前言

五一假期期间，写了几篇SSD源码解读系列文章，阅读完源码之后，SSD的很多细节都弄懂了，经过这几个月的工作和学习，对SSD又有了新的理解，这里跟大家一起分享我对SSD的一些理解，欢迎一起讨论交流。

目录

SSD效果为什么这么好

虽然SSD这个算法出来已经两年了，但是至今依旧是目标检测中应用最广泛的算法，虽然后面有很多基于SSD改进的算法，但至今也没有哪一种检测算法在速度和精度上能够完全碾压SSD的。那么为什么SSD效果这么好？SSD效果好主要有三点原因：

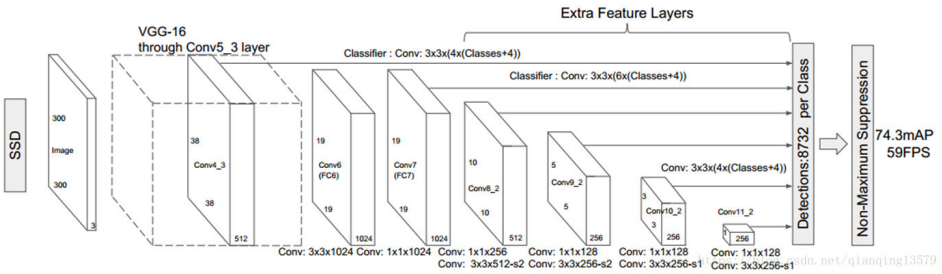
1. 多尺度

2. 设置了多种宽高比的default box(anchor box)

3. 数据增强

下面一一对他们进行分析。

原因1：多尺度



这里写图片描述

由SSD的网络结构可以看出，SSD使用6个不同特征图检测不同尺度的目标。低层预测小目标，高层预测大目标。

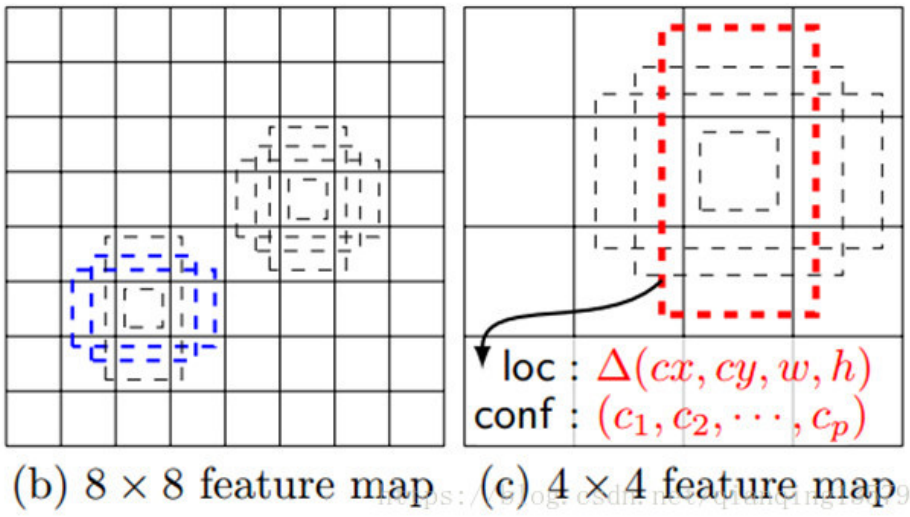
Prediction source layers from:						mAP use boundary boxes?		# Boxes
conv4_3	conv7	conv8_2	conv9_2	conv10_2	conv11_2	Yes	No	
✔	✔	✔	✔	✔	✔	74.3	63.4	8732
✔	✔	✔	✔	✔		**74.6**	63.1	8764
✔	✔	✔	✔			73.8	68.4	8942
✔	✔	✔				70.7	69.2	9864
✔	✔					64.2	64.4	9025
	✔					62.4	64.0	8664

Table 3: **Effects of using multiple output layers.**

这里写图片描述

作者在论文中通过实验验证了，采用多个特征图做检测能够大大提高检测精度，从上面的表格可以看出，采用6个特征图检测的时候，mAP为74.3%,如果只采用conv7做检测，mAP只有62.4%。

原因2：设置了多种宽高比的default box



这里写图片描述

在特征图的每个像素点处，生成不同宽高比的default box(anchor box),论文中设置的宽高比为{1,2,3,1/2,1/3}。假设每个像素点有k个default box，需要对每个default box进行分类和回归，其中用于分类的卷积核个数为c*k(c表示类别数)，回归的卷积核个数为4*k。

SSD300中default box的数量:（38*38*4 + 19*19*6 + 10*10*6 + 5*5*6 + 3*3*4 + 1*1*4）= 8732

讲到这里，我想对于刚学习SSD的朋友，一定有这些疑惑：

1. 为什么要设置default box?

2. 为什么要设置多种宽高比的default box?

3. 为什么在6个特征图上使用3x3的卷积核进行卷积就可以做检测了呢？

这些问题曾经困恼我很长时间，当我看完SSD的源码，并通过很长时间的实践，对这些问题豁然开朗，下面分享自己对这些问题的一些理解，希望对你理解SSD能够提供一些帮助。

理论感受野和有效感受野

这里先简单说明一下理论感受野和有效感受野的概念，更加详细的介绍参考论文:Understanding the Effective Receptive Field in Deep Convolutional Neural Networks[1]。这两个概念的理解对于理解default box(anchor)非常重要。

影响某个神经元输出的输入区域就是理论感受野，也就是我们平时说的感受野，但该输入区域的每个像素点对输出的重要性不同，越靠近中心的像素点影响越大，呈高斯分布，也就是说只有中间的一小部分区域对最后的输出有重要的影响，这个中间的一小部分区域就是有效感受野。

这里写图片描述

图a中，整个黑色区域就是理论感受野(TRF)，中间呈高斯分布的白色点云区域就是有效感受野(ERF)

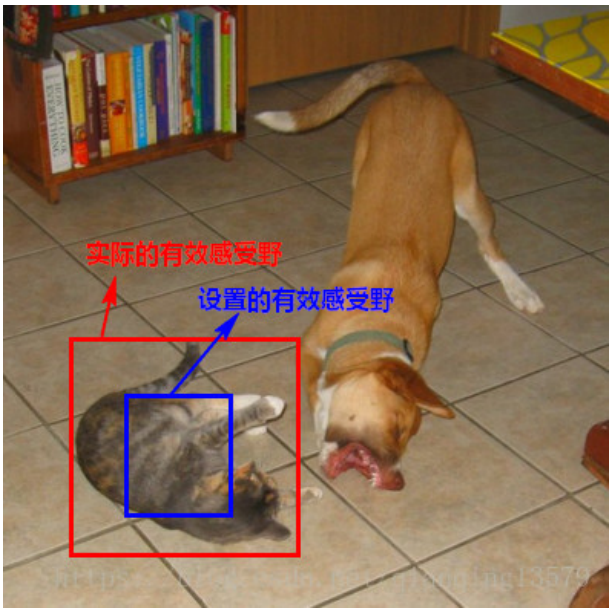
图b中，图中黑色虚线区域对应理论感受野，蓝色虚线部分对应有效感受野，红色实线框是anchor大小,他比理论感受野小很多，但是能够匹配有效感受野。

为什么要设置default box?

SSD在6个特征图上使用2组3x3的卷积核分别做分类和boundingbox回归,所以SSD是一个全卷积神经网络。我们知道每个特征图上每个像素点对应一个理论感受野，所以SSD相当于对原图中所有的理论感受野作分类和回归，由于有效感受野在理论感受野中有重要的影响，其他区域的影响可以忽略，所以这里我们认为SSD是对有效感受野作分类和回归，那么问题来了，既然是对所有的有效感受野做分类和回归，那每个有效感受野的分类的label和回归的label是如何确定的呢?default box就是用来干这个的。

每一层的default box设置了每一层特征图的有效感受野，然后使用这些default box与ground truth进行匹配来确定特征图上每个像素点的实际的有效感受野的label(包含分类label和回归label)，分别用于分类和boundingbox回归。说的简单点，default box就是用来确定特征图上每个像素点实际的有效感受野的label的。

既然default box是确定实际有效感受野的label的，所以如果default box设置的有效感受野能够很好的匹配实际的有效感受野，SSD模型效果就会很好，如果两者相差较大，模型效果就会很差。



这里写图片描述

上图中，某一层特征图的某个像素点对应的**实际有效感受野是红色区域**，这个实际有效感受野的label应该是猫，但是SSD训练时这个红色区域的label是由default box确定的，**如果default box设置的有效感受野对应的是蓝色区域**，通过对default box与ground truth进行匹配我们发现，蓝色区域的label不是猫，而是背景，这样由default box确定的labe与实际有效感受野的真实的label就匹配不上了，如果用这个label作为红色区域的真实label就不对了，训练效果就会很差。

由于default box只要匹配实际的有效感受野就可以了，而实际的有效感受野要比理论感受野小很多，所以SSD中每一层的default box的大小可以比理论感受野小很多。作者在论文中也提到了这一点：

Feature maps from different levels within a network are known to have different (empirical) receptive field sizes .Fortunately, within the SSD framework, the default boxes do not necessary need to correspond to the actual receptive fields of each layer. We design the tiling of default boxes so that specific feature maps learn to be responsive to particular scales of the objects.

大意就是：SSD中default box不必响应实际的感受野，default box只对特定尺度的目标响应。也就是说，SSD的default box只要能够响应有效感受野就可以了。

所以在训练SSD的时候，default box大小的设置非常重要。目前实际的有效感受野的大小还不能精确计算出来，如何让default box设置的有效感受野更好的匹配实际的有效感受野还需要进一步研究。这一点作者在论文中也提到了：

An alternative way of improving SSD is to design a better tiling of default boxes so that its position and scale are better aligned with the receptive field of each position on a feature map. We leave this for future work.

了解了default box的作用后，我们就很容易知道SSD的本质了

SSD对6个特征图上所有的default box进行分类和回归，其实就是对6个特征图对应的实际的有效感受野进行分类和回归，说得更加通俗一点，这些有效感受野其实就是原图中的滑动窗口，所以SSD本质上就是对所有滑动窗口进行分类和回归。这些滑动窗口图像其实就是SSD实际的训练样本。知道SSD的原理后我们发现深度学习的目标检测方法本质与传统的目标检测方法是相同的，都是对滑动窗口的分类。

注：

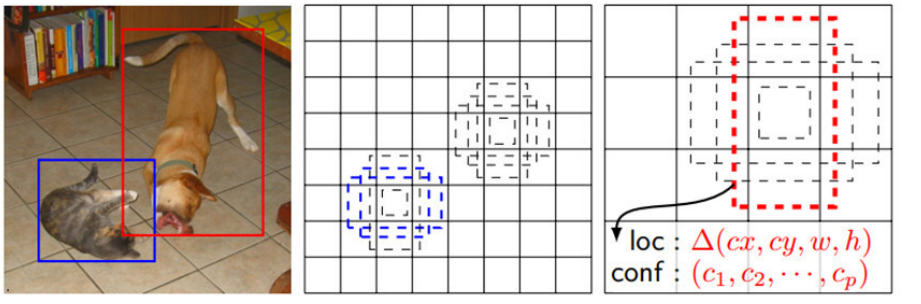
1. 这里要好好理解这两个概念：“每一层实际的有效感受野”和“default box设置的有效感受野”

2. 注意全卷积神经网络与非全卷积神经网络的区别，一般的分类网络比如AlexNet只需要对整幅图像提取特征然后做分类，感受野是整幅图像，所以最后会用全连接层，而SSD中，由于要对每一个感受野做分类，所以只能用卷积层。

default box的匹配

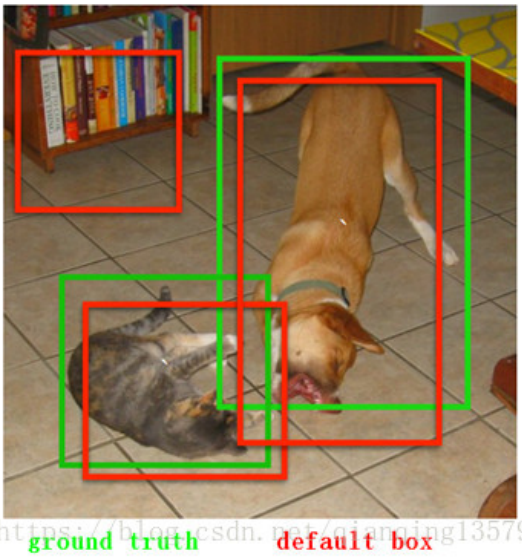
现在我们知道了default box是用来确定label的，那么是如何确定label的呢？

在训练阶段，SSD会先寻找与每个default box的IOU最大的那个ground truth（大于IOU阈值0.5），这个过程叫做匹配。如果一个default box找到了匹配的ground truth,则该default box就是正样本，该default box的类别就是该ground truth的类别，如果没有找到，该default box就是负样本。图1(b)中8x8特征图中的两个蓝色的default box匹配到了猫，该default box的类别为猫，图1(c)中4x4特征图中的一个红色的default box匹配到了狗，该default box的类别为狗。图2显示了实际的匹配过程，两个红色的default box分别匹配到了猫和狗，左上角的default box没有匹配，即为负样本。



这里写图片描述

图1



这里写图片描述

图 2

关于匹配更多的细节，参考Caffe源码multibox_loss_layer.cpp中的FindMatches()函数，前面的博客:SSD源码解读3-MultiBoxLossLayer中也讲到了该函数。

为什么要设置多种宽高比的default box?

我们知道default box其实就是SSD的实际训练样本，如果只设置了宽高比为1的default box,最多只有1个default box匹配到，如果设置更多宽高比的default box，将会有更多的default box匹配到，也就相当于有更多的训练样本参与训练，模型训练效果越好，检测精度越高。

	SSD300			
more data augmentation?		✔	✔	✔	✔
include $\{\frac{1}{2}, 2\}$ box?	✔		✔	✔	✔
include $\{\frac{1}{3}, 3\}$ box?	✔			✔	✔
use atrous?	✔	✔	✔		✔
VOC2007 test mAP	65.5	71.6	73.7	74.2	74.3

这里写图片描述

作者实验结果表明，增加宽高比为1/2,2,1/3,3的default box，mAP从71.6%提高到了74.3%。

如何选择default box的scale和aspect ratio?

假设我们用m个feature maps做预测，那么对于每个featuer map而言其default box的scale是按以下公式计算的。

$$S_k = S_{min} + \frac{S_{max} - S_{min}}{m-1}(k-1)$$

这里S_{min}是0.2，表示最低层的scale是0.2,S_{max}是0.9，表示最高层的scale是0.9。宽高比$\alpha_r = 1, 2, 3, 1/2, 1/3$,因此每个default box的宽$w_k^a = S_k\sqrt{\alpha_r}$，高$h_k^a = S_k/\sqrt{\alpha_r}$,当aspect ratio为1时，作者还增加一种scale的default box：$S_k' = \sqrt{S_k S_k + 1}$,因此，对于每个feature map cell而言，一共有6种default box。

示例：
假设m=6，即使用6个特征图做预测, 则每一层的scale:0.2,0.34,0.48,0.62,0.76,0.9
对于第一层，scale=0.2,对应的6个default box为：

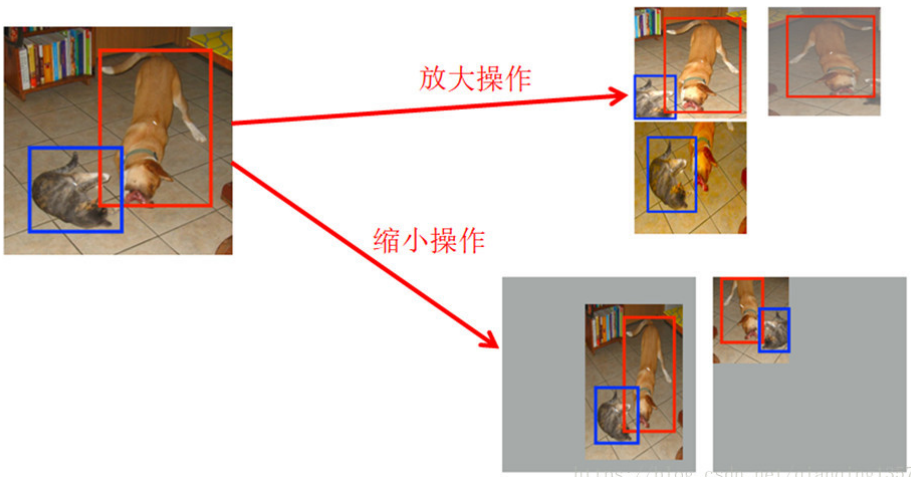
宽高比	宽	高
1	0.200000	0.200000
2	0.282843	0.141421
3	0.346410	0.115470
1/2	0.141421	0.282843
1/3	0.115412	0.346583
最后增加的default box	0.260768	0.260768

注：表格中每个宽高比的default box的实际宽和高需要乘以输入图像的大小，如SSD300,则需要使用上面的数值乘以300得到default box实际大小。

Caffe源码中default box的宽高比以及scale的设置参考prior_box_layer.cpp，前面的博客：SSD源码解读2-PriorBoxLayer也对该层进行过解读。

原因3：数据增强

SSD中使用了两种数据增强的方式
放大操作: 随机crop，patch与任意一个目标的IOU为0.1,0.3,0.5,0.7,0.9，每个patch的大小为原图大小的[0.1,1],宽高比在1/2到2之间。能够生成更多的尺度较大的目标
缩小操作: 首先创建16倍原图大小的画布，然后将原图放置其中，然后随机crop，能够生成更多尺度较小的目标



	SSD300				
more data augmentation?		✔	✔	✔	✔
include $\{\frac{1}{2}, 2\}$ box?	✔		✔	✔	✔
include $\{\frac{1}{3}, 3\}$ box?	✔			✔	✔
use atrous?	✔	✔	✔		✔
VOC2007 test mAP	65.5	71.6	73.7	74.2	74.3

作者实验表明，增加了数据增强后，mAP从65.5提高到了74.3！

数据增强对应Caffe源码annotated_data_layer.cpp，前面的博客:SSD源码解读1-数据层AnnotatedDataLayer也对该层进行过解读。

SSD的缺点及改进

SSD主要缺点：SSD对小目标的检测效果一般，作者认为小目标在高层没有足够的信息。

论文原文：
This is not surprising because those small objects may not even have any information at the very top layers. Increasing the input size (e.g. from 300× 300 to 512× 512) can help improve detecting small objects, but there is still a lot of room to improve.

对SSD的改进可以从下面几个方面考虑：
1. 增大输入尺寸
2. 使用更低的特征图做检测
3. 设置default box的大小，让default box能够更好的匹配实际的有效感受野