Mimicking Very Efficient Network for Object Detection
http://www.sohu.com/a/160377591_114877
缺点：参数量和速度都可以压缩，但是性能不会比原网络好
虽然没有用到ImageNet去做预训练，但使用预训练过的网络去监督学习，相当于还是使用了ImageNet中的特征信息

摘要

这篇论文，设计了一种全卷积特征模仿框架去训练一个非常高效的基于CNN的检测器，这个检测器不需要ImageNet的预训练而且性能跟大的慢的模型差不多。
在训练中加入大网络的高维特征的监督帮助小网络更好的学习目标表示。
从整个feature map中采样，并使用一个转换层，把小网络的特征映射成跟大网络特征维度相同。
训练小网络：优化从两个网络的相同的域的feature maps中采样的特征的相似度。

1.介绍

深度卷积网络的物体检测方法：Faster R-CNN[28], R-FCN [6] and SSD
Fast R-CNN是从空白开始训练的（没有预训练），AlexNet在Pascal VOC 2007上 AP是40.4%
经过ImageNet预训练的AlexNet 的AP是56.8%
在多GPU上训练一个ImageNet的分类器需要几周的时间
核心思想：如果我们已经有了一个网络，满足了对检测性能的要求，那么我们可以用这个网络监督训练另一个网络用于检测任务。
标准分类任务中用到这种思想的论文是：
J. Ba and R. Caruana. Do deep nets really need to be deep? In Advances in neural information processing systems, pages 2654–2662,2014.
G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.
Faster R-CNN [28], R-FCN [6], SSD [25] and YOLO[27].都是计算features map然后从特征图中解译检测结果。
检测器可以分为特征提取和特征解译
大型网络之间的不同主要在于特征提取
模仿监督应该加载特征提取提的feture map的产生中
真实值的监督加在特征解译中
小网络产生的特征要**转化**成新特征，最小化新特征和大网络产生的特征的欧几里德距离。
真实值的监督训练和Fast R-CNN一样，联合分类和定位的损失

训练过程：
用大网络从训练图像中提取feature map,用feature map和检测的注释联合训练小网络（小网络是空白初始化的）

存在的问题是feature map是高维的，直接模仿，不能收敛。解决方法：优化从域中采样的特征

---

feature map模仿方法的扩展：
1. 尺度上，CNN检测，如果把图像的长宽变成原来的一般，计算量会变为原来的1/4,但性能会下降。 解决：

feature map的一个转换，把feature map 上采样到大尺度，然后让小网络模仿转换过的feature map
2. 把模仿方法扩展成两个过程，从而进一步提高性能

性能下降一点，可以达到4.5倍的速度和16倍的参数的压缩

2.相关工作

RCNN系列把目标检测任务分为两个问题：生成候选域和分类候选域
YOLO和SSD是把目标检测当作一步去做的
RCNN系列和YOLO,SSD都要在计算feature map上花费很多计算量
文章中实现的模仿方法是基于Faster R-CNN ， R-FCN,但可以很容易的扩展到YOLO,SSD和其他基于feature map的方法
学生网络可以从老师网络的中间结果中学习
空白初始化的网络跟在ImageNet上预训练的网络相比，性能相差比较大
这篇论文详细分析了ImageNet的特征：M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? arXiv print arXiv:1608.08614, 2016.
模仿方法可以和其它的加速网络的方法联合使用去进一步加速网络

3.目标检测中的模仿

1.Logits 模仿学习

---

logits:softmax之前的预测 要优化的小网络的L2 logits损失函数是：

$$L(w)=\frac{1}{2T}\sum_t\left\|g(x^t;W)-Z^{(t)}\right\|_2^2$$

训练数据是：

$$(x^{(1)},z^{(1)}),\ldots,(x^{(T)},z^{(T)})$$

$g(x^t;W)$是对第t个训练数据的预测
训练一个小网络去匹配整个大网络的输出是不对的，在目标检测框架中很难通过原始的logits匹配把大网络学到的只是转移给小网络，这种方法模拟的小网络比通过ImageNet预训练和微调的网络差。

2.Feature map 模仿学习

全卷积网络中，整个图片在深度卷积网络中只会被传送一次，然后在feature map上提取候选窗口的特征
主要模仿输出的feature map
在全卷积网络中模仿大网络的feature map激活
神经网络的最后一个卷积层中的特征不仅包含了响应的重点而且包含了空间信息 全卷积网络得到的特征是高维的，很难直接去回归，包含整个图片的响应信息，当图片中物体少，小时，在feature map的这个物体域上的响应非常弱 物体所在的局部域的特征包含更有用的信息
去模仿从候选域中采样的特征去解决高维feature map难以回归问题，让小网络集中学习感兴趣的特征的域

RoI:感兴趣的域
局部区域特征可以通过不同比例和大小的边框在feature map上采样得到，空间金字塔池化方法：SSP
SSP:计算完整的特征图一次，然后池化子窗口特性
使用一个转化层把小网络的feature map采样的特征退回成和大网络相同的维度
SSP层连接到最后一个卷积层后面，可以消除全连接层对固定大小的限制
不同大小的输入-->SPP-->固定大小的输出
小网络的损失函数定义为：

$$\mathcal{L}(W) = \lambda_1 \mathcal{L}_m(W) + \mathcal{L}_{gt}(W), \quad (2)$$

$$\mathcal{L}_m(W) = \frac{1}{2N} \sum_i \|u^{(i)} - r(v^{(i)})\|_2^2, \quad (3)$$

$$\mathcal{L}_{gt}(W) = \mathcal{L}_{cls}(W) + \lambda_2 \mathcal{L}_{reg}(W), \quad (4)$$

- L_m:特征模仿的L2损失，
- L_{gt}：候选域网络的分类和回归损失
- u^i:从大网络的features map中采样的第i个候选框的特征
- v^i:从小网络采样的特征
- r:回归函数，把v的维度转换成跟u一样

在训练阶段L_m可能会非常大，所以两种损失的权重要仔细的选择
正则化L_m:

$$\mathcal{L}_m(W) = \frac{1}{2N} \sum_i \frac{1}{m_i} \|u^{(i)} - r(v^{(i)})\|_2^2, \quad (5)$$

m_i:第i个候选域特征的维度，不同候选域的维度不同
两种损失的权重就可以设置为1了

3.网络架构和实现细节

训练过程可以分为两个阶段：
1. 训练候选域网络，RPN 用于产生候选域 用特征模仿的方法
2. 微调faster-RCNN或者R-FCN在RPN上用训练数据 模仿训练框架:

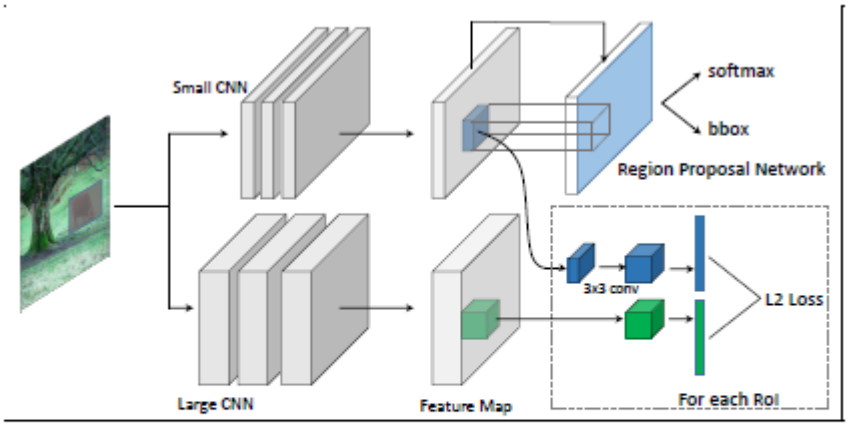

Figure 1: Overall architecture of feature mimic by proposal sampling. A Region Proposal Network generates candidate RoIs, which then used to extract features from the feature maps.

大网络是被很好的初始化的，小网络是随即初始化的

4.两个阶段模仿

在第二个阶段加入logits匹配监督，这样可以把从大网络学到的候选域信息和分类信息都传送给小网络
在损失函数中加入分类和边界框的L2损失。这样就把L_m变为$L(w)$

5.尺度上的模仿

当输入大小变小时提高性能
物体尺度比较小时，目标检测算法表现差
hierarchical feature fusion[3] and hole algorithms用来提高物体检测性能，但是耗费大量时间
小物体检测性能差的原因是在最后通过卷积网络下采样的的feature map中小物体的特征太小
可以简单的在最后的feature map上加一个反卷积层去放大feature map,然后模仿

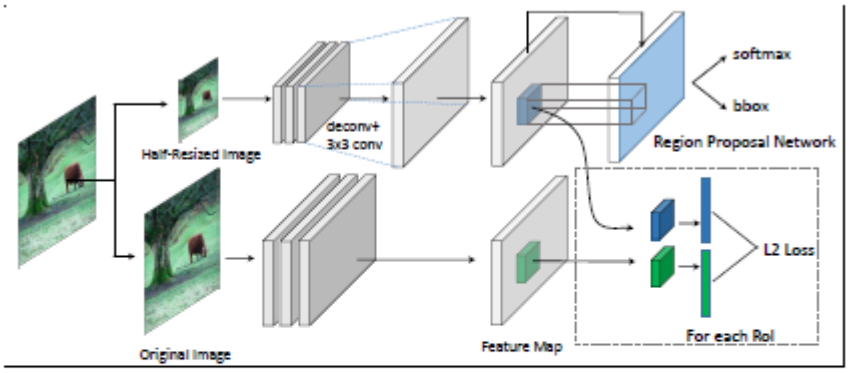

Figure 2: Overall architecture of feature mimic by proposal sampling on different input scales. A Region Proposal Network generates candidate RoIs, which then used to extract features from the up-sampled feature map and the feature map of the network with large input.

大图片通过网络的步长是16，一半大小的图片通过的步长是8， 从而产生一个相似大小的feature map

4.实验

1. Caltech行人检测
   评价方法：FPPI(False Positive Per Image)
   N:样本个数
   L:目标的准确位置
   P：检测到的目标的位置
   如果图像中没有目标，而检测到n个目标，则FP+n
   如果图像有目标，检测到目标，但是P没有击中L，则FP+1

$$FPPI = \frac{FP}{N}$$

   困难：图片的大小是640*480像素，但是行人的高度是低于80像素
   解决方法：调节图像的大小，使得短边为1000像素，在这种尺度下训练RPN和R-FCN.纵横比有两种2：1，3：1
   1/n-Inception network：深度和Inception networkyi一样，但每个层只包含1/n的数量的滤波器

2. PASCAL VOC 2007目标检测
   PASCAL VOC 2007 test set ,VOC 2007训练验证集 16k,VOC 2012 训练验证集 16k. 详细的测试结果在论文的附录里有
   大网络是Inception network architecture 的，在VOC2007数据集上的mAP是75.7%
   评价RPN：每张图片300个候选框的recall ;IoU 设为0.7的正例的recll
   RPN的性能跟原来的大网络的差不多

Method	Recall@.5	Recall@.7
Inception RPN	97.26%	85.36%
1/2-Inception-from-scratch	91.18%	70.66%
1/2-Inception-finetune-ImageNet	96.8%	83%
1/2-Inception-mimic	97.13%	85.58%

Table 11: RPN results on PASCAL VOC 2007 given up to 300 proposals per image. Recall@.7 means the IoU threshold to determine true positives is set to 0.7.

基于Faster RCNN
首先在第一阶段，从大的模型中模仿一个小的RPN，然后在第二阶段微调。
第一阶段训练的RPN只能预测候选框，而不能分类。
在共享的features map上模仿是最好的选择

5.收获

这种方法，可以把达模型的mAP提高2-3个点，但时间和参数数量少很多。缺点就是性能方面相对大网络没有很大的提高，因为训练小网络用了大网络的知识和实际的训练数据，相当于比训练大网络用到的知识多

？？？？改进方法：可以从多个大的网络中学习，这样虽然时间和参数会多一些，但能不能比较好的提高性能。
？？？？统一的一个网络，可以设置参数，决定模仿网络的复杂度，通过调节参数平衡时间和性能，从而可以满足不同的应用 ，参数就是决定：需要的大网络的数量或者类似的
