在不在COCO的156类中，YOLO9000获得16.0 mAP。但是YOLO可以检测超过200个类;它预测超过9000个不同目标类别的检测。它仍然实时运行。

1、引言

通用目标检测应该快速，准确，并且能够识别各种各样的目标。自从引入神经网络以来，检测框架已经变得越来越快速和准确。然而，大多数检测方法仍然局限于一小组目标。

与分类和标记等其他任务的数据集相比，当前目标检测数据集是有限的。最常见的检测数据集包含数十到数十万的图像，具有几十到几百个标签。分类数据集具有数百万个具有数十或数十万类别的图像。

我们希望检测可以缩放到目标分类的级别。然而，用于检测的标记图像比用于分类或标记的标记（标签通常由用户免费提供）昂贵得多。因此，我们不太可能在不久的将来看到与分类数据集相同规模的检测数据集。

我们提出了一种新方法利用我们已经拥有的大量分类数据，并使用它来扩大当前检测系统的范围。我们的方法使用目标分类的层次视图，允许我们将不同的数据集合在一起。

我们还提出了联合训练算法，允许我们在检测和分类数据上训练目标检测器。我们的方法利用标记的检测图像来学习精确地定位目标，同时使用分类图像来增加其词汇和鲁棒性。

使用这种方法，我们训练YOLO9000，一个实时目标检测器，可以检测超过9000不同的目标类别。首先，我们改进基本的YOLO检测系统，以产生YOLOv2，一个最先进的，实时检测器。然后我们使用我们的数据集组合方法和联合训练算法来训练来自ImageNet的超过9000个类的模型以及来自COCO的检测数据。

我们的所有代码和预训练模型都可以在<http://pjreddie.com/yolo9000/>在线获得。

[https://blog.csdn.net/weixin\\_35654926/article/details/72473024](https://blog.csdn.net/weixin_35654926/article/details/72473024)

1/12

联系我们



请扫描二维码联系客服

✉ webmaster@csdn.net

☎ 400-660-0108

💬 QQ客服 🗨 客服论坛

关于 招聘 广告服务 网站地图  
©2018 CSDN版权所有 京ICP证09002463号  
🔍 百度提供支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

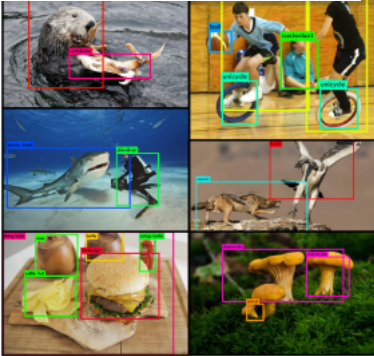


Figure 1: YOLOv2. YOLOv2 can detect a wide variety of object classes in real-time.

## 2、更好

相对于现有技术的检测系统，YOLO具有各种缺点。与fast RCNN相比，YOLO的误差分析显示YOLO产生大量的定位误差。此外，与基于候选区域的方法相比，YOLO具有相对较低的召回率。因此，我们主要集中在改进召回率和定位，同时保持分类精度。

计算机视觉通常趋向于更大，更深的网络。更好的性能通常取决于训练更大的网络或将多个模型组合在一起。然而，使用YOLOv2，我们需要一个更加精确的检测器使得它仍然很快。我们不是扩展我们的网络，而是简化网络，使表示更容易学习。我们从过去的工作中融合了我们自己的新概念的各种想法，以提高YOLO的性能。结果总结可以在表2中找到。

**批量标准化。**批量归一化导致收敛的显著改善，而不需要其他形式的正则化。通过在YOLO中的所有卷积层上添加批量归一化，我们在mAP中获得超过2%的改进效果。批量规范化也有助于规范模型。使用批次标准化，我们可以从模型中dropout，而不会过度拟合。

**高分辨率分类器。**所有最先进的检测方法使用ImageNet预训练分类器。从AlexNet开始，大多数分类器对小于 $256 \times 256$ 的输入图像进行操作[8]。原来的YOLO在 $224 \times 224$ 分辨率上训练分类器网络，并将分辨率增加到448以用于检测。这意味着网络必须同时切换到学习目标检测并调整到新的输入分辨率。

对于YOLOv2，我们首先在分辨率为 $448 \times 448$ 的分辨率下对ImageNet上的10个epoch进行微调。这种网络时间可以在较高分辨率输入上调整滤波器。然后我们在检测时微调所得到的网络。这种高分辨率分类网络使我们增加了近4%的mAP。

**使用anchor box进行卷积。**YOLO直接使用卷积特征提取器顶部的完全连接的层来预测边界框的坐标。相比于直接预测坐标，faster RCNN使用手动挑选的先验预测边界框[15]来预测左边。仅使用卷积层，faster RCNN中的区域建议网络（RPN）预测anchor box的偏移和置信度。由于预测层是卷积的，因此RPN在特征图中的每个位置处预测这些偏移。预测偏移而不是预测坐标简化了问题，并使网络更容易学习。

我们从YOLO中删除全连接层，并使用anchor box预测边界框。首先，我们消除一个池化层，使网络的卷积层的输出更高的分辨率。我们还缩小网络将输入尺寸为416而不是 $448 \times 448$ 。我们这样做是因为我们想要特征图中大小为奇数，所以有一个中心单元格。目标，特别是大目标，倾向于占据图像的中心，所以在中心有一个单一的位置是很好的预测这些目标，而不是四个位置都在中心附近。YOLO的卷积层将图像下采样32倍，所以通过使用输入图像416，我们得到 $13 \times 13$ 的输出特征图。

当我们移动到anchor box时，我们也将类预测机制与空间位置解耦，而代之以预测每个anchor box的类和目标。在YOLO之后，目标预测在假设有一个目前前提下仍然预测ground truth的IOU和提出的框和类预测预测该类的条件概率。

使用anchor box我们得到一个小的精度下降。YOLO每个图像只预测98个box，但使用anchor box我们的模型预测超过一千个box。没有anchor box，我们的中间模型获得69.5 mAP，召回率为81%。使用anchor box我们的模型获得69.2 mAP，召回率为88%，mAP少量减少，召回率的增加意味着我们的模型有更多的改进空间。

联系我们



请扫描二维码联系客服

webmaster@csdn.net

400-660-0108

QQ客服 客服论坛

关于 招聘 广告服务 网站地图  
 ©2018 CSDN版权所有 京ICP证09002463号  
 百度提供支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心



**Figure 2: Clustering box dimensions on VOC and COCO.** We run k-means clustering on the dimensions of bounding boxes to get good priors for our model. The left image shows the average IOU we get with various choices for  $k$ . We find that  $k = 5$  gives a good tradeoff for recall vs. complexity of the model. The right image shows the relative centroids for VOC and COCO. Both sets of priors favor thinner, taller boxes while COCO has greater variation in size than VOC.

图2: VOC和COCO的聚类框尺寸。我们在边界框的维度上运行k均值聚类以获得我们的模型的先验。左图显示了对于k的各种选择得到的平均IOU。我们发现 $k = 5$ 给出了召回率和模型复杂性的良好权衡。右图显示了VOC和COCO的相对质心。两种方法都有利于更薄, 更高的盒子, 而COCO在尺寸上比VOC有更大的变化。

我们将平均IOU与我们的聚类策略和表1中的精选anchor box进行比较。只有5个先验的质心类似于9个anchor box, 平均IOU为61.0, 而9个anchor box为60.9。如果我们使用9个质心, 我们看到一个高得多的平均IOU。这表明使用k-means生成边界框以更好的表示开始模型, 并使任务更容易学习。

Box Generation	#	Avg IOU
Cluster SSE	5	58.7
Cluster IOU	5	61.0
Anchor Boxes [15]	9	60.9
Cluster IOU	9	67.2

表1: VOC 2007上最接近的先验的平均IOU。VOC 2007上目标的平均IOU, 与使用不同生成方法的其最接近的未修改先验。聚类提供比使用手挑选的先验更好的结果。

直接位置预测。当YOLO使用anchor box时我们遇到第二个问题: 模型不稳定, 特别是在早期迭代时。大多数不稳定性来自预测box的 $(x, y)$ 位置。在候选区域网络中, 网络预测值 $tx$ 和 $ty$ 和 $(x, y)$ 中心坐标计算为:

$$x = (tx * wa) - xa, y = (ty * ha) - ya$$

例如,  $tx = 1$ 的预测将使框向右移动anchor box的宽度,  $tx = -1$ 的预测将使其向左移动相同的量。

这种公式是不受约束的, 因此任何anchor box可以在图像中的任何点结束, 而不管预测box的位置。使用随机初始化模型需要很长时间才能稳定到预测可感知的偏移。

相比于预测偏移, 我们遵循YOLO的方法并预测相对于网格单元的位置的位置坐标。这将ground truth限制在0和1之间。我们使用逻辑激活函数来约束网络的预测落在该范围内。

网络预测输出要素图中每个单元格的5个边界框。网络为每个边界框预测 $tx, ty, th, tw$ 和 $to$ 这5个坐标。如果单元从图像的左上角偏移 $(x, y)$ 并且边界框先前具有宽度和高度, 则预测对应于:

由于我们约束位置预测, 参数化更容易学习, 使得网络更稳定。使用维度集群以及直接预测边界框中心位置使YOLO比具有anchor box的版本提高了近5%的mAP。

细粒度特征。该修改的YOLO版本在 $13 \times 13$ 特征图上检测。虽然这对于大目标是足够的, 但是它可以用于定位较小目标的更细粒度特征中受益。Faster RCNN和SSD在网络中的各种特征映射上运行它们的提议网络以获得一系列分辨率。我们采取不同的方法, 只是添加一个传递层, 这个层能够将其他 $26 \times 26$ 分辨率的层融合起来。

联系我们



请扫描二维码联系客服

 webmaster@csdn.net

 400-660-0108

 QQ客服  客服论坛

关于 招聘 广告服务 网站地图

©2018 CSDN版权所有 京ICP证09002463号

 百度提供支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

from cluster centroids. We predict the center coordinates of the box relative to the location of filter application using a sigmoid function.

道而不是堆叠到空间位置，将较高分辨率特征与低分辨率特征相连，类似于ResNet中的标识映射。  
<2048特征映射，其可以与原始特征连接。我们的检测器在这个扩展的特征图的顶部运行，以便它可  
性能提高。

图3：具有维度先验和位置预测的边界框。我们将box的宽度和高度预测为来自聚类中心的偏移。我们使用sigmoid函数预测框相对于过滤器应用的位置的中心坐标。

多尺度训练。原始的YOLO使用448×448的输入分辨率。添加anchor box后，我们将分辨率更改为416×416。然而，由于我们的模型只使用卷积层和池化层，它可以在运行中调整大小。我们希望YOLOv2能够在不同大小的图像上运行，因此我们将其训练到模型中。

相比于固定输入图像大小，我们每隔几次迭代更改网络。每迭代10个batch我们的网络随机选择一个新的图像尺寸大小。由于我们的模型以32的因子下采样，我们从以下32的倍数中抽取：{320,352, ..., 608}。因此，最小的选项是320×320，最大的是608×608.我们调整网络的大小，并继续训练。

这种训练方法迫使网络学习在各种输入维度上很好地预测。这意味着相同的网络可以预测不同分辨率的检测。网络在更小的尺寸下运行更快，因此YOLOv2在速度和精度之间提供了一个简单的折衷。

在低分辨率下，YOLOv2作为一个便宜且相当准确的检测器。在288×288分辨率下它运行超过90 FPS而且mAP几乎与Fast RCNN一样好。这使其成为较小的GPU，高帧率视频或多个视频流的理想选择。

在高分辨率下，YOLOv2是一种最先进的检测器，在VOC 2007上具有78.6 mAP，同时仍然在实时速度以上运行。YOLOv2与其他框架在VOC 2007上的比较见表3。

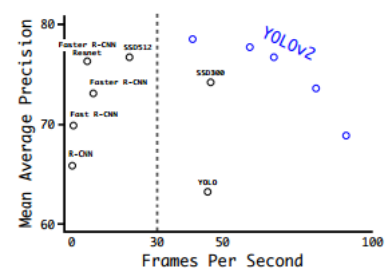


Figure 4: Accuracy and speed on VOC 2007.

进一步的实验。我们训练YOLOv2在VOC 2012上进行检测。表4显示了YOLOv2与其他现有技术检测系统的比较性能.YOLOv2得到73.4 mAP, 而运行速度远远快于其他方法。我们还对COCO进行训练，并与表5中的其他方法进行比较。在VOC指标 (IOU = 0.5) 上，YOLOv2获得44.0 mAP,与SSD和faster RCNN相当。

联系我们



请扫描二维码联系客服  
✉ webmaster@csdn.net  
☎ 400-660-0108  
👤 QQ客服    💬 客服论坛

关于   招聘   广告服务   网站地图

©2018 CSDN版权所有 京ICP证09002463号

🔍 百度提供支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

↓P	FPS
.0	0.5
.2	7
.4	5
.4	45
.3	46
.8	19
.0	91
.7	81
.8	67
.8	59
.6	40

DC 2007.  
ion meth-  
y tradeoff  
tually the  
d at a dif-  
X Titan X

表3: P ASCAL VOC 2007的检测框架.YOLOv2比现有检测方法更快, 更准确。它也可以运行在不同的分辨率上, 以便在速度和精度之间轻松权衡。每个YOLOv2条目实际上是相同的训练模型具有相同的权重, 只是在不同的尺寸进行评估。所有的时间信息是在Geforce GTX Titan X (原始, 而不是Pascal模型)。

3、更快

快速检测。我们希望检测准确, 但我们也希望检测速度快。大多数检测应用, 如机器人或自动驾驶汽车, 都依赖于低延迟预测。为了最大化性能, 我们设计YOLOv2从头开始快。

大多数检测框架依赖于VGG-16作为基本特征提取器[17]。VGG-16是一个功能强大, 精确的分类网络, 但它是不必要的复杂。VGG-16的卷积层需要306.6亿浮点操作用于在224×224分辨率的单个图像上的单次通过。

YOLO框架使用基于Googlenet架构的自定义网络[19]。这个网络比VGG-16快, 只使用85.2亿次操作进行正向传递。但是, 它的精度略差于VGG-16。对于单一目标, 在224×224分辨率上的top-5精度, YOLO的定制模型ImageNet获得88.0%, 而VGG-16为90.0%。

Darknet-19。我们提出了一个新的分类模型作为YOLOv2的基础。我们的模型建立在网络设计的先前工作以及在该领域的常识基础上。类似于VGG模型, 我们使用大多数3×3的过滤器, 并在每个池化步骤后将通道数量加倍[17]。在网络中的网络(NIN)中的工作之后, 我们使用全局平均池进行预测以及1×1滤波器以压缩3×3卷积之间的特征表示[9]。我们使用批次归一化来稳定训练, 加速收敛, 并正则化模型[7]。

我们的最终模型, 称为Darknet-19, 有19卷积层和5个最大池化层。有关完整说明, 请参见表6。Darknet-19只需要55.8亿次操作来处理图像, 但在ImageNet上实现了72.9%的top-1精度和91.2%的top-5精度。

为分类器训练。我们使用以0.1的起始学习速率的随机梯度下降, 使用4的幂的多项式速率衰减, 0.0005的权重衰减和0.9的动量, 我们使用Darknet神经网络框架在标准ImageNet 1000类分类数据集上训练网络[13]160个时期。在训练期间, 我们使用标准数据增加技巧, 包括随机作物, 旋转, 以及色调, 饱和度和曝光移位。

如上所述, 在我们对224×224的图像的初始训练之后, 我们在更大的尺寸如448上微调我们的网络。对于这种微调, 我们用上述参数训练, 但是仅仅10个时期, 并且以的收益率开始。在这个更高的分辨率下, 我们的网络实现了top-1精度为76.5%, top-5精度为93.3%。

为检测器训练。我们通过去除最后的卷积层并且替代地添加具有1024个滤波器的三个3×3卷积层来修改该网络, 每个随后是具有我们需要检测所需的输出数量的最后的1×1卷积层。对于VOC, 我们预测5个box, 每个具有5个坐标, 每个box20个类, 因此125个过滤器。我们还添加了从最后的3×3×512层到第二到最后的卷积层的传递层, 使得我们的模型可以使用细粒度特征。

我们训练网络160个时期, 开始学习率为, 在60和90个时期将其除以10。我们使用0.0005的权重衰减和0.9的动量。我们使用类似的数据增强YOLO和SSD随机作物, 颜色转移等。我们使用相同的培训策略COCO和VOC。





联系我们



请扫描二维码联系客服

✉ webmaster@csdn.net

☎ 400-660-0108

💬 QQ客服 🗨 客服论坛

关于 招聘 广告服务 网站地图  
©2018 CSDN版权所有 京ICP证09002463号  
🔍 百度提供支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

数据集的图像。当我们的网络看到标记为检测的图像时，我们可以基于完整的YOLOv2损失函数反向反向传播从结构的分类特定部分的损失。

只有常用目标和常规标签，如“dog”或“boat”。分类数据集具有更宽和更深的标签范围。ImageNet有“犬”，“约克夏犬”和“贝灵顿梗犬”。如果我们想训练两个数据集，我们需要一种连贯的方式来合并这些

用softmax层来计算最终的概率分布。使用softmax假定类是互斥的。这提出了组合数据集的问题，例COCO，因为类“Norfolk terrier”和“dog”不是互斥的。

互斥的数据集。这种方法忽略了我们所知道的关于数据的所有结构，例如所有的COCO类是相互排斥

at中提取的，WordNet是一个语言数据库，用于构建概念及其关系[12]。在WordNet中，“诺福克猎犬”是一种“猎犬”，是一种“狗”，是一种“犬”分类假设一个平面结构到标签，但是对于组合数据集，

WordNet被构造为有向图，而不是树，因为语言是复杂的。例如，“狗”既是“犬”的一种类型，也是“家畜”的类型，它们都是WordNet中的同义词。不是使用完整的图结构，我们通过从ImageNet中的概念构建层次树来简化问题。

为了构建这个树，我们检查ImageNet中的视觉名词，看看他们通过WordNet图到根节点的路径，在这种情况下是“物理目标”。许多synsets只有一条路径通过图，所以首先我们添加所有这些路径到我们的树。然后我们迭代地检查我们剩下的概念，并添加尽可能少地生长树的路径。因此，如果一个概念有两个到根的路径，一个路径会给我们的树添加三个边，而另一个只添加一个边，我们选择较短的路径。

最终的结果是WordTree，一个视觉概念的层次模型。要使用WordTree执行分类，我们预测在每个节点的条件概率的给定synset的同义词的每个下位词的概率。例如，在“terrier”节点，我们预测：

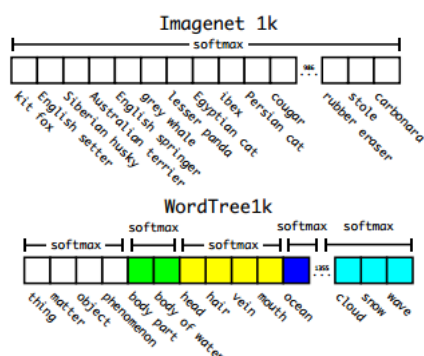
如果我们要计算特定节点的绝对概率，我们只需遵循通过树到达根节点的路径，并乘以条件概率。因此，如果我们想知道图片是否是诺福克梗犬，我们计算：

为了分类的目的，我们假设图像包含一个目标：Pr（物理目标）= 1。

为了验证这种方法，我们训练使用1000类ImageNet构建的WordTree上的Darknet-19模型。为了构建WordTree1k，我们在所有的中间节点中添加将标签空间从1000扩展到1369。在训练期间，我们沿着树传播ground truth标签，以便如果图像被标记为“诺福克梗犬”，它也被标记为“狗”和“哺乳动物”等。为了计算条件概率，我们的模型预测了1369个值的向量，并且我们计算作为相同概念的下位词的所有系统的softmax，参见图5。

使用与以前相同的训练参数，我们的分层Darknet-19实现71.9%的top-1精度和90.4%的top-5精度。尽管增加了369个附加概念，并且我们的网络预测了一个树结构，我们的准确度只有轻微下降。以这种方式执行分类也具有一些益处。性能在新的或未知的目标类别上正常降级。例如，如果网络看到一只狗的图片，但不确定它是什么类型的狗，它仍然会预测具有高置信度的“狗”，但具有较低的置信度散布在上下文词。

这个公式也用于检测。现在，不是假设每个图像都有一个目标，我们使用YOLOv2的目标预测器来给我们Pr（物理目标）的值。检测器预测边界框和概率树。我们遍历树，在每个分割中采用最高置信度路径，直到我们达到某个阈值，我们预测目标类。



**Figure 5: Prediction on ImageNet vs WordTree.** Most ImageNet models use one large softmax to predict a probability distribution. Using WordTree we perform multiple softmax operations over co-hyponyms.

图5：ImageNet对WordTree的预测。大多数ImageNet模型使用一个大的softmax来预测概率分布。使用WordTree，我们对同义词执行多个softmax操作。

与WordTree的数据集组合。我们可以使用WordTree以合理的方式将多个数据集组合在一起。我们只需将数据集中的类别映射到树中的同义词。图6显示了使用WordTree组合来自ImageNet和COCO的标签的示例。WordNet极其多样化，因此我们可以将此技术用于大多数数据集。

联系我们



请扫描二维码联系客服

✉ webmaster@csdn.net

☎ 400-660-0108

💬 QQ客服 客服论坛

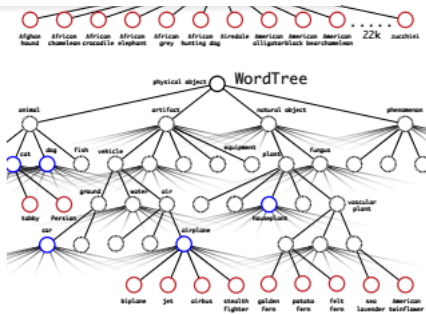
关于 招聘 广告服务 网站地图  
©2018 CSDN版权所有 京ICP证09002463号  
百度提供支持

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心



**Figure 6: Combining datasets using WordTree hierarchy.** Using the WordNet concept graph we build a hierarchical tree of visual concepts. Then we can merge datasets together by mapping the classes in the dataset to synsets in the tree. This is a simplified view of WordTree for illustration purposes.

图6：使用WordTree层次结构组合数据集。使用WordNet概念图，我们构建了一个视觉概念的分层树。然后我们可以通过将数据集中的类映射到树中的synsets来将数据集合并在一起。这是WordTree的简化视图用于说明的目的。

当它看到一个分类图像，我们只反向分配损失。为此，我们只需找到预测该类的最高概率的边界框，然后仅计算其预测树上的损失。我们还假设预测框与grountruth标签重叠的IOU至少0.3，并且基于该假设反向传播物体损失。

使用这种联合训练，YOLO9000使用COCO中的检测数据学习找到图像中的目标，并使用ImageNet中的数据学习分类各种各样的这些目标。

我们在ImageNet检测任务上评估YOLO9000。ImageNet的检测任务共享44个具有COCO的目标类别，这意味着YOLO9000只看到大多数测试图像的分类数据，而不是检测数据。YOLO9000获得19.7 mAP整体与16.0 mAP对不相交的156目标类，它从未见过任何标记的检测数据。这个mAP高于DPM实现的结果，但YOLO9000是在不同的数据集训练，只有部分监督[4]。它还同时检测9000个其他目标类别，都是实时的。

当我们分析YOLO9000在ImageNet上的性能时，我们看到它学习了新的物种，但很难学习类别，如服装和设备。

新动物更容易学习，因为目标预测与COCO中的动物很好地一致。相反，COCO没有任何类型的衣服的边界框标签，只有人，所以YOLO9000努力模拟类似“太阳镜”或“游泳裤”的类别。

表7：ImageNet上的YOLO9000最佳和最差类。具有来自156个弱监督类的最高和最低AP的类。YOLO9000学习各种动物的好模型，但努力与新的类，如服装或设备。

## 5、结论

我们介绍YOLOv2和YOLO9000，实时检测系统。YOLOv2是最先进的，并且比其他检测系统在各种检测数据集中更快。此外，它可以以各种图像大小运行，以提供速度和精度之间的平滑权衡。

YOLO9000是一个通过联合优化检测和分类检测9000多个目标类别的实时框架。我们使用WordTree来组合来自各种来源的数据和我们的联合优化技术同时训练ImageNet和COCO。YOLO9000是关闭检测和分类之间的数据集大小差距的强大步骤。

我们的许多技术泛化到目标检测之外。ImageNet的ImageTree表示为图像分类提供了更丰富，更详细的输出空间。使用分层分类的地形组合在分类和分割领域将是有益的。诸如多尺度训练的训练技术可以在各种视觉任务中提供益处。

对于未来的工作，我们希望使用类似的技术弱监督图像分割。我们还计划使用更强大的匹配策略来改进我们的检测结果，以在训练期间将弱标签分配给分类数据。计算机视觉有大量的标记数据。我们将继续寻找方法，将不同的数据源和结构的数据结合在一起，形成更强大的视觉世界模型。