SSD(Single Shot MultiBox Detector)算法理解

原创 Main Theme 2018-06-06 17:53:31 8783 收藏 23
分类专栏： 深度学习用于目标检测论文

1、算法概述

SSD(Single Shot MultiBox Detector)是ECCV2016的一篇文章，属于one - stage套路。在保证了精度的同时，又提高了检测速度，相比当时的Yolo和R-CNN是最好的目标检测算法了，可以达到实时检测的要求。在Titan X上，SSD在VOC2007数据集上的mAP值为74.3%，检测速度为59fps。

SSD算法在传统的基础网络（比如VGG）后添加了5个特征图尺寸依次减小的卷积层，对5个特征图的输入分别采用2个不同的3*3的卷积核进行卷积，输出分类用给的confidence，每个default box生成21个类别的confidence；一个输出回归用的localization，每个default box生成4个坐标值，最后将5个特的结果合并（Contact），送入loss层。

多说一句：SSD算法是我平常用的最多的检测算法，但有一个问题是对小目标，尤其是密集小目标的检测效果不好，而且有时检测结果中会出现重叠是对于一般的检测目标，比如车牌、行人和验证码什么的，检测准确率还是很高的。而且其检测速度达59FPS比Faster R-CNN系列高了很多，对检测速度求的任务场景首选SSD算法。

2、SSD算法特色

（1）在基础网络（VGG）后添加了辅助性的层进行多尺度卷积图的预测结果融合；

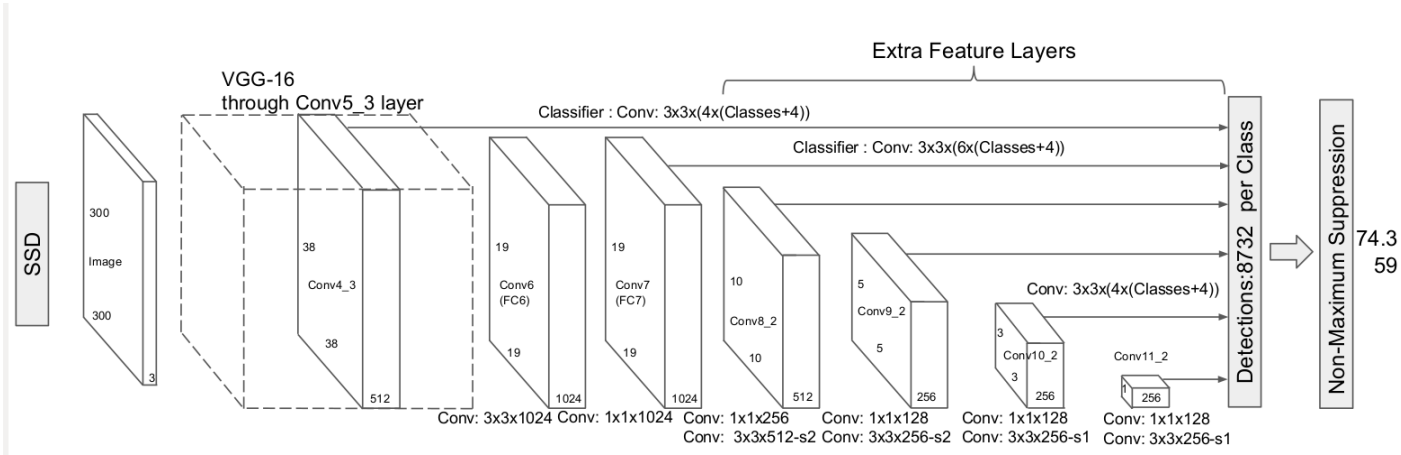
（2）提出了类似Anchor的Default boxes，解决了输入图像目标大小尺寸不同的问题，同时提高了精度，可以理解为一种特征金字塔；

（3）相比于Faster R-CNN，SSD提出了一个彻底的end to end的训练网络，保证了精度的同时大幅度提高了检测速度，且对低分辨率的输入图像的效果

3、具体细节

3.1 添加辅助层结构

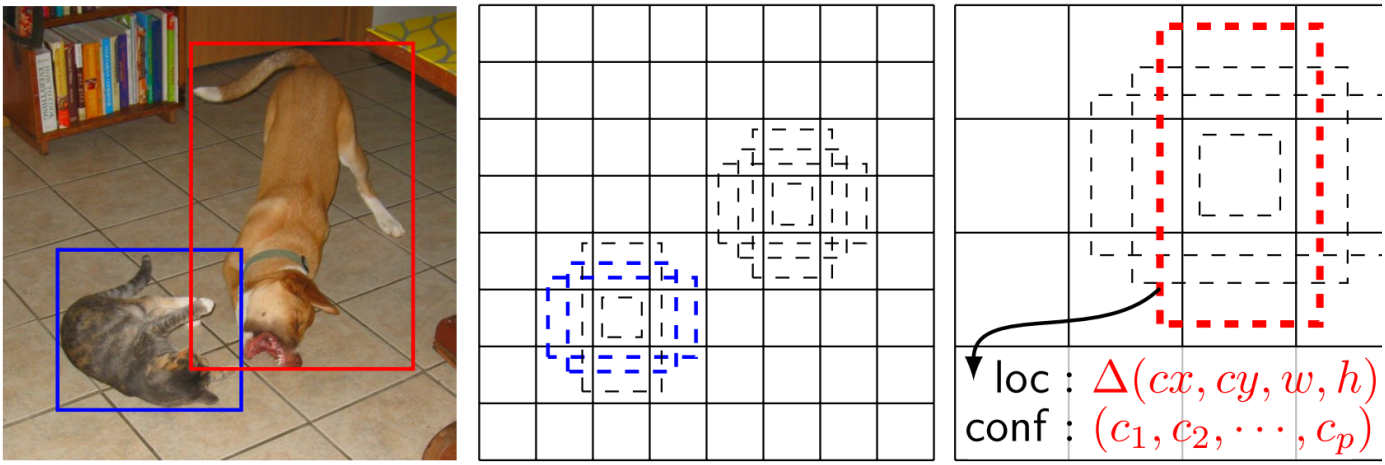
具体结构图如下图所示：



将VGG19的FC6和FC7改成卷积层，又在后面添加了三个尺寸大小逐级减小的卷积层和一个平均池化层。具体用于分类回归的层有：Conv4_3、Co Conv8_2、Conv9_2、Conv10_2和Pool11。最后contact后传给loss层。利用不同层次的特征图来预测offset和confidence，可以检测不同尺寸的物体。

3.2 Default box

是文章的核心部分。这一部分的讲解具体可以看这篇博客：https://www.cnblogs.com/xuanyuyt/p/7447111.html。default box如下图所示：



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

feature map被分成了许多小格子，如4*4、8*8等，每一个格子是feature map的一个cell。每一个feature map的cell上都有一系列固定大小的不同尺寸box，叫default box，上图中虚线的矩形框就是default box。坐标的类别的预测都是基于default box（代码中似乎在default box的基础上进行了处理编程了box）预测的。假设每个feature map的大小是m*n,即feature map的cell为m*n个，每一个default box都要预测C个类别的score和4个offset，假设每个featu对应K个default box，则这张m*n大小的feature map上要产生m*n*K*(4+c)个输出，这也意味着在这张m*n大小的特征图上需要用m*n*k*(c+4)个3*3的卷积积得到最后的m*n*K*(4+c)个输出。当然这些feature map是3.1中提到的参与最终回归预测的5个层。每一个m*n*K*(4+c)个输出都对应一个3*3的卷积核，的5个层的输出全部都执行上述3*3的卷积操作后，将得到的特征图合并（采用类似Inception模块里的Contact，是通道合并而不是卷积图对应的数值相加

default box的尺寸选择（摘自博客：https://blog.csdn.net/u010167269/article/details/52563573）：

所幸的是，SSD 结构中，default boxes 不必要与每一层 layer 的 receptive fields 对应。本文的设计中，feature map 中定的位置，来负责图像中特定的区域，以及物体特定的尺寸。加入我们用 m 个 feature maps 来做 predictions，每一eature map 中 default box 的尺寸大小计算如下：

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

其中，s_{min} 取值 0.2，s_{max} 取值 0.95，意味着最低层的尺度是 0.2，最高层的尺度是 0.95，再用不同 aspect ratio 的 default boxes a_r 来表示：$a_r = \{1, 2, 3, \frac{1}{2}, \frac{1}{3}\}$，则每一个 default boxes 的 width、height 就可以计算出来：

$$w_k^a = s_k\sqrt{a_r}$$
$$h_k^a = s_k/\sqrt{a_r}$$

对于 aspect ratio 为 1 时，本文还增加了一个 default box，这个 box 的 scale 是 $s'k = \sqrt{s_k s_{k+1}}$。所以最终，在每个 feature map loc n 上，有 6 个 default boxes。

每一个 default box 的中心，设置为：$\left(\frac{i+0.5}{|f_k|}, \frac{j+0.5}{|f_k|}\right)$，其中，$|f_k|$ 是第 k 个 feature map 的大小，同时 $i, j \in [0, |f_k|)$。

在结合 feature maps 上，所有 不同尺度、不同 aspect ratios 的 default boxes，它们预测的 predictions 之后。可以见，我们有许多个 predictions，包含了物体的不同尺寸、形状。如下图，狗狗的 ground truth box 与 4×4 feature r 中的红色 box 吻合，所以其余的 boxes 都看作负样本。

3.3 Matching Strategy

这一步是说训练需要的default box如何与GT框匹配的问题。MultiBox中用的是best jaccard overlap来配对，jaccard overlap跟IOU的概念类似，都是上并集。MultiBox中采用jaccard overlap最大值的default box与GT（Ground Truth）配对。SSD中只要jaccard overlap大于0.5的default box都可以看做是本，因此一个GT可以与多个default box配对。当然，小于0.5的default box就看做是负例了。

3.4 Hard Negative Mining

经过上述的Matching Strategy可能产生多个与GT匹配的正样例的和数量更多的负例。负样例的数目远远多于正样例的数目，使正负样例数目不平衡，导致难以收敛。解决方法是：选取负样例的default box，将他们的得分从大大小进行排序，选取的得分最高的前几个负样例的default box，最终使正负样例比1：3。

3.5 损失函数

损失函数由分类和回归两部分组成，具体可参考博客：https://www.cnblogs.com/xuanyuyt/p/7447111.html

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)) \qquad (1)$$

参数解释：

（1）x: 令i表示第i个默认框，j表示第j个真实框，p表示第p个类，xi,j p={0，1}表示第i个prior box与类别p的第j个GT相匹配的jaccard overlap系数，若不匹则系数为0；

（2）c: 类别分类的置信值；

（3）l: 预测框参数，即box的中心坐标位置和box的宽和高；

（4）g：GT框的参数，同上；

（5）N：与阈值大于0.5的GT框相匹配的default box(prior box)的个数；

（6）a（阿尔法）:权重项，在prototxt中设置loc_weight对应权重项，默认为1。实际问题中检查对于你的样本，回归和分类问题哪个更难，调整loc_weig练。Lloc是Faster R-CNN中的Smooth L1 loss，Lconf是Softmax Loss。

3.6 Data Augmentation

对每张训练图像做如下的数据增广：

（1）采用原始图像；

（2）在原图的基础上随机采样一个patch, jaccard overlap的值随机为{0.1, 0.3, 0.5, 0.7, 0.9};

（3）在原图的基础上随机采样一个patch,采样的patch的scale随机在【0.1, 1】中取，aspect ratio随机在【1/2， 2】之间取；

这样一个样本被上面3个batch_sampler采样器采样后会生成多个候选样本，然后从中随机选一个样本送入网络中训练。

测试时由于会产生大量的Bounding boxes，采用NMS（非极大值抑制），阈值设置为0.01。

具体的实验结果请看论文，需要注意的是，论文在最后关于小目标的识别也做了对应的Data Augmentation。

参考博客：

https://blog.csdn.net/u010167269/article/details/52563573

https://www.cnblogs.com/xuanyuyt/p/7447111.html

推荐一篇SSD使用的教学博客：

https://blog.csdn.net/u014696921/article/details/53353896

优质评论可以帮助作者获得更高权重

记忆如阳: 博主你好，我之前也一直在用ssd，但是最近我想试一些最新的算法，所以想请教一下，你知道今年cvpr、eccv、iccv这些检测算法里边，能达到实时性的且精度和召回率最优的文章是哪个吗 3 年前 回复 •••

Main Theme 博主 回复： 不好意思，以后不从事这个方向了，所以最新的文章没有跟进。目前较为经典而且不算过时的算法，FPN和RetinaNet可以研究一下。 年前 回复 •••