**Yolo1的整理解剖**

**你只需看一次：统一的实时对象检测**

论文作者.Joseph Redmon

摘要

我们提出了YOLO，一种全新的目标检测方法。以往的目标检测工作将分类器重新用于检测任务。相反，我们将目标检测定义为对空间分离的边界框及其类别概率的回归问题。单个神经网络直接从完整图像中一次性预测边界框和类别概率。由于整个检测流程是一个单一网络，可以直接端到端优化检测性能。

我们的统一架构速度极快。基础YOLO模型以每秒45帧的速度实时处理图像。其轻量版本Fast YOLO更是达到每秒155帧，同时仍保持其他实时检测器两倍的mAP（平均精度）。与最先进的检测系统相比，YOLO的定位误差较多，但对背景的误检（假阳性）更少。此外，YOLO能学习到非常通用的物体表征。当从自然图像泛化到艺术作品等其他领域时，YOLO的表现优于DPM和R-CNN等方法。

1. 引言

人类只需扫视图像便能立即知晓其中的物体、位置及其交互关系。人类视觉系统快速且精准，使我们无需过多思考即可完成驾驶等复杂任务。快速、精准的目标检测算法将使计算机无需专用传感器即可驾驶汽车，辅助设备能够向用户实时传递场景信息，并释放通用响应式机器人系统的潜力。

现有检测系统通过改造分类器实现检测。为了检测物体，这些系统将针对该物体的分类器在测试图像的不同位置和尺度上进行评估。例如，可变形部件模型（DPM）采用滑动窗口方法，在整个图像上均匀分布的位置运行分类器[10]。

更近期的R-CNN等方法使用区域提议方法首先生成图像中的潜在边界框，然后在这些框上运行分类器。分类后，需通过后处理优化边界框、消除重复检测并根据场景中的其他物体重新评分[13]。这些复杂流程速度缓慢且难以优化，因为每个独立组件需单独训练。

我们将目标检测重新定义为从图像像素到边界框坐标及类别概率的单一回归问题。使用我们的系统YOLO（You Only Look Once），只需对图像“看一次”即可预测存在的物体及其位置。

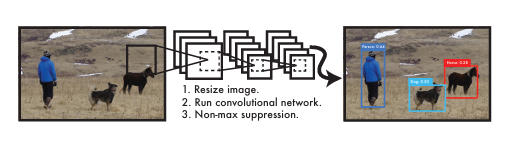
 （图一）

图 1：YOLO 检测系统。使用 YOLO 处理图像简单明了。我们的系统 （1） 将输入图像的大小调整为 448 × 448，（2） 对图像运行单个卷积网络工作，以及 （3） 根据模型的置信度对结果检测进行阈值处理。

YOLO的设计简洁直观（见图1）。单一卷积网络同时预测多个边界框及这些框的类别概率。YOLO在完整图像上训练，并直接优化检测性能。这种统一模型相较于传统方法具有多项优势。

第一，YOLO速度极快。由于将检测视为回归问题，无需复杂流程。测试时只需对新图像运行神经网络即可预测检测结果。基础网络在Titan X GPU上以45帧/秒的速度运行（无需批处理），快速版本可达155帧/秒。这意味着我们能够以低于25毫秒的延迟实时处理视频流。此外，YOLO的mAP是其他实时系统的两倍以上。

第二，YOLO在预测时对图像进行全局推理。不同于滑动窗口和基于区域提议的技术，YOLO在训练和测试时均能看到完整图像，因此隐式编码了类别的上下文信息及其外观。顶尖检测方法Fast R-CNN[14]因缺乏全局上下文会将图像背景误判为物体，而YOLO的背景误检率仅为Fast R-CNN的一半。

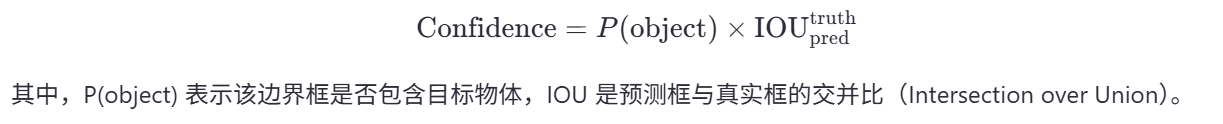
第三，YOLO学习到物体的泛化表征。当在自然图像上训练并在艺术作品上测试时，YOLO以显著优势超越DPM和R-CNN。由于YOLO泛化能力强，应用于新领域或意外输入时更不易失效。

YOLO在精度上仍落后于最先进的检测系统。虽然能快速识别图像中的物体，但对某些物体（尤其是小物体）的精确定位仍有困难。我们通过实验进一步分析了这些权衡。所有训练和测试代码均已开源，多种预训练模型可供下载。

2.统一检测

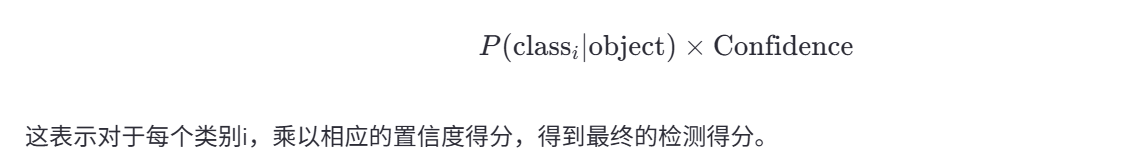
我们将目标检测的各个组件整合到单一神经网络中。网络利用图像全局特征预测每个边界框，并同时预测图像中所有类别的所有边界框。这意味着网络能对全图及其中所有物体进行全局推理。YOLO的设计实现了端到端训练、实时速度与高平均精度的平衡。

系统将输入图像划分为S×S个网格。若物体中心落入某个网格单元，则该单元负责检测该物体。通常为7\*7=49个单元格；groundtruth的中心落在那个单元格，就由那个单元格负责这个物体的检测即预测框的拟合整理。

每个网格单元预测**B（通常为2）**个边界框及这些框的置信度。置信度反映模型对框内存在物体的信心以及预测框的准确性。正式定义为：若单元内无物体，置信度应为零即C=0；否则，置信度应等于预测框与真实框的交并比（IOU在0-1之间）。

每个边界框包含5个预测值：x, y, w, h和置信度c。 (x, y)坐标表示框中心相对于网格单元的边界位置，宽度w和高度h相对于整张图像预测。置信度表示预测框与真实框的IOU。

每个网格单元还预测C（在版本1中为20）个条件类别概率Pr(Class i | Object),这些概率以网格单元包含物体为条件。无论B的数量如何，每个网格单元仅预测一组类别概率（20个类别概率中最高者）。

测试时，我们将条件类别概率与各框置信度相乘：即类别概率与交并比的积为最终得分，

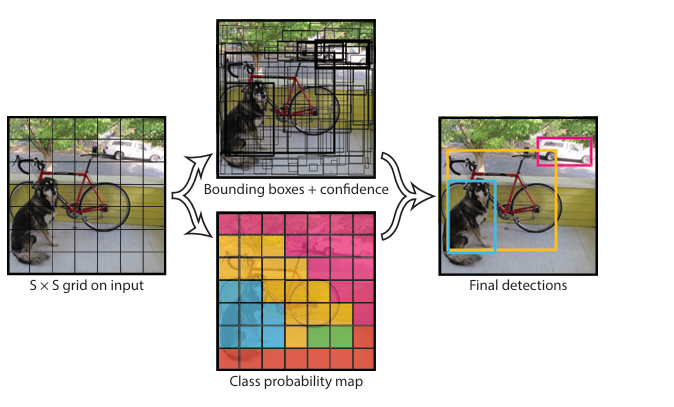
得到每个框的类别特异性置信度。这些分数同时编码了该类别出现在框中的概率以及预测框与物体的匹配程度。  (图二)

图 2：模型。我们的系统将检测建模为回归问题。它将图像划分为 S ×S 网格，并针对每个网格单元格预测 B 边界框、这些框的置信度和 C 类概率。这些预测被编码为 S ×S×（B×5+C）的张量。

在Pascal VOC上评估YOLO时，使用S=7, B=2。Pascal VOC有20个标注类别，因此C=20。最终预测为7×7×（2\*5+20）的张量即三维向量。

2.1网络设计

我们将该模型实现为卷积神经网络，并在Pascal VOC检测数据集[9]上评估。网络的前端卷积层提取图像特征，全连接层预测输出概率和坐标。

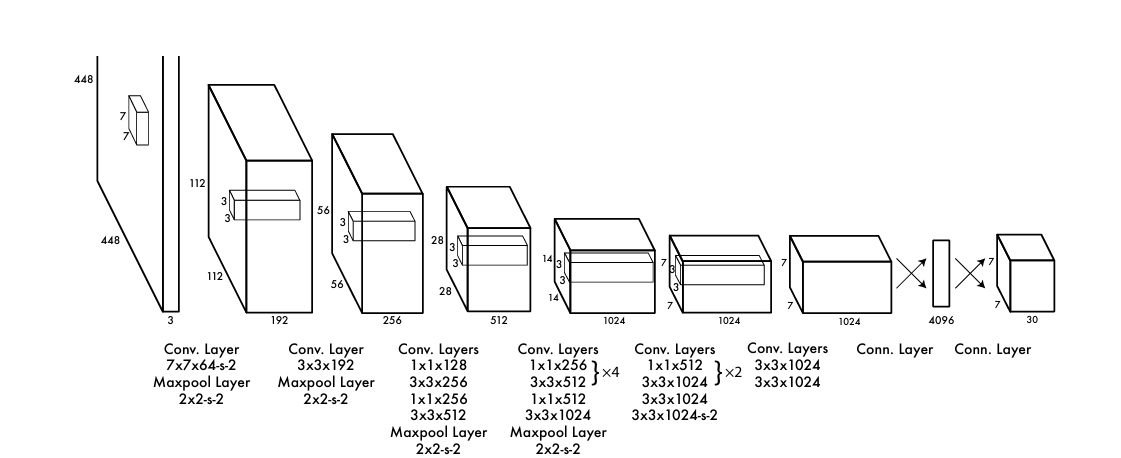
网络架构受GoogLeNet图像分类模型启发[33]。网络包含24个卷积层和2个全连接层。未使用GoogLeNet的Inception模块，而是采用类似Lin et al [22]的1×1降维层接3×3卷积层。完整架构见图3。 (图3)

图 3：架构。我们的检测网络有 24 个卷积层，后跟 2 个全连接层。交替使用 1 × 1 卷积层可减少前一层的特征空间。我们以一半的分辨率（224 × 224 输入图像）对 ImageNet 分类任务上的卷积层进行预训练，然后将分辨率提高一倍进行检测。

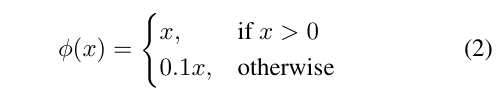
我们还训练了Fast YOLO以探索快速检测的极限。Fast YOLO使用更少的卷积层（9层而非24层）和更少的滤波器。除网络规模外，YOLO与Fast YOLO的训练和测试参数完全相同。

2.2训练

我们首先在ImageNet 1000类竞赛数据集[29]上预训练卷积层。预训练使用图3的前20个卷积层，后接平均池化层和全连接层。训练约一周后，在ImageNet 2012验证集上达到单裁剪图top-5精度88%，与Caffe模型库中的GoogLeNet模型相当[24]。

然后将模型转换为检测任务。Ren表明，在预训练网络中增加卷积层和全连接层可提升性能[28]。遵循其方法，我们添加4个卷积层和2个随机初始化权重的全连接层。检测需要细粒度视觉信息，因此将输入分辨率从224 ×224提升至448×448。

最终层预测类别概率和边界框坐标。我们将边界框的宽度和高度归一化为图像尺寸的比例（0到1之间），并将x和y坐标参数化为特定网格单元的偏移量（同样限制在0到1之间）。

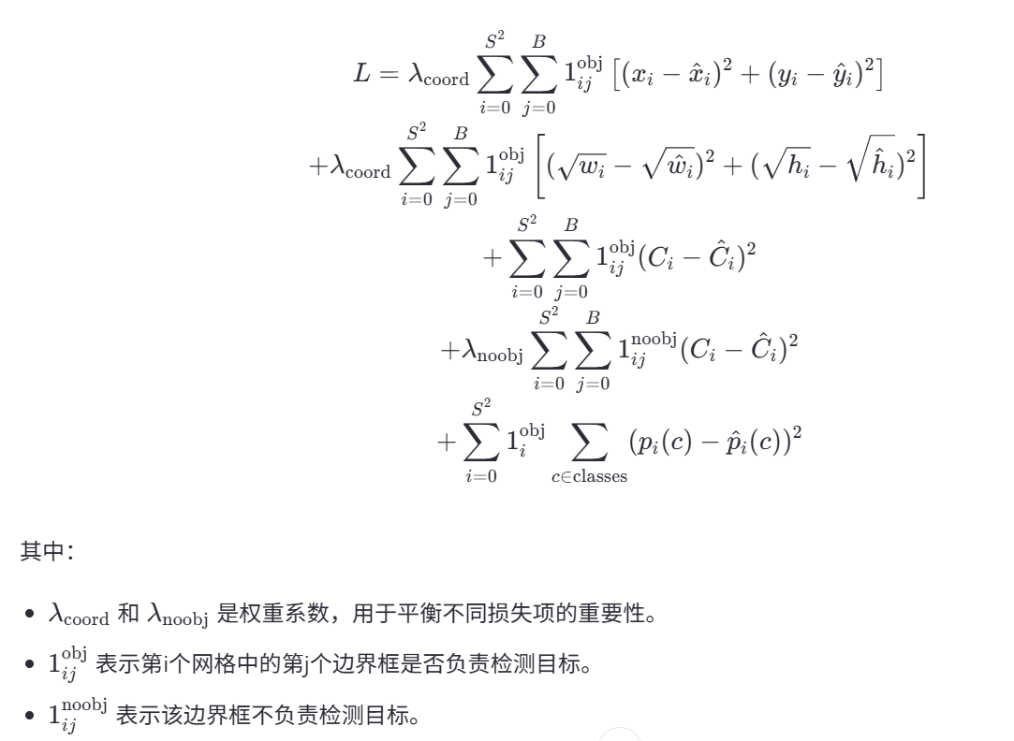
最终层使用线性激活函数，其余层使用带泄漏的修正线性激活： 

我们通过输出端的均方误差进行优化。虽然均方误差易于优化，但与最大化平均精度的目标不完全一致。它将定位误差与分类误差同等加权，而分类误差可能并不理想。此外，在每个图像中，许多网格单元不包含任何对象。这会将这些单元格的 “置信度” 分数推向零，通常会压倒包含对象的单元格的梯度。这可能会导致模型不稳定，从而导致训练在早期出现分歧。

为此，我们增加边界框坐标预测的损失权重，并降低不含物体框的置信度损失权重，通过参数λcoord and λnoobj实现。我们设置 λcoord = 5 和 λnoobj = 5。

和平方误差在大框和小框中的权重也相等。我们的误差指标应该反映出，大盒子中的小偏差比小盒子中的小偏差更不重要。为了部分解决这个问题，我们预测边界框宽度和高度的平方根，而不是直接预测宽度和高度。

YOLO 预测每个网格单元格的多个边界框。在训练时，我们只希望一个边界框预测器负责每个对象。我们分配一个预测器“负责”预测对象，根据哪个预测具有最高的当前 IOU 和真实值。这会导致边界框预测变量之间的特化。每个预测器在预测某些大小、纵横比或对象类别方面做得更好，从而提高了整体召回率。

训练时，我们优化以下多部分损失函数：

请注意，如果该网格单元中存在对象，则损失函数仅惩罚分类误差（因此前面讨论的条件类概率）。如果该预测器“负责”地面实况框（即在该网格单元格中具有最高的 IOU），它也只会惩罚边界框坐标错误。

我们在Pascal VOC 2007和2012的训练集和验证集上训练约135轮。测试2012时，训练数据也包含VOC 2007测试集。训练使用批量大小64、动量0.9和权重衰减0.0005。

学习率调度如下：前几轮从0.001缓慢升至0.01（避免梯度不稳定导致模型发散），随后以0.01训练75轮，0.001训练30轮，最后0.0001训练30轮。

为防止过拟合，使用Dropout和数据增强。首个全连接层后使用Dropout（比率0.5）[18]。数据增强包括随机缩放和平移（最多原图20%），以及HSV色彩空间中随机调整曝光和饱和度（因子1.5）。为避免过拟合，我们使用 dropout 和广泛的数据增强。在第一个连接层之后 rate=.5 的 dropout 层会阻止层之间的协适应[18]。对于数据增强，我们引入了随机缩放和高达原始图像大小的 20% 的转换。我们还在 HSV 色彩空间中随机调整图像的曝光和饱和度，最高可达 1.5 倍。

2.3推理

与训练类似，测试图像只需一次网络前向传播。在Pascal VOC上，网络每张图像预测98个边界框及各类别概率。由于只需单次推理，YOLO测试速度极快。

网格设计强制边界框预测的空间多样性。通常物体明显落入某个网格单元，网络仅为其预测一个框。但大型物体或靠近多个单元边界的物体可能被多个单元准确定位。非极大值抑制（NMS）可用于消除重复检测。虽然对性能影响不如R-CNN或DPM显著，但NMS仍能提升2-3%的mAP。

2.4 YOLO的局限性

YOLO对边界框预测施加了强空间约束，因为每个网格单元仅预测两个框且只能有一个类别。这种约束限制了模型可预测的邻近物体数量。模型对成群的小物体（如鸟群）检测困难。

由于模型从数据中学习预测边界框，难以泛化到新长宽比或配置的物体。此外，因架构中存在多个下采样层，预测边界框时使用的特征相对粗糙。由于我们的模型学习从数据中预测边界框，因此它很难推广到具有新的或不寻常的纵横比或配置的对象。我们的模型还使用相对粗略的特征来预测边界框，因为我们的架构从输入图像中有多个下采样层。

最后，尽管损失函数近似检测性能，但均方误差平等对待大小框的误差。大框的小误差通常无害，但小框的小误差对IOU影响更大。YOLO的主要误差来源是定位错误。

3. 与其他检测系统的比较

目标检测是计算机视觉的核心问题。检测流程通常从提取图像鲁棒特征（Haar[25]、SIFT[23]、HOG[4]、卷积特征[6]）开始，然后使用分类器[35, 21, 13, 10]或定位器[1, 31]在特征空间中识别物体。这些分类器或定位器以滑动窗口方式遍历图像或在区域子集上运行[34, 15, 38]。我们将YOLO与几种顶尖检测框架对比，突出关键异同。

可变形部件模型（DPM）：使用滑动窗口进行检测[10]，其流程包含特征提取、区域分类、高分区域框预测等独立步骤。YOLO用单一卷积网络替代所有组件，同步执行特征提取、框预测、NMS和上下文推理。网络动态训练特征并针对检测任务优化，从而获得比DPM更快、更精确的模型。

R-CNN：及其变体使用区域提议而非滑动窗口。Selective Search[34]生成潜在边界框，卷积网络提取特征，SVM评分，线性模型调整框，NMS消除重复检测。该流程复杂、速度慢（测试时每图超40秒[14]）。

YOLO与R-CNN的相似之处在于均利用卷积特征对提议框评分。但YOLO通过网格单元的空间约束减少重复检测，且每图仅生成98个框（远低于Selective Search的约2000个）。所有组件整合为单一联合优化模型。

其他快速检测器：如Fast/Faster R-CNN通过共享计算和神经网络生成区域提议加速R-CNN框架[14][27]，但仍无法实时。

许多研究优化DPM流程[30][37][5]，如加速HOG计算、级联和使用GPU，但仅30Hz DPM[30]真正实时。

YOLO摒弃传统流程，通过设计实现快速。单类别检测器（如人脸）可高度优化，而YOLO是通用检测器，能同时检测多类物体。

Deep MultiBox：训练卷积网络生成区域提议[8]，但仅支持单目标检测且需额外分类步骤。YOLO则是完整检测系统。

OverFeat：训练卷积网络进行定位并适配为检测器[31]，但仍是分离系统。其优化目标为定位而非检测，且缺乏全局上下文。

MultiGrasp：设计与YOLO类似[26]，但仅预测单一抓取区域，无需处理物体尺寸、位置或类别。YOLO则预测多类多物体的边界框和类别概率。

4. 实验

4.1 与其他实时系统的比较

在Pascal VOC 2007上，YOLO的Fast版本（Fast YOLO）以155帧/秒和52.7% mAP成为最快的检测器，精度是其他实时方法的两倍以上。基础YOLO以45帧/秒达到63.4% mAP。使用VGG-16的YOLO精度更高但速度显著下降。

4.2 VOC 2007误差分析

YOLO的主要误差为定位错误（占所有误差的19%），而Fast R-CNN的背景误检率更高（13.6% vs YOLO的4.75%）。结合YOLO与Fast R-CNN（用YOLO剔除背景误检），mAP提升3.2%至75.0%。

4.3-4.5 泛化性与结果

在艺术作品数据集（Picasso和People-Art）上，YOLO的泛化性能优于R-CNN和DPM。尽管像素级差异大，但YOLO对物体尺寸和形状的建模使其在艺术作品中仍能有效检测。

5. 实时野外检测

YOLO连接摄像头时可保持实时性能（含图像采集和结果显示），演示视频和代码详见项目主页。

6. 结论

YOLO是统一的目标检测模型，结构简单、可端到端训练。Fast YOLO是文献中最快的通用检测器，YOLO则推动实时检测的技术前沿。其强泛化能力使其适用于依赖快速鲁棒检测的应用场景。