上海大学无人艇工程研究院

——环境感知组

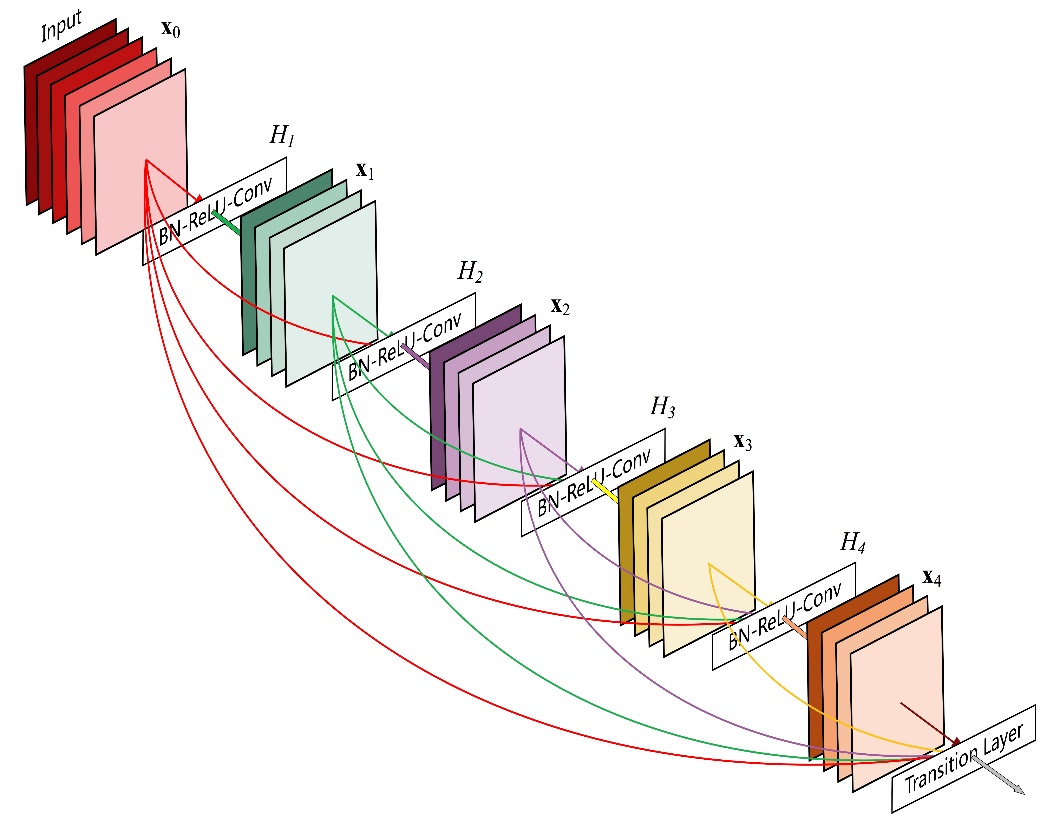
# Learning Deeply Supervised Object Detectors from Scratch

论文：Learning Deeply Supervised Object Detectors from Scratch

作者：Zhiqiang Shen Zhuang Liu Jianguo Li

机构：复旦大学、清华大学

出处：2017年ICCV



注：**加粗**的作者为重点关注研究者

图注：本算法的核心示意图

Date：2017.12.13

## 版本更新记录

|  |  |  |  |
| --- | --- | --- | --- |
| **日 期** | **更新人** | **主要更新内容描述** | **版本号** |
| 2017年12月13日 | 陈加宏 | 完成大致框架搭建 | V1.0.0 |
| 2017年12月02日 | 陈加宏 | 完成算法细节的总结 | V1.0.1 |

目 录

[Multi-kernel Correlation Filter for Visual Tracking—MKCF 1](#_Toc500018927)

[版本更新记录 2](#_Toc500018928)

[1、概述 3](#_Toc500018929)

[1.1 前言 3](#_Toc500018930)

[1.2 创新点 3](#_Toc500018931)

[2、细节 3](#_Toc500018932)

[2.1 主要流程 3](#_Toc500018933)

[2.2 数学模型 3](#_Toc500018934)

[2.3 模型求解 3](#_Toc500018935)

[3、实验 3](#_Toc500018936)

[3.1 代码框架 3](#_Toc500018937)

[3.2 实验结果及分析 3](#_Toc500018938)

[3.3 优缺点总结 3](#_Toc500018939)

[3.4 今后工作 3](#_Toc500018940)

## 1、概述

这是一篇ICCV2017的文章，我觉得非常有意思，因为DSOD算法并不是在mAP上和其他检测算法做比较，看谁的算法更有效或者速度更快，而是从另一个角度切入说明模型微调和直接训练检测模型的差异其实是可以减小的，也就是说训练一个检测模型可以不需要大量的数据和预训练好的模型。这篇文章的核心内容如果用一句话概括那就是：**DSOD可以从0开始训练模型，不需要预训练模型，而且效果可以和模型微调后的模型媲美**。

### 1.1 前言

STOA的目标检测算法的模型训练都是要进行模型预训练的，比如在Imagenet上的分类网络，但是由于目标检测和图像分类任务在目标损失函数、类别分布等方面的不同会造成一些训练偏差，如果数据类型不同比如RGB图像和深度图像会加重这种偏差。采用模型微调的策略可以适当的降低这种影响，但是不能从根本上解决这种偏差。那解决这种偏差的方法可以是从给定的数据中直接训练模型，而不进行预训练过程，之前在这方面的研究失败的原因是网络训练的损失函数较为复杂，同时目标检测的数据集较小。在本文提出的DSOD算法中使用了一些技巧使得目标检测的模型训练可以直接从数据上进行训练，比如dense layer-wise连接结构可以训练出一个性能比较好的检测器。

为什么要从零开始训练一个检测模型，而不是微调一个预训练的模型呢，这是本文的出发点，而作者也在文中介绍了三个主要原因：1、是预训练的模型一般是在分类图像数据集比如Imagenet上训练的，不一定可以迁移到检测模型的数据上比如医疗图像；2、预训练的模型，其结构都是固定的，所以如果你要再修改的话比较麻烦；3、预训练的分类网络的训练目标和检测目标不一致，因此预训练的模型对于检测算法而言不一定是最优的选择。第2点我在实际中有体会过，自己修改一些检测模型，新添的层只能随机初始化，这些层和预训练好的层合并在一起的效果并不一定比不添加的好。

基于上面3个原因，作者就想能不能从零开始训练一个检测模型呢。如果这样可以的话，要怎么设计网络结构使得检测的效果可以跟在预训练模型上训练检测模型的效果媲美呢。于是就有了本文的算法DSOD。

DSOD是在SSD算法的基础上进行改进的，可以简单理解为SSD+DenseNet=DSOD作者文中也曾尝试从零开始训练基于区域候选的检测算法比如Faster RCNN、R-FCN等，发现这样的模型很难收敛；而基于无候选的检测算法比如SSD却可以收敛，虽然最终的效果一般，因此最后采用基于无候选的检测模型SSD。可以看原文的这句话：Our DSOD300 achieves 77.7% mAP, which is much better than the SSD300S that is trained from scratch using VGG16 (69.6%) without deep supervision. It is also much better than the fine-tuned results by SSD300 (75.8%)。在实验中作者是在VOC2007 trainval和2012 trainval数据集上进行训练模型，然后在VOC2007 testset上测试模型的效果。DSOD的参数量是SSD的一半，是Faster RCNN的十分之一。

### 1.2 创新点

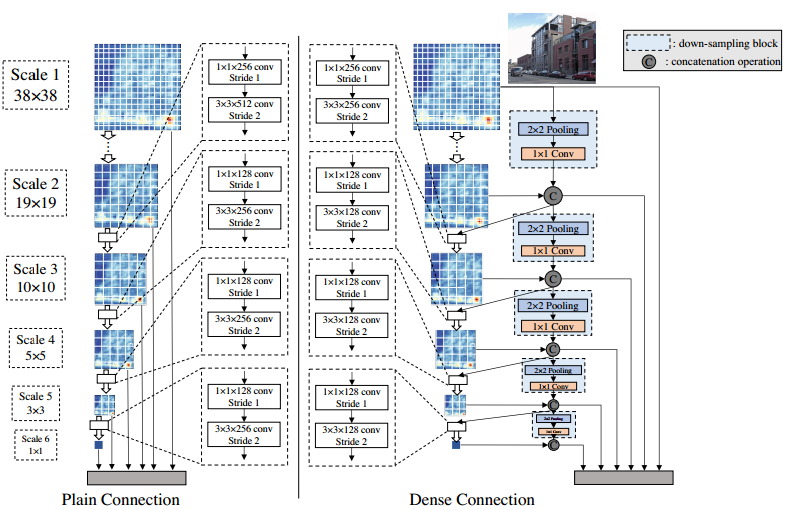
本文贡献：

1、提出DSOD目标检测算法，是第一个可以直接在数据集上进行模型训练而不需要使用图像分类网络预训练的目标检测算法；

2、介绍并验证了在DSOD网络框架中采用的各种技巧可以有效的提升检测效果；

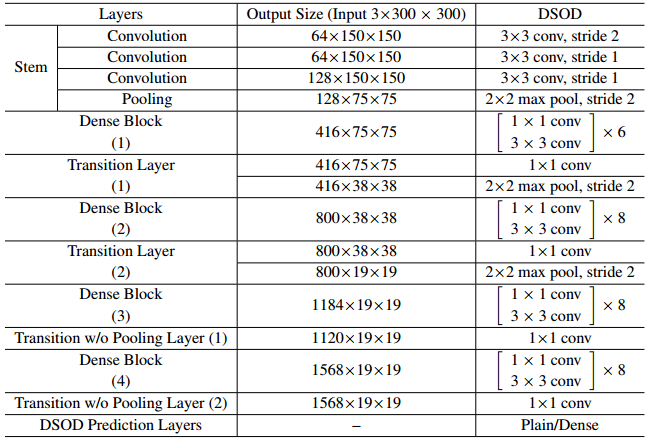
3、在数据集VOC2007、VOC2012、MS COCO上通过检测速度和模型紧凑型方面的对比来验证了本文算法的有效性；

## 2、DSOD网络框架



上图左边的plain connection表示SSD算法中的特征融合操作，这里对于300\*300的输入图像而言，一共融合了6种不同scale的特征。在每个虚线矩形框内都有一个1\*1的卷积和一个3\*3的卷积操作，这其实就是一个bottleneck，也就是1\*1的卷积主要起到降低channel个数从而降低3\*3卷积计算量的作用。

上图右边的dense connection表示本文引入densenet思想的特征融合操作。dense connection部分左边的虚线矩形框部分和plain connection的右边虚线矩形框部分很像，差别在于channel个数dense connection中3\*3的channel个数是对应plain connection中3\*3的channel个数的一半，主要是因为在plain connection中，每个bottleneck的输入直接是前一个bottleneck的输出，但是在dense connection中，每个bottleneck的输入是前面所有bottleneck的输出的concate。dense connection部分右边的矩形框是down sampling block，包含2\*2的max pooling降采样作用和一个1\*1的卷积降低channel个数的作用，作者也提到先进行降采样再进行1\*1卷积主要可以减少计算量。因此可以看出DSOD是SSD+DenseNet的结果。



DSOD的网络结构图如上所示。Dense Block就是DenseNet网络的子模块。这里的stem block部分由3\*3卷积和2\*2的max pool层组成，后面作者的实验有证明这个stem block的存在可以提高mAP。

DSOD 网络分两个部分：用于特征提取的backbone，用于目标预测的front-end。backbone子网络 类似于DenseNets，由stem block, four dense blocks, two transition layers and two transition w/o pooling layers 组成。设计原则：

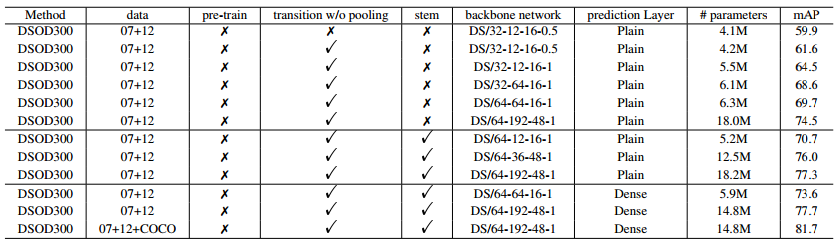
原则1：Proposal-free 基于候选区域提取的方法无法从零开始训练，不收敛。只有 proposal-free 方法可以从零开始训练并收敛。 我们猜测原因是RoI pooling 需要好的初始化输入，于是我们选择了基于SSD来改进；

原则2：Deep Supervision 主要是参考 DenseNet 中 更多的 skip connections 实现 supervised signals 传递 ；Transition w/o Pooling Layer 用这个层来增加 dense blocks 数量。 原来的 DenseNet 的dense blocks 数量是固定的；

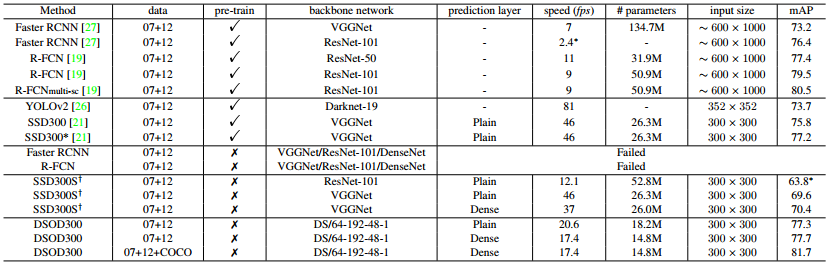
原则3： Stem Block 受 Inception-v3 [33]and v4 [31] 的启发 我们定义 Stem Block 为三个 3×3卷积层和一个 2×2 最大池化层， 我们发现这么设计可以提升性能；

原则4：Dense Prediction Structure 和 SSD类似，我们也是在6个尺度特征图进行目标检测，再融合，但是有SSD 有所区别的是怎么利用特征图的，我们这里是 Learning Half and Reusing Half，对照上图看。

## 3、实验



上表是在VOC2007 testset上不同参数的实验结果。第3，4行的对比可以看出bottlenect的channel个数越多，mAP相对越高。第5、6行的对比可以看出growth rate从16变成48，可以提高4.8%的mAP。第6，9行的对比可以看出stem block的存在可以提高2.8%的mAP。



上表的信息很丰富。前面两个block是在预训练模型上继续训练检测模型，后面3个block是没有预训练模型的，也就是直接开始训练检测模型。粗略地看第1,2,5三个block的对比，DSOD可以在没有预训练模型的情况下取得和其他检测算法在预训练模型上训练的效果。第4个block可以看做baseline，因为作者也对比都没有预训练模型的情况下检测算法的优劣，可以看出第5个block比第4个block的提升非常明显。第5个block的前两个DSOD300的对比主要是预测的结构是plain还是dense的，或者简单讲是否包含down-sampling block和concate操作，可以看出dense结构比plain结构在mAP上有提升，但是速度慢了，因为前者包含down-sampling等操作。

## 3、实验

DSOD是在SSD算法的基础上引入了Densenet的思想，也就是改造了一些层的输入。文章的核心内容可以用原文的一句话来概括：A better model structure might enable similar or better performance compared with complex models trained from large data.另外通过结合densenet，也使得DSOD的参数数量大大减少。如果专注于特殊图像的检测或者苦于没有预训练模型，可以参考这篇文章的思想，而且已经开源了代码。