DSSD : Deconvolutional Single Shot Detector论文笔记

整理：李英杰

paper:
github:

改进

作为SSD的第一个改进分支，本文的主要改进有两点：

- 把SSD的基准网络从VGG换成了Resnet-101，增强了特征提取能力；
- 使用反卷积层（deconvolution layer ）增加了大量上下文信息。最终提升了目标检测精度，尤其是小物体的检测精度。

DSSD以513 * 513的图片输入，在VOC2007上的mAP是81.5%，而SSD为80.6%，在COCO数据集上mAP也达到了33.2%，以下为成绩对比图。

Method	network	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster [24]	VGG	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
ION [1]	VGG	75.6	79.2	83.1	77.6	65.6	54.9	85.4	85.1	87.0	54.4	80.6	73.8	85.3	82.2	82.2	74.4	47.1	75.8	72.7	84.2	80.4
Faster [14]	Residual-101	76.4	79.8	80.7	76.2	68.3	55.9	85.1	85.3	**89.8**	56.7	87.8	69.4	88.3	88.9	80.9	78.4	41.7	78.6	79.8	85.3	72.0
MR-CNN [10]	VGG	78.2	80.3	84.1	78.5	70.8	68.5	88.0	85.9	87.8	60.3	85.2	73.7	87.2	86.5	85.0	76.4	48.5	76.3	75.5	85.0	81.0
R-FCN [3]	Residual-101	80.5	79.9	**87.2**	81.5	72.0	**69.8**	86.8	88.5	89.8	**67.0**	**88.1**	74.5	**89.8**	**90.6**	79.9	81.2	**53.7**	81.8	81.5	85.9	79.9
SSD300*[18]	VGG	77.5	79.5	83.9	76.0	69.6	50.5	87.0	85.7	88.1	60.3	81.5	77.0	86.1	87.5	83.97	79.4	52.3	77.9	79.5	87.6	76.8
SSD 321	Residual-101	77.1	76.3	84.6	79.3	64.6	47.2	85.4	84.0	88.8	60.1	82.6	76.9	86.7	87.2	85.4	79.1	50.8	77.2	**82.6**	**87.3**	76.6
DSSD 321	Residual-101	78.6	81.9	84.9	80.5	68.4	53.9	85.6	86.2	88.9	61.1	83.5	78.7	86.7	88.7	86.7	79.7	51.7	78.0	80.9	87.2	79.4
SSD512*[18]	VGG	79.5	84.8	85.1	81.5	73.0	57.8	87.8	88.3	87.4	63.5	85.4	73.2	86.2	86.7	83.9	82.5	55.6	81.7	79.0	86.6	80.0
SSD 513	Residual-101	80.6	84.3	87.6	**82.6**	71.6	59.0	88.2	88.1	89.3	64.4	85.6	76.2	88.5	88.9	87.5	83.0	53.6	83.9	82.2	87.2	81.3
DSSD 513	Residual-101	**81.5**	**86.6**	86.2	**82.6**	**74.9**	62.5	**89.0**	**88.7**	88.8	65.2	87.0	**78.7**	88.2	89.0	**87.5**	**83.7**	51.1	**86.3**	81.6	85.7	**83.7**

Table 3: **PASCAL VOC2007 `test` detection results.** R-CNN series and R-FCN use input images whose minimum dimension is 600. The two SSD models have exactly the same settings except that they have different input sizes (321 × 321 vs. 513 × 513). It order to fairly compare models, although Faster R-CNN with Residual network [14] and R-FCN [3] provide the number using multiple cropping or ensemble method in testing. We only list the number without these techniques.

DSSD模型

使用Resnet-101替换VGG

- 把VGG换成Resnet-101（下图的1部分）。这里，作者在conv5-x区块后面增加了一些层（SSD Layers），然后会在conv3-x，conv5-x以及SSD Layers预测分类概率和边框偏移。如果仅仅是换网络的话，mAP居然还下降了一个百分点，只有增加上下文信息，精度才会有较大提升。
- 利用中间层的上下文信息（下图的2部分），方法就是把红色层做反卷积操作，使其和上一级蓝色层尺度相同，再把二者融合在一起，得到的新的红色层用来做预测。如此反复，仍然形成多尺度检测框架。在图中越往后的红色层分辨率越高，而且包含的上下文信息越丰富，综合在一起，使得检测精度得以提升。

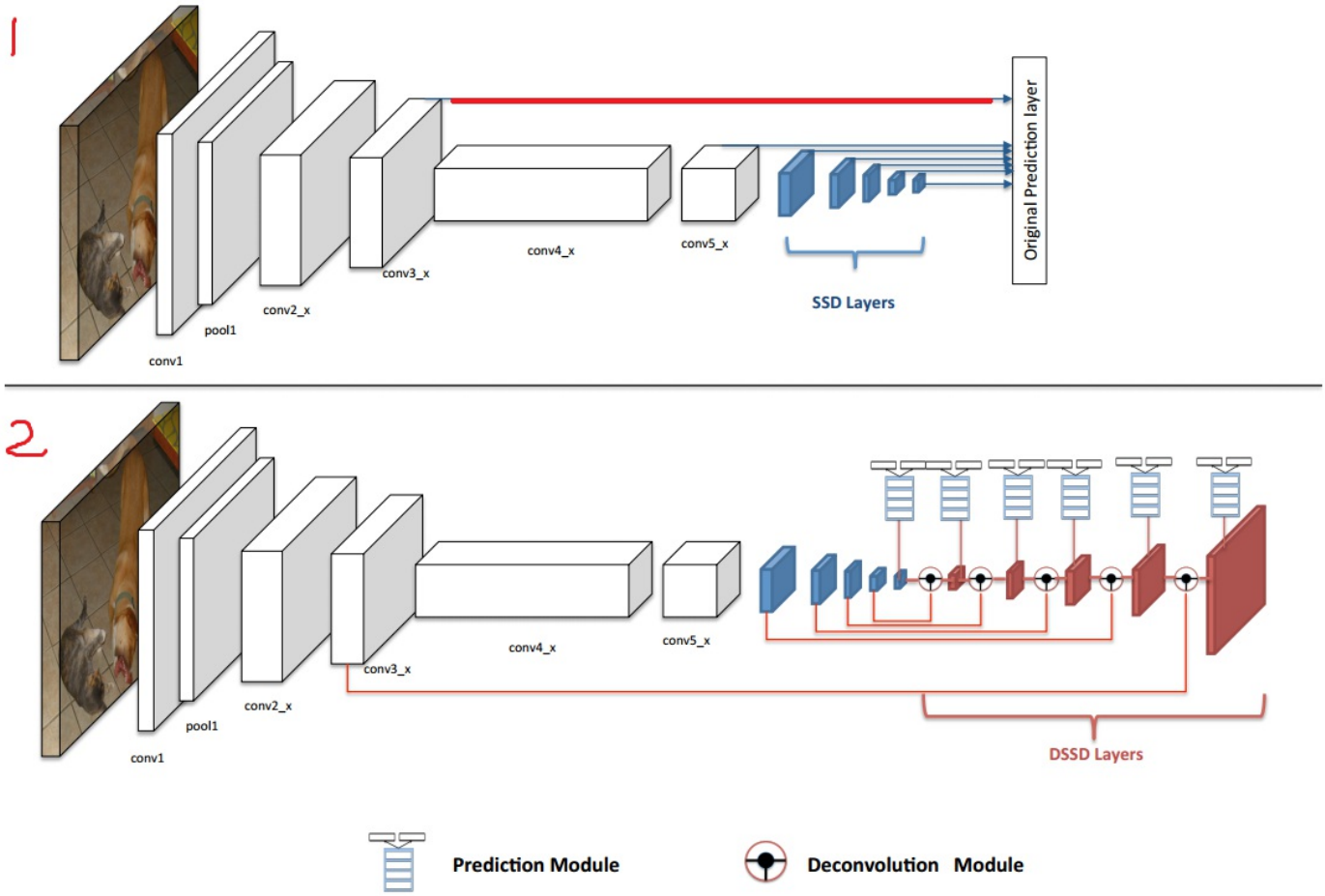

Figure 1: **Networks of SSD and DSSD on residual network.** The blue modules are the layers added in SSD framework, and we call them SSD Layers. In the bottom figure, the red layers are DSSD layers.

预测模块

- **MS-CNN方法指出，改进每个任务的子网可以提高准确性。**根据这一思想，作者在每一个预测层后**增加残差模块，** 并且对于多种方案进行了对比，如下图所示。结果表明，增加残差预测模块后，高分辨率图片的检测精度比原始SSD提升明显。

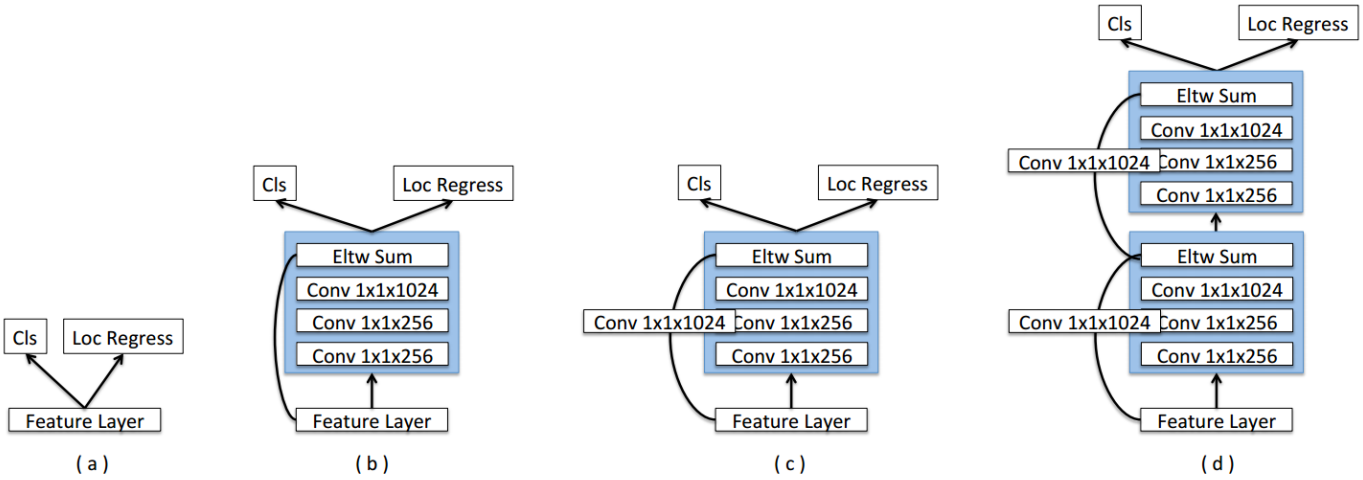

Figure 2: Variants of the prediction module

作者尝试了这么四种prediction module，其中:

- (a)是SSD用的，直接在feature layer上预测
- (b)是设计成residual block的预测模块

- (c)相对比就是把identity mapping换成了1x1卷积
- (d)是stacked (c)

反卷积模块

- 为了引入更多的高级上下文信息，作者在SSD+Resnet-101之上，采用反卷积层来进行预测，和原始SSD是不同的，最终形成Hourglass型网络。添加额外的反卷积层以连续增加后面特征图的分辨率
- 为了加强特征，作者在沙漏形网络中采用了跳步连接（skip connection）方法。

> 通常来说，模型在编码和解码阶段应该包含对称的层，但由于两个原因，作者使解码（反卷积）的层比较浅：
>
> > 检测只算是基础目标，还有很多后续任务，因此必须考虑速度，对称的卷积层影响速度。
> >
> > 目前并没有现成的包含解码（反卷积）的预训练模型。

作者引入了如下图所示的反卷积模块，为了整合浅层特征图和反卷积层的信息。

- 作者受到论文**Learning to Refine Object Segments**的启发，认为用于精细网络的反卷积模块的分解结构达到的精度可以和复杂网络一样，并且更有效率。

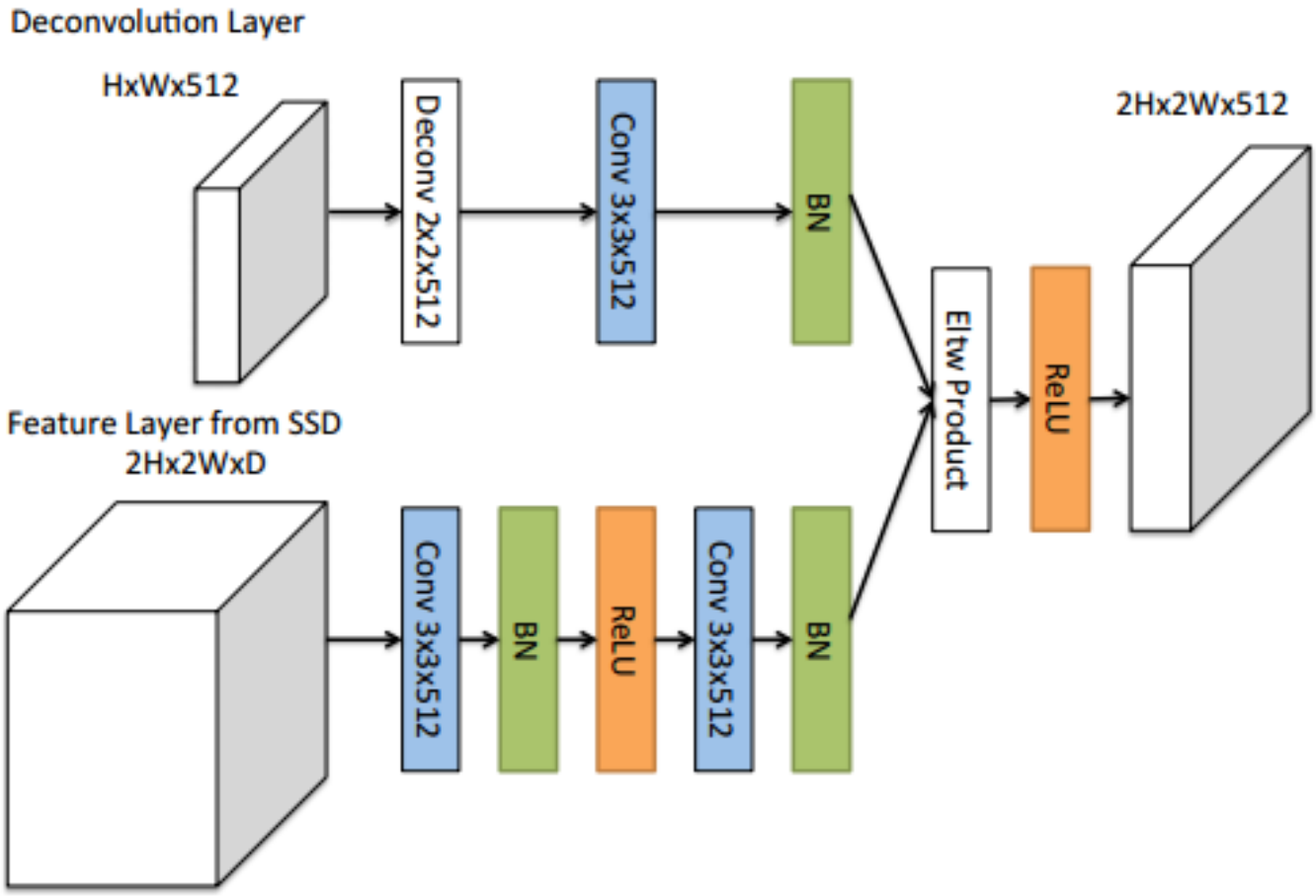

Figure 3: Deconvolution module

> 作者对其进行了一定的修改,如图：
>
> > 在每个卷积层后添加批归一化层；
> >
> > 使用基于学习的反卷积层而不是简单地双线性上采样；
> >
> > 作者测试了不同的结合方式，元素求和（element-wise sum）与元素点积（element-wise product）方式，实验证明点积计算能得到更好的精度。

网络训练

训练技巧大部分和原始SSD类似。
首先，依然采用了SSD的default boxes，把重叠率高于0.5的视为正样本。再设置一些负样本，使得正负样本的比例为1:3。

训练中使Smooth L1+Softmax联合损失函数最小。

训练前依然需要数据扩充（包含了hard example mining技巧）。

另外原始SSD的default boxes维度是人工指定的，可能不够高效，为此，作者在这里采用K-means聚类方法重新得到了7种default boxes维度，得到的这些boxes维度更具代表性