基于yolov2算法的深度学习目标识别

摘要：

我们基于yolov2算法进行目标识别，yolov2是yolo的改进版。yolo是一种

把目标识别问题转化为回归问题的深度学习方法，有别于rcnn系列目标识别算法，也有别于基于DPM的传统目标识别算法。rcnn系列算法 采取proposal+分类的思路，proposal提供位置信息，分类对proposal进行分类。DPM是一种基于 提取HOG特征+使用窗口滑动HOG特征+分类窗口 的目标识别算法。yolo使用卷积神经网络进行图片的特征提取，在输出层回归bounding box的位置和box的类别。yolov2在yolo的基础上，提出了一系列改进的方法，在处理速度不变的情况下，提高了检测的精度。我们主要改进了yolov2的网络结构，去掉一些冗余的中间层，加快识别的速度，精度基本可以保持一致。

介绍

人类只需要看一眼，便可以立马识别出图像中的物体是什么，并且知道他们在哪里。人类的视力系统识别效率如此的高和快速，可以让我们很快识别出复杂的物体。快速，高效的目标检测算法可以让计算机实时识别出视频中各种物体，如车辆，行人，可以作为了机器人的视觉系统。

现在目标识别系统，都是基于分类来做的。为了识别目标，需要使用固定大小的窗口滑动整个图片，使用分类模型判断每个窗口所属类别的自信度，DPM目标识别模型便是依据这样的思路。rcnn系列模型先产生很多后选框，分类每个后选框，最后通过预测的张量调整每个后选框的位置来得出最后的结果。

yolo将目标识别任务作为回归问题来处理，直接根据图像的像素值来预测目标区域的位置以及所属类别的概率。只需要扫描一次图片便可以识别出目标。检测过程如图1，相比较其他目标识别方法，yolo具有以下优点。

图1： 目标识别过程： 1）将图片调整为448\*448， 2） 使用卷积网络提取特征 3）通过模型得出的自信度来识别目标。

1）检测速度快，可以实时检测目标，在Titan X GPU上可以跑45帧每秒，并且识别效果是其他实时识别的两倍。

2）滑动窗口和列举后选框都是基于局部信息来进行分类，经常会把背景识别为目标，yolo通过全局图像信息来识别目标，可以减少背景的错误识别率。

3）yolo泛化能力强，当预测没有训练过的目标时，效果会好些。

yolov2在yolo的基础上，提出了一系列方法改进了，在保持原有的速度的同时，提高了识别精度。主要在一下方面做了改进：

　　　1）把图片输入分辨率改为416 \* 416，目的是让后面产生的卷积特征图宽高都为奇数，这样就可以产生一个center cell。统计发现大物体通常占据了图像的中间位置，可以只用一个中心的cell来预测这些物体的位置，否则就要用中间的4个cell来进行预测，这个技巧可稍稍提升效率。

1）在每个的卷基层后面添加了batch normalization层，来归一化中间数据，使得输入数据，中间数据分布大致相同。通过添加batch normalization层，极大的加快了收敛速度。

2） 提高了输入图片的分辨率，AlexNet输入图片会被resize到不足256 \* 256，这导致分辨率不够高，给检测带来困难。yolo使用分辨率为448\*448的图片作为输入。

3）借鉴了Faster R-CNN中的anchor思想，在卷积特征图上进行滑窗采样，每个中心预测5种不同大小和比例的后选框。由于都是卷积不需要reshape，很好的保留的空间信息，最终特征图的每个特征点和原图的每个cell一一对应。而且用预测相对偏移（offset）取代直接预测坐标简化了问题，方便网络学习。

我们改进了网络结构，去除一些冗余的层。主要将几个卷基层去掉了。直接训练检测网络，没有先训练分类网络，再微调检测网络，取得的效果基本和yolov2保持一致。速度更快，权重模型从200M降到48M，速度更快。

使用卷积层产生候选框

YOLOv1使用全连接层的输出进行bounding box的预测，即要把1470\*1的全链接层reshape为7\*7\*30特征），这样导致丢失了很多空间信息，定位不准去。YOLOv2借鉴了Faster R-CNN中的anchor思想： 简单理解为 在卷积特征图上进行滑窗采样，每个中心预测9种不同大小和比例的候选框。由于卷基层是直接预测，不需要reshape，很好的保留了空间信息，最终特征图的每个特征点和原图的每个cell一一对应。如图2

维度聚类

Yolov2借鉴了anchor的思想，Faster-RCNN中anchor boxes的个数和宽高维度是手动精选的先验框。为了能够选择了更好的、更有代表性的boxes维度，yolov2使用K-means聚类方法，通过对数据集中的ground true box做聚类，找到ground true box的统计规律。以聚类个数k为anchor boxs个数，以k个聚类中心box的宽高维度为anchor box的维度。

聚类的真正想要的是产生好的IOU得分的boxes。因此采用了如下距离度量：

　　　d(box; centroid) = 1 - IOU(box; centroid)

聚类结果如图3：

图3： 随着k的增大，IOU也在增大，但是复杂度也在增加。所以平衡复杂度和IOU之后，最终得到k值为5。

更快速

大多数检测网络依赖于VGG-16作为特征提取网络，VGG-16是一个强大而准确的分类网络，但是确过于复杂，计算量太大，速度慢。224 \* 224的图片进行一次前向传播，其卷积层就需要多达306.9亿次浮点数运算。

YOLO使用的是基于Googlenet的自定制网络，比VGG-16更快，一次前向传播仅需85.2亿次运算，不过它的精度要略低于VGG-16。224 \* 224图片取 single-crop, top-5 accuracy，YOLO的定制网络得到88%（VGG-16得到90%）。

Darknet-19

YOLOv2使用了一个新的分类网络作为特征提取部分，参考了前人的工作经验。类似于VGG，网络使用了较多的3 \* 3卷积核，在每一次池化操作后把通道数翻倍。借鉴了network in network的思想，网络使用了全局平均池化（global average pooling）做预测，把1 \* 1的卷积核置于3 \* 3的卷积核之间，用来压缩特征。使用batch normalization稳定模型训练，加速收敛，正则化模型。

最终得出的基础模型就是Darknet-19，包含19个卷积层、5个最大值池化层（max pooling layers ）。Darknet-19处理一张照片需要55.8亿次运算，imagenet的top-1准确率为72.9%，top-5准确率为91.2%。

网络结构如图4

更强大

yolov2作者 提出了一种在分类数据集和检测数据集上联合训练的机制。使用检测数据集的图片去学习检测相关的信息，例如bounding box 坐标预测，是否包含物体以及属于各个物体的概率。使用仅有类别标签的分类数据集图片去扩展可以检测的种类。

训练过程中把监测数据和分类数据混合在一起。当网络遇到一张属于检测数据集的图片就基于YOLOv2的全部损失函数（包含分类部分和检测部分）做反向传播。当网络遇到一张属于分类数据集的图片就仅基于分类部分的损失函数做反向传播。

这种方法有一些难点需要解决。检测数据集只有常见物体和抽象标签，例如 “狗”，“船”。分类数据集拥有广而深的标签范围（例如ImageNet就有一百多类狗的品种，包括 “Norfolk terrier”, “Yorkshire terrier”, and “Bedlington terrier”等. ）。必须按照某种一致的方式来整合两类标签。

大多数分类的方法采用softmax层，考虑所有可能的种类计算最终的概率分布。但是softmax假设类别之间互不包含，但是整合之后的数据是类别是有包含关系的，例如 “Norfolk terrier” 和 “dog”。 所以整合数据集没法使用这种方式（softmax 模型），作者最后采用一种不要求互不包含的多标签模型（multi-label model）来整合数据集。这种方法忽略了数据集的结构（例如 COCO数据集的所有类别之间是互不包含的）

层次式分类

ImageNet的标签参考WordNet（一种结构化概念及概念之间关系的语言数据库）。如图5

很多分类数据集采用扁平化的标签。而整合数据集则需要结构化标签。

WordNet是一个有向图结构，因为语言是复杂的（例如“dog”既是“canine”又是“domestic animal”），为了简化问题，作者从ImageNet的概念中构建了一个层次树结构来代替图结构方案。

创建层次树：

遍历ImageNet的所有视觉名词

对每一个名词，在WordNet上找到从它所在位置到根节点的路径。 许多同义词集只有一条路径。所以先把这些路径加入层次树结构。

然后迭代检查剩下的名词，得到路径，逐个加入到层次树。路径选择办法是：如果一个名词有两条路径到根节点，其中一条需要添加3个边到层次树，另一条仅需添加一条边，那么就选择添加边数少的那条路径。

最终结果是一颗 WordTree （视觉名词组成的层次结构模型）。用WordTree执行分类时，预测每个节点的条件概率。例如： 在“terrier”节点会预测：图6

如果想求得特定节点的绝对概率，只需要沿着路径做连续乘积。例如 如果想知道一张图片是不是“Norfolk terrier ”需要计算：图7

分类时假设 图片包含物体：Pr(physical object) = 1.

为了验证这种方法作者在WordTree（用1000类别的ImageNet创建）上训练了Darknet-19模型。为了创建WordTree1k作者天添加了很多中间节点，把标签由1000扩展到1369。训练过程中ground truth标签要顺着向根节点的路径传播：例如 如果一张图片被标记为“Norfolk terrier”它也被标记为“dog” 和“mammal”等。为了计算条件概率，模型预测了一个包含1369个元素的向量，而且基于所有“同义词集”计算softmax，其中“同义词集”是同一概念的下位词。图8：

使用相同的训练参数，层次式Darknet-19获得71.9%的top-1精度和90.4%top-5精度。尽管添加了369个额外概念，且让网络去预测树形结构，精度只有略微降低。按照这种方式执行分类有一些好处，当遇到新的或未知物体类别，预测精确度降低的很温和（没有突然巨幅下降）。例如：如果网络看到一张狗的图片，但是不确定狗的类别，网络预测为狗的置信度依然很高，但是，狗的下位词（“xx狗”）的置信度就比较低。

这个策略野同样可用于检测。不在假设每一张图片都包含物体，取而代之使用YOLOv2的物体预测器（objectness predictor）得到Pr(physical object)的值。检测器预测一个bounding box和概率树（WordTree）。沿着根节点向下每次都走置信度最高的分支直到达到某个阈值，最终预测物体的类别为最后的节点类别。

使用wordTree聚合数据

可以使用WordTree把多个数据集整合在一起。只需要把数据集中的类别映射到树结构中的同义词集合（synsets）。使用WordTree整合ImageNet和COCO的标签如下图：图9

联合训练分类和检测

使用WordTree整合了数据集之后就可以在数据集（分类-检测数据）上训练联合模型。我们想要训练一个检测类别很大的检测器所以使用COCO检测数据集和全部ImageNet的前9000类创造一个联合数据集。为了评估我们使用的方法，也从ImageNet detection challenge 中向整合数据集添加一些还没有存在于整合数据集的类别。相应的WordTree有9418个类别。由于ImageNet是一个非常大的数据集，所以通过oversampling COCO数据集来保持平衡，使ImageNet：COCO = 4：1。

使用上面的数据集训练YOLO9000。采用基本YOLOv2的结构，anchor box数量由5调整为3用以限制输出大小。

当网络遇到一张检测图片就正常反向传播。其中对于分类损失只在当前及其路径以上对应的节点类别上进行反向传播。

当网络遇到一张分类图片仅反向传播分类损失。在该类别对应的所有bounding box中找到一个置信度最高的（作为预测坐标），同样只反向传播该类及其路径以上对应节点的类别损失。反向传播objectness损失基于如下假设：预测box与ground truth box的重叠度至少0.31IOU。

采用这种联合训练，YOLO9000从COCO检测数据集中学习如何在图片中寻找物体，从ImageNet数据集中学习更广泛的物体分类。

作者在ImageNet detection task上评估YOLO9000。ImageNet detection task和COCO有44个物体类别是相同的。这意味着YOLO9000只从大多数测试数据集中看到过分类数据而非检测数据。最终整体精度为19.7mAP，在从未见过的156个物体检测数据类别上精度为16.0mAP。这个结果高于DPM，但是YOLO9000是在不同数据集上进行半监督训练。而且YOLO9000可以同时实时检测9000多种其它物体类别。

作者也分析了YOLO9000在ImageNet上的性能，发现可以学习新的动物表现很好，但是学习衣服和设备这类物体则不行。因为从COCO数据集上动物类别那里学习到的物体预测泛化性很好。但是COCO数据集并没有任何衣服类别的标签数据（只有"人"类别），所以YOLO9000很难对“太阳镜”，“游泳裤”这些类别建模。

改进yolov2网络结构：

我们以Darknet-19网络结构为基础，对它进行了改进，去掉了一些中间层，虽然效果有点受影响，但是处理速度更快，更加满足实时性的要求。

　　　 改进后的网络结构，包含13个卷积层、5个最大值池化层（max pooling layers ）。改进后的网络模型的权重只有56M，Darknet-19网络的模型权重200M，处理时间也是Darknet-19的2倍， 网络结构更小， 处理速度更快，可以更好的满足实时性要求。

网络结构如图

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Filters | Size/Stride | Output |
| Convolutional  Maxpool  Convolutional  Maxpool  Convolutional  Convolutional  Maxpool  Convolutional  Convolutional  Maxpool  Convolutional  Convolutional  Convolutional  Maxpool  Convolutional  Convolutional  Convolutional | 32  64  128  128  256  256  512  512  512  512  512  1024 | 3 × 3  2 × 2 / 2  3 × 3  2 × 2 / 2  3 × 3  3 × 3  2 × 2 / 2  3 × 3  3 × 3  2 × 2 / 2  3 × 3  1 × 1  3 × 3  2 × 2 / 2  3 × 3  1 × 1  3 × 3 | 416 × 416  208 × 208  208 × 208  104 × 104  104 × 104  104 × 104  52 × 52  52 × 52  52 × 52  26 × 26  26 × 26  26 × 26  26 × 26  13 × 13  13 × 13  13 × 13  13 × 13 |
| RegionLoss |  |  |  |

训练框架caffe：

 Caffe，全称Convolutional Architecture for Fast Feature Embedding，是一个计算CNN相关算法的框架，由Yangqing Jia老师编写和维护的，代替了之前的decaf工具。Caffe是用C++和Python实现的，并提供了C++、Python、Matlab的接口，目前有Linux 和Windows版。

它是一个清晰、高效的深度学习框架，它是开源的，核心语言是C++，既可以在CPU上运行也可以在GPU上运行。它的license是BSD 2-Clause。

改进网络的训练：

我们没有使用yolov2那种复杂的训练数据集，我们直接使用Voc2012 voc2007作为训练的数据集。我们直接在voc数据集上，进行目标识别的训练，没有向yolov2那样，先训练分类，然后使用训练好的权重模型，微调目标识别模型。

我们使用caffe框架作为我们的训练平台。使用多阶段学习率方式进行训练，学习率在100内为0.0001，23000内为0.001，35000内为 0.0001，45000内为0.00001，设置平均损失为25，设置batch\_size为64，并使用GPU来加速训练。

数据集：

PASCAL VOC为图像识别和分类提供了一整套标准化的优秀的数据集，从2005年到2012年每年都会举行一场图像识别challenge。

我们采用PASCAL VOC2012作为例子。

下载VOC2012完之后解压，可以在VOCdevkit目录下的VOC2012中看到如下的文件夹：

　　　图11

　　　其中在图像物体识别上着重需要了解的是Annotations、ImageSets和JPEGImages。

　　　JPEGImages：JPEGImages文件夹中包含了PASCAL VOC所提供的所有的图片信息，包括了训练图片和测试图片。

　　　图11

这些图像都是以“年份\_编号.jpg”格式命名的。

图片的像素尺寸大小不一，但是横向图的尺寸大约在500\*375左右，纵向图的尺寸大约在375\*500左右，基本不会偏差超过100。在之后的训练中，第一步就是将这些图片都resize到416\*416，所有原始图片不能离这个标准过远。

这些图像就是用来进行训练和测试验证的图像数据。

Annotations：

Annotations文件夹中存放的是xml格式的标签文件，每一个xml文件都对应于JPEGImages文件夹中的一张图片。

　　　图12

　　　Xml文件会包含 所对应的图片名字，图形尺寸，类别等其他属性信息。

ImageSets：

ImageSets存放的是每一种类型的challenge对应的图像数据。

在ImageSets下有四个文件夹：

　　　图13

其中Action下存放的是人的动作（例如running、jumping等等，这也是VOC challenge的一部分）

Layout下存放的是具有人体部位的数据（人的head、hand、feet等等，这也是VOC challenge的一部分）

Main下存放的是图像物体识别的数据，总共分为20类。

Segmentation下存放的是可用于分割的数据。

　　　因为与目标识别无关，这里便不再介绍。

　　　Voc2012 voc2007

测试图片：

总结：

目标识别从传统的人为组织特征，到现在的通过大量的数据来让机器学习最佳的特征，准确率也在大幅度的提高，特征的标示性，抽象性比人为的组织的特征更精确。目标识别从滑动窗口的分类，到学习候选框的位置然后进行分类，目标识别的精确度越来越高，效果更好。在和训练数据集相似的图片上，目标识别的效果基本可以达到人类的视觉识别的高度，但是相比于人类的视觉系统，泛化能力太弱了，对光线也太敏感了。 Yolov2算法虽然在实时识别方面，可以满足要求，我们通过简化网络结构，将模型的权重参数降到原来的一半，更能满足实时性的要求。

训练实现：

摘要 大致写的什么

贡献点 点一下

2 具体介绍一下

贡献点 展开

数据集

3 怎么实现 caffe

训练方式 我们的方法

4 实验效果

5 总结

参考文献：

压缩的

yolo一二

rcnn系列

dpm

google net

alxnet

Voc数据集

batch normalization

network in network

Caffe

@misc{pascal-voc-2012,

author = "Everingham, M. and Van~Gool, L. and Williams, C. K. I. and Winn, J. and Zisserman, A.",

title = "The {PASCAL} {V}isual {O}bject {C}lasses {C}hallenge 2012 {(VOC2012)} {R}esults",

howpublished = "http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html"}