《Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition》 讲了SSP用于图像分类和检测，着重看检测

1.介绍

流行的ＣＮＮ需要一个固定大小的输入，限制了输入图片的比例和大小
常用的把图片变成固定大小的方法：剪裁和扭曲

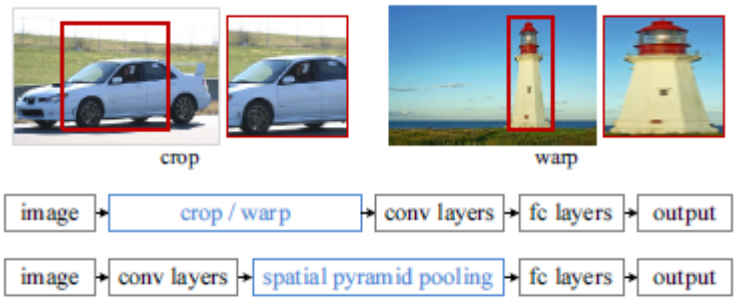

Figure 1: Top: cropping or warping to fit a fixed size. Middle: a conventional CNN. Bottom: our spatial pyramid pooling network structure.

剪裁：不能包含完整的物体
扭曲：会造成几何形变

为什么ＣＮＮ需要固定大小的输入？
CNN由卷积层和全连接层组成
卷积层：以滑动窗口的方式工作，输出的feature map代表了激活的空间分布，不需要固定图片大小，可以输出任意大小的feature map
全连接层：需要固定大小的输入，在网络的深层阶段

SSP加在最后一个conv后，SSP层产生一个固定大小的输出，然后输入到fc层
spatial pyramid matching:SPM来自BoW,SPP来自SPM
SPP对深度ＣＮＮ的好处：
1. 不管输入的大小，产生固定大小的输出，可以和滑动窗口池化同时使用
2. SSP使用多级空间bin，对物体的变形比单窗口大小更鲁棒
3. *（？？？什么叫池化到可变尺度）*由于输入尺度的灵活性，SSP可以把特征精确的池化到可变尺度

SSP-net训练和测试时都可以使用可变尺度的图片或窗口
用可变大小的图片训练可以增加尺度不变性，减小over-fiting
尺度不变性就是对于纵横比改变，大小改变，旋转变换依然能够检测出来
对于单个网络接受可变大小的输入，通过多个共享参数的网络去近似，这些网络使用固定大小的输入训练的
多尺度的收敛速度跟单尺度的差不多，但是精确率更高
SSP跟特殊的CNN设计是正交的
在RCNN中使用SSP,可以对每张图片只计算一次feature map，加快速度100倍
RCNN中使用SSP可以达到0.5秒处理一张图片
SSP-net可以促进一些更深更大的网络，相对于no-SSP
代码：http://research.microsoft.com/en-us/um/people/kahe/

2.有SSP的深度网络

卷基层和feature maps

流行的卷积网络有 7 层， 5 层卷积层， 2 层全连接层，输出到一个N-way的softMax分类器。
是因为全连接层需要输入一个固定长度的向量，所以才要求输入图片的大小是固定的。
卷基层是可以接受任意大小的输入的，使用滑动的滤波器，输出具有大概相同的纵横比,输出叫做feature maps，feature maps包含响应的强度和空间位置

空间金字塔池化层

卷基层接受任意大小的输入，产生可变大小的输出
分类器（SVM/softmax）和全连接层需要固定长度的向量作为输入
空间金字塔池化的优点是可以在空间bins上保持空间信息
空间bins的大小跟图片大小成比例
bins的数量是固定的，跟图片大小无关
最后一个卷积层后面的池化层用SSP替换
网络结构:

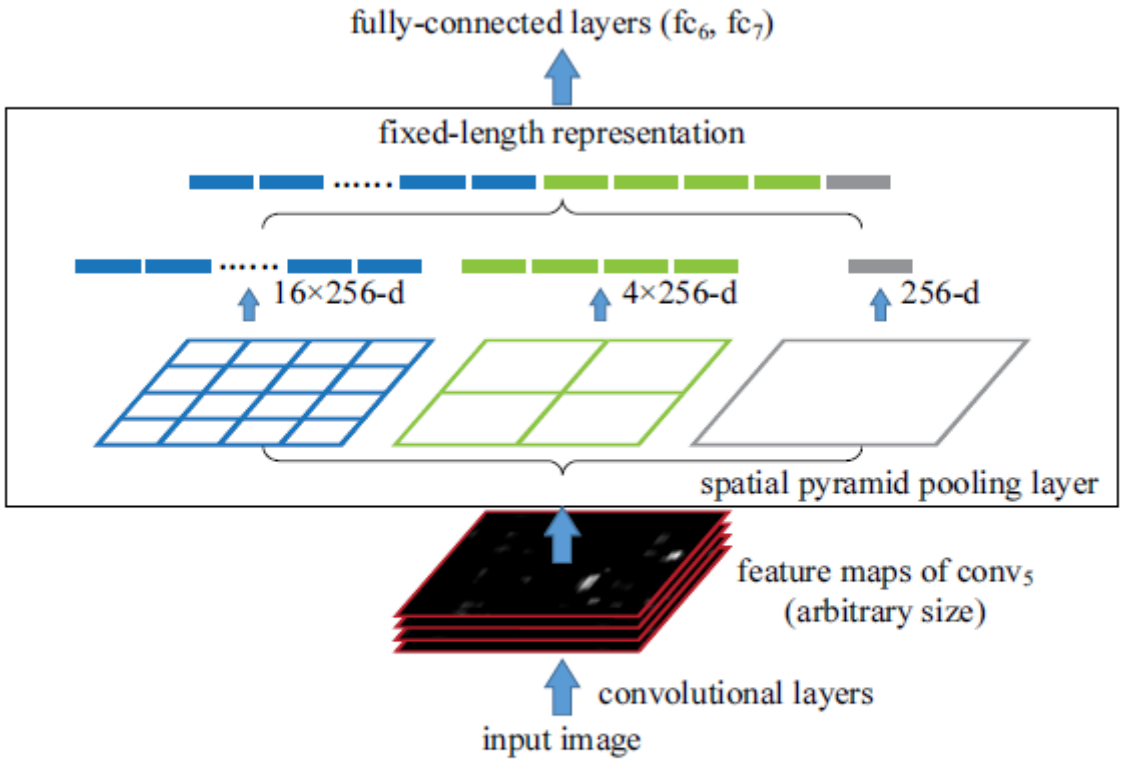

Figure 3: A network structure with a **spatial pyramid pooling layer**. Here 256 is the filter number of the $conv_5$ layer, and $conv_5$ is the last convolutional layer.

空间池化层的输出是kM维的向量，k是滤波器的数量，M是bins的数量
*（？？？是通过这样来达到任意固定长度的输出吗？比如，全连接层的输入需要\(5376=21256=16*256+4256+256\)这样的话就只能是256的整数倍 ）*
输入图片是不同尺度时，网路也会以不同尺度来提取特征
粗糙的金字塔层只有一个bin,包含整个图片，相当于全局池化

训练网络

？？不能用反向传播来训练吗？

单尺度训练

从图片中裁剪一个224×224大小的图片作为输入
裁剪是为了扩充数据*（？？？裁剪能扩充数据吗？）*
conv5输出大大小为$a \times a$,金字塔层有$n \times n$个bins,则窗口大小为$win = \lceil a/n \rceil$步长为$str = \lfloor a/n \rfloor$

多尺度训练

输入是一组预先定义好大小的图片，$180 \times 180, 224 \times 224$
把224×224 大小的域**调整**(不是剪裁)成180×180
180-network和224-network有相同长度的输出，每一层的参数也是一样的
实现接收两个不同大小输入的网络是通过两个接受固定大小的网络共享参数来实现的
为了节省总的开销，是先训练一个网络，然后把另一转换到这个网络
训练分为单尺度和多尺度，测试是在任意尺度的

3.SSP-Net用于图像分类

用于分类就大概看一下，主要关注用于物体检测

在ImageNet2012数据集上的分类实验

水平翻转和颜色变化扩展数据集
两个全连接层用到了Dropout 只是把最后一个卷积层后面的池化层换成SSP层
使用的是4级金字塔，$6 \times 6, 3 \times 3, 2 \times 2, 1 \times 1$
一共50个bins

4.SSP-net用于目标检测

R-CNN:在特征提取上有时间瓶颈限制
提取整张图片的feature map一次，然后为每个feature map的候选窗口使用空间金字塔池化等到一个固定长度的特征
由于只计算了一次feature map，所以大大的加快了速度
SSP-net是从feature map的提取候选窗口的特征，而R-CNN是直接从图像的的域中提取

关于selective serch的论文：“Segmentation as selective search for object recognition,” in ICCV, 2011.
http://blog.csdn.net/mao_kun/article/details/50576003 SSP-net可以提取大小的窗口的特征

检测算法

使用fast selective search 方法为每个图片收集2000个候选框
调整图片的尺寸min(w，h) = s，从整张图片中提取feature map
暂时使用ZF-5（单尺度训练的）的SSP-net模型
对于每个候选窗口使用4级金字塔$(1\times1, 2\times2, 3\times3, 6\times6)$ 共50个bin 去池化特征，为每个窗口产生一个12,800$(50*256)$ 维的特征 ,把这些特征输入到全连接层，为每个类训练一个2分类SVM分类器
SVM训练的实现：真实的窗口作为正样本，跟正样本的IoU小于30%的窗口作为负样本，并加入hard negativ mining,对于负样本，如果跟其他窗口的重叠率超过70%则移除。
为20个类训练分类器只要不到一个小时，
用分类器给窗口打分，然后在这些打过分的窗口上使用非极大值抑制
通过多尺度特征提取可以进一步提升，$min(w,h)=s\in S=\{480,576,688,864,1200\}$
为每个尺度，在conv5计算feature map 为每个候选框选择一个单一的尺度s使得每个候选框的像素数接近224×224,只用从这个尺度的feature map提取的特征
微调预训练的网络：只微调全连接层，
conv5后面接了一个空间金字塔池化层，后面有fc6,fc7两个全连接层，fc8是一个21类的分类器
fc8用方差为0.01的高斯分布初始化
学习率为0.0001,调整这三个全连接层的学习率为0.00001
正样本是跟真实的窗口IoU在【0.5,1】,负样本是在【0.1,0.5）
每个min-batch中有２５％是正样本
250k min-batch的学习率是0.0001,50k min-batch的学习率是0.00001
使用bounding-box 回归，特征是来自conv5的，窗口和真实窗口的IoU至少是50%

检测结果

Pascal VOC 2007数据集上

	SPP (1-sc) (ZF-5)	SPP (5-sc) (ZF-5)	R-CNN (Alex-5)
$pool_5$	43.0	44.9	44.2
fc_6	42.5	44.8	46.2
$ftfc_6$	52.3	53.7	53.1
$ftfc_7$	54.5	55.2	54.2
$ftfc_7$ bb	58.0	**59.2**	58.5
conv time (GPU)	0.053s	0.293s	8.96s
fc time (GPU)	0.089s	0.089s	0.07s
total time (GPU)	0.142s	0.382s	9.03s
speedup (*vs.* RCNN)	**64**×	**24**×	-

Table 9: Detection results (mAP) on Pascal VOC 2007. “ft” and “bb” denote fine-tuning and bounding box regression.

五尺度的SSP-net比R-CNN好0.7%，快24倍
使用ZF-5预训练的R-CNN和SPP-net的性能一样，但SSP-net更快,性能一样的原因是ZF-5的架构比AlexNet的好

复杂性和运行时间

用ZF-5预训练的R-CNN,SPP-net在为每张图片计算feature map所用的时间，1尺度的SSP加快了270倍，5尺度的加快了４９倍

regression.

	SPP (1-sc) (ZF-5)	SPP (5-sc) (ZF-5)	R-CNN (ZF-5)
ftfc_7	54.5	55.2	55.1
ftfc_7 bb	58.0	**59.2**	**59.2**
conv time (GPU)	0.053s	0.293s	14.37s
fc time (GPU)	0.089s	0.089s	0.089s
total time (GPU)	0.142s	0.382s	14.46s
speedup (*vs.* RCNN)	**102**×	**38**×	-

Table 10: Detection results (mAP) on Pascal VOC 2007, using the same pre-trained model of SPP (ZF-5).

更快的候选框产生方法：EdgeBoxes(更快0.2s每张图片，Selective search１到２ｓ)
训练阶段使用：Selective search 和EdgeBoxes
测试阶段使用：EdgeBoxes

组合模型用于检测

只是初始化不同的两个模型
首先用每个模型为测试图片的候选框打分，然后在这两个打过分的集合中使用非极大值抑制
组合模型提高了mAP,原因是卷积层的互补性

ILSVRC 2014检测

200类，训练集450k,验证集20k,测试集40k
困难：
1. DET的训练数据是CLS的1/3
2. 2. DET的类别是CLS的1/5,使用提供的有标签的子集做预训练，在DET训练集上训练一个499类的分类网络
3. CLS上物体的尺度是图片长度的0.8,DET上是0.5,调整图片大小$min(w,h)=400$,随机的裁剪一个224×224的视图用于训练，只有跟真实物体的重叠度大于50%才裁剪

就算数据量一样，用更多的类预训练网络更好，原因是提高了特征的质量
训练一个　499-category Overfeat-7 SPP-net
验证集产生正负样本，用selective search产生候选框，训练集只用真实样本产生正样本
用训练和验证集的样本微调fc和训练SVM,用验证集训练bounding-box 回归
6个相似的模型结合起来mAP达到35.11%，排名第二，速度是８GPU小时