R-CNN 论文详解

&创新点

1.采用 CNN 网络提取图像特征，从经验驱动的人造特征范式 HOG、SIFT 到数据驱动的表示学习范式，提高特征对样本的表示能力；

2.采用大样本下有监督预训练+小样本微调的方式解决小样本难以训练甚至过拟合等问题。

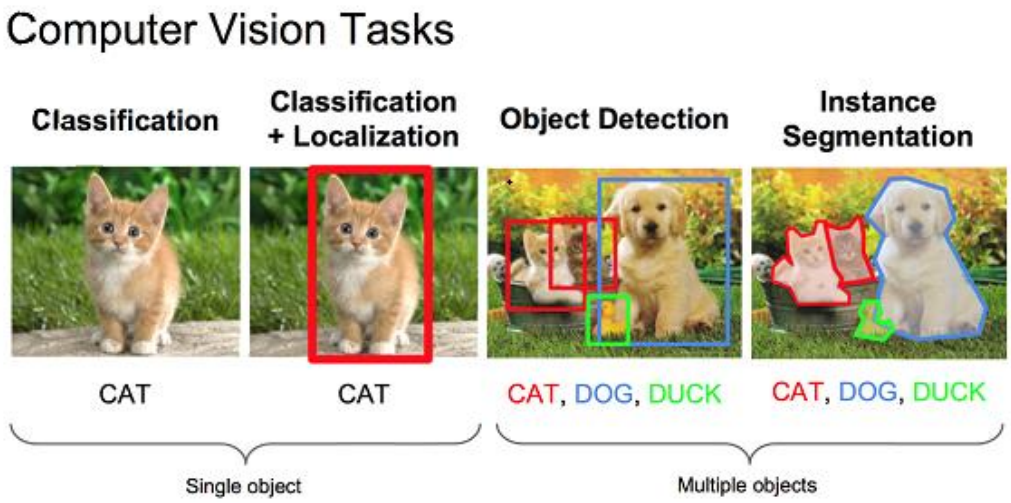
&问题是什么

1.近 10 年以来，以人工经验特征为主导的物体检测任务 mAP【物体类别和位置的平均精度】提升缓慢；

2.随着 ReLu 激励函数、dropout 正则化手段和大规模图像样本集 ILSVRC 的出现，在 2012 年 ImageNet 大规模视觉识别挑战赛中，Hinton 及他的学生采用 CNN 特征获得了最高的图像识别精确度；

3.上述比赛后，引发了一股“是否可以采用 CNN 特征来提高当前一直停滞不前的物体检测准确率“的热潮。

【写给小白：一图理解图像分类，图像定位，目标检测和实例分割】



&如何解决问题

。测试过程

1.输入一张多目标图像，采用 selective search 算法提取约 2000 个建议框；

2.先在每个建议框周围加上 16 个像素值为建议框像素平均值的边框，再直接变形为 227×227 的大小；

3.先将所有建议框像素减去该建议框像素平均值后【预处理操作】，再依次将每个 227×227 的建议框输入 AlexNet CNN 网络获取 4096 维的特征【比以前的人工经验特征低两个数量级】，2000 个建议框的 CNN 特征组合成 2000×4096 维矩阵；

4.将 2000×4096 维特征与 20 个 SVM 组成的权值矩阵 4096×20 相乘【20 种分类，SVM 是二分类器，则有 20 个 SVM】，获得 2000×20 维矩阵表示每个建议框是某个物体类别的得分；

5.分别对上述 2000×20 维矩阵中每一列即每一类进行非极大值抑制剔除重叠建议框，得到该列即该类中得分最高的一些建议框；

6.分别用 20 个回归器对上述 20 个类别中剩余的建议框进行回归操作，最终得到每个类别的修正后的得分最高的 bounding box。

。解释分析

1.selective search

采取过分割手段，将图像分割成小区域，再通过颜色直方图，梯度直方图相近等规则进行合并，最后生成约 2000 个建议框的操作，具体见博客。

2.为什么要将建议框变形为 227×227？怎么做？

本文采用 AlexNet CNN 网络进行 CNN 特征提取，为了适应 AlexNet 网络的输入图像大小：227×227，故将所有建议框变形为 227×227。
那么问题来了，如何进行变形操作呢？作者在补充材料中给出了四种变形方式：

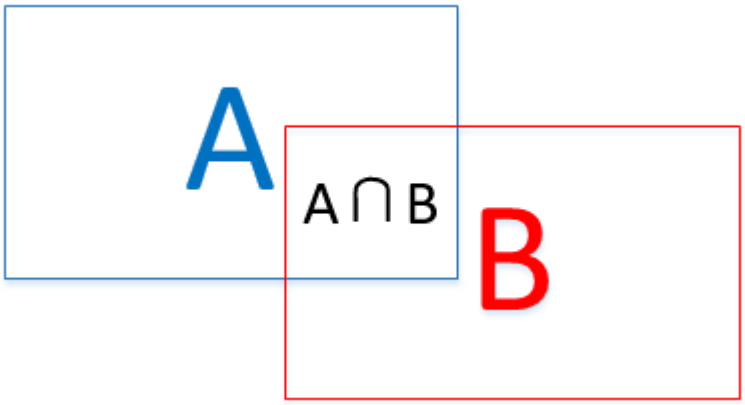
① 考虑 context【图像中 context 指 RoI 周边像素】的各向同性变形，建议框像周围像素扩充到 227×227，若遇到图像边界则用建议框像素均值填充，下图第二列；
② 不考虑 context 的各向同性变形，直接用建议框像素均值填充至 227×227，下图第三列；
③ 各向异性变形，简单粗暴对图像就行缩放至 227×227，下图第四列；
④ 变形前先进行边界像素填充【padding】处理，即向外扩展建议框边界，以上三种方法中分别采用 padding=0 下图第一行，padding=16 下图第二行进行处理；

经过作者一系列实验表明采用 padding=16 的各向异性变形即下图第二行第三列效果最好，能使 mAP 提升 3-5%。

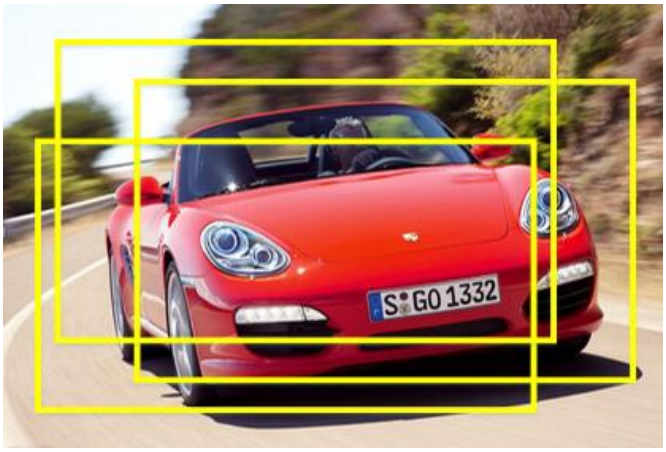
3.CNN 特征如何可视化？

文中采用了巧妙的方式将 AlexNet CNN 网络中 Pool5 层特征进行了可视化。该层的 size 是 6×6×256，即有 256 种表示不同的特征，这相当于原始 227×227 图片中有 256 种 195×195 的感受视野【相当于对 227×227 的输入图像，卷积核大小为 195×195，padding=4，step=8，输出大小(227-195+2×4)/8+1=6×6】;
文中将这些特征视为”物体检测器”，输入 10million 的 Region Proposal 集合，计算每种 6×6 特征即“物体检测器”的激活量，之后进行非极大值抑制【下面解释】，最后展示出每种 6×6 特征即“物体检测器”前几个得分最高的 Region Proposal，从而给出了这种 6×6 的特征图表示了什么纹理、结构，很有意思。

4.为什么要进行非极大值抑制？非极大值抑制又如何操作？
先解释什么叫 IoU。如下图所示 IoU 即表示(A∩B)/(A∪B)



在测试过程完成到第 4 步之后，获得 2000×20 维矩阵表示每个建议框是某个物体类别的得分情况，此时会遇到下图所示情况，同一个车辆目标会被多个建议框包围，这时需要非极大值抑制操作去除得分较低的候选框以减少重叠框。



具体怎么做呢？

① 对 2000×20 维矩阵中每列按从大到小进行排序；

② 从每列最大的得分建议框开始，分别与该列后面的得分建议框进行 IoU 计算，若 IoU>阈值，则剔除得分较小的建议框，否则认为图像中存在多个同一类物体；

③ 从每列次大的得分建议框开始，重复步骤②；

④ 重复步骤③直到遍历完该列所有建议框；

⑤ 遍历完 2000×20 维矩阵所有列，即所有物体种类都做一遍非极大值抑制；

⑥ 最后剔除各个类别中剩余建议框得分少于该类别阈值的建议框。【文中没有讲，博主觉得有必要做】

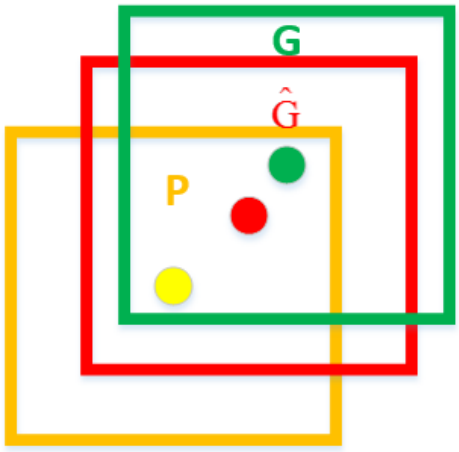
5.为什么要采用回归器？回归器是什么有什么用？如何进行操作？

首先要明确目标检测不仅是要对目标进行识别，还要完成定位任务，所以最终获得的 bounding-box 也决定了目标检测的精度。

这里先解释一下什么叫定位精度：定位精度可以用算法得出的物体检测框与实际标注的物体边界框的 IoU 值来近似表示。

如下图所示，绿色框为实际标准的卡宴车辆框，即 Ground Truth；黄色框为 selective search 算法得出的建议框，即 Region Proposal。即使黄色框中物体被分类器识别为卡宴车辆，但是由于绿色框和黄色框 IoU 值并不大，所以最后的目标检测精度并不高。采用回归器是为了对建议框进行校正，使得校正后的 Region Proposal 与 selective search 更接近， 以提高最终的检测精度。论文中采用 bounding-box 回归使 mAP 提高了 3~4%。

那么问题来了，回归器如何设计呢？



如上图，黄色框口P表示建议框Region Proposal，绿色窗口G表示实际框Ground Truth，红色窗口$\hat{G}$表示Region Proposal进行回归后的预测窗口，现在的目标是找到P到$\hat{G}$的线性变换【当Region Proposal与Ground Truth的IoU>0.6时可以认为是线性变换】，使得$\hat{G}$与G越相近，这就相当于一个简单的可以用最小二乘法解决的线性回归问题，具体往下看。

让我们先来定义P窗口的数学表达式：$P^i = (P_x^i, P_y^i, P_w^i, P_h^i)$，其中$(P_x^i, P_y^i)$表示第一个i窗口的中心点坐标，$P_w^i$，$P_h^i$分别为第i个窗口的宽和高；G窗口的数学表达式为：$G^i = (G_x^i, G_y^i, G_w^i, G_h^i)$；$\hat{G}$窗口的数学表达式为：$\hat{G}^i = (\hat{G}_x^i, \hat{G}_y^i, \hat{G}_w^i, \hat{G}_h^i)$。以下省去i上标。

这里定义了四种变换函数，$d_x(P)$，$d_y(P)$，$d_w(P)$，$d_h(P)$。$d_x(P)$和$d_y(P)$通过平移对x和y进行变化，$d_w(P)$和$d_h(P)$通过缩放对w和h进行变化，即下面四个式子所示：

$$\hat{G}_x = P_w d_x(P) + P_x \tag{1}$$

$$\hat{G}_y = P_h d_y(P) + P_y \tag{2}$$

$$\hat{G}_w = P_w exp(d_w(P)) \tag{3}$$

$$\hat{G}_h = P_h exp(d_h(P)) \tag{4}$$

每一个$d_*(P)$【*表示x，y，w，h】都是一个AlexNet CNN网络Pool5层特征$\phi_5(P)$的线性函数，即$d_*(P) = w_*^T \phi_5(P)$，这里w_*^T就是所需要学习的回归参数。损失函数即为：

$$Loss = argmin \sum_{i=0}^{N} (t_*^i - \hat{w}_*^T \phi_5(P^i))^2 + \lambda ||\hat{w}_*||^2 \tag{5}$$

损失函数中加入正则项$\lambda ||\hat{w}_*||^2$是为了避免归回参数w_*^T过大。其中，回归目标t_*由训练输入对(P, G)按下式计算得来：

$$t_x = (G_x - P_x)/P_w \tag{6}$$

$$t_y = (G_y - P_y)/P_h \tag{7}$$

$$t_w = log(G_w/P_w) \tag{8}$$

$$t_h = log(G_h/P_h) \tag{9}$$

①构造样本对。为了提高每类样本框回归的有效性，对每类样本都仅仅采集与Ground Truth相交IoU最大的Region Proposal，并且IoU>0.6的Region Proposal作为样本对(P^i, G^i)，一共产生20对样本对【20个类别】；
②每种类型的回归器单独训练，输入该类型样本对N个：$\{(P^i, G^i)\}_{i=1\cdots N}$以及$P^i_{i=1\cdots N}$所对应的AlexNet CNN网络Pool5层特征$\phi_5(P^i)_{i=1\cdots N}$；
③利用(6)-(9)式和输入样本对$\{(P^i, G^i)\}_{i=1\cdots N}$计算$t^i_{*i=1\cdots N}$；
④利用$\phi_5(P^i)_{i=1\cdots N}$和$t^i_{*i=1\cdots N}$，根据损失函数(5)进行回归，得到使损失函数最小的参数w^T_*。

。训练过程

1.有监督预训练

样本	来源
正样本	ILSVRC2012
负样本	ILSVRC2012

ILSVRC 样本集上仅有图像类别标签，没有图像物体位置标注；
采用 AlexNet CNN 网络进行有监督预训练，学习率=0.01；
该网络输入为 227×227 的 ILSVRC 训练集图像，输出最后一层为 4096 维特征->1000 类的映射，训练的是网络参数。

2.特定样本下的微调

样本	来源
正样本	Ground Truth+与Ground Truth相交IoU>0.5的建议框【由于Ground Truth太少了】
负样本	与Ground Truth相交IoU≤0.5的建议框

PASCAL VOC 2007 样本集上既有图像中物体类别标签，也有图像中物体位置标签；
采用训练好的 AlexNet CNN 网络进行 PASCAL VOC 2007 样本集下的微调，学习率=0.001【0.01/10 为了在学习新东西时不至于忘记之前的记忆】；
mini-batch 为 32 个正样本和 96 个负样本【由于正样本太少】；
该网络输入为建议框【由 selective search 而来】变形后的 227×227 的图像，修改了原来的 1000 为类别输出，改为 21 维【20 类+背景】输出，训练的是网络参数。

3.SVM 训练

样本	来源
正样本	Ground Truth
负样本	与Ground Truth相交IoU＜0.3的建议框

由于 SVM 是二分类器，需要为每个类别训练单独的 SVM；
SVM 训练时输入正负样本在 AlexNet CNN 网络计算下的 4096 维特征，输出为该类的得分，

训练的是 SVM 权重向量；
由于负样本太多，采用 hard negative mining 的方法在负样本中选取有代表性的负样本，该方法具体见。

4.Bounding-box regression 训练

样本	来源
正样本	与Ground Truth相交IoU最大的Region Proposal，并且IoU>0.6的Region Proposal

输入数据为某类型样本对N个：$\{(P^i, G^i)\}_{i=1\cdots N}$以及$P^i_{i=1\cdots N}$所对应的AlexNet CNN网络Pool5层特征$\phi_5(P^i)_{i=1\cdots N}$，输出回归后的建议框Bounding-box，训练的是$d_x(P)$，$d_y(P)$，$d_w(P)$，$d_h(P)$四种变换操作的权重向量。具体见前面分析。

。解释分析

1.什么叫有监督预训练？为什么要进行有监督预训练？

有监督预训练也称之为迁移学习，举例说明：若有大量标注信息的人脸年龄分类的正负样本图片，利用样本训练了 CNN 网络用于人脸年龄识别；现在要通过人脸进行性别识别，那么就可以去掉已经训练好的人脸年龄识别网络 CNN 的最后一层或几层，换成所需要的分类层，前面层的网络参数直接使用为初始化参数，修改层的网络参数随机初始化，再利用人脸性别分类的正负样本图片进行训练，得到人脸性别识别网络，这种方法就叫做有监督预训练。这种方式可以很好地解决小样本数据无法训练深层 CNN 网络的问题，我们都知道小样本数据训练很容易造成网络过拟合，但是在大样本训练后利用其参数初始化网络可以很好地训练小样本，这解决了小样本训练的难题。
这篇文章最大的亮点就是采用了这种思想，ILSVRC 样本集上用于图片分类的含标注类别的训练集有 1millon 之多，总共含有 1000 类；而 PASCAL VOC 2007 样本集上用于物体检测的含标注类别和位置信息的训练集只有 10k，总共含有 20 类，直接用这部分数据训练容易造成过拟合，因此文中利用 ILSVRC2012 的训练集先进行有监督预训练。

2.ILSVRC 2012 与 PASCAL VOC 2007 数据集有冗余吗？

即使图像分类与目标检测任务本质上是不同的，理论上应该不会出现数据集冗余问题，但是作者还是通过两种方式测试了 PASCAL 2007 测试集和 ILSVRC 2012 训练集、验证集的重合度：第一种方式是检查网络相册 IDs，4952 个 PASCAL 2007 测试集一共出现了 31 张重复图片，0.63%重复率；第二种方式是用 GIST 描述器匹配的方法，4952 个 PASCAL 2007 测试集一共出现了 38 张重复图片【包含前面 31 张图片】，0.77%重复率，这说明 PASCAL 2007 测试集和 ILSVRC 2012 训练集、验证集基本上不重合，没有数据冗余问题存在。

3.可以不进行特定样本下的微调吗？可以直接采用 AlexNet CNN 网络的特征进行 SVM 训练吗？

文中设计了没有进行微调的对比实验，分别就 AlexNet CNN 网络的 pool5、fc6、fc7 层进行特征提取，输入 SVM 进行训练，这相当于把 AlexNet CNN 网络当做万精油使用，类似 HOG、SIFT 等做特征提取一样，不针对特征任务。实验结果发现 f6 层提取的特征比 f7 层的 mAP 还高，pool5 层提取的特征与 f6、f7 层相比 mAP 差不多；
在 PASCAL VOC 2007 数据集上采取了微调后 fc6、fc7 层特征较 pool5 层特征用于 SVM 训练提升 mAP 十分明显；
由此作者得出结论：不针对特定任务进行微调，而将 CNN 当成特征提取器，pool5 层得到的特征是基础特征，类似于 HOG、SIFT，类似于只学习到了人脸共性特征；从 fc6 和 fc7 等全连接层中所学习到的特征是针对特征任务特定样本的特征，类似于学习到了分类性别分类年龄的个性特征。

4.为什么微调时和训练 SVM 时所采用的正负样本阈值【0.5 和 0.3】不一致？

微调阶段是由于 CNN 对小样本容易过拟合，需要大量训练数据，故对 IoU 限制宽松：Ground Truth+与 Ground Truth 相交 IoU>0.5 的建议框为正样本，否则为负样本；
SVM 这种机制是由于其适用于小样本训练，故对样本 IoU 限制严格：Ground Truth 为正样本，与 Ground Truth 相交 IoU＜0.3 的建议框为负样本。

5.为什么不直接采用微调后的 AlexNet CNN 网络最后一层 SoftMax 进行 21 分类【20 类+背景】？

因为微调时和训练 SVM 时所采用的正负样本阈值不同，微调阶段正样本定义并不强调精准的位置，而 SVM 正样本只有 Ground Truth；并且微调阶段的负样本是随机抽样的，而 SVM 的负样本是经过 hard negative mining 方法筛选的；导致在采用 SoftMax 会使 PSACAL VOC 2007 测试集上 mAP 从 54.2%降低到 50.9%。

&结果怎么样

1.PASCAL VOC 2010 测试集上实现了 53.7%的 mAP；

2.PASCAL VOC 2012 测试集上实现了 53.3%的 mAP；

3.计算 Region Proposals 和 features 平均所花时间：13s/image on a GPU；53s/image on a CPU。

&还存在什么问题

1.很明显，最大的缺点是对一张图片的处理速度慢，这是由于一张图片中由 selective search 算法得出的约 2k 个建议框都需要经过变形处理后由 CNN 前向网络计算一次特征，这其中涵盖了对一张图片中多个重复区域的重复计算，很累赘；

2.知乎上有人说 R-CNN 网络需要两次 CNN 前向计算，第一次得到建议框特征给 SVM 分类识

别，第二次对非极大值抑制后的建议框再次进行 CNN 前向计算获得 Pool5 特征，以便对建议框进行回归得到更精确的 bounding-box，这里文中并没有说是怎么做的，博主认为也可能在计算 2k 个建议框的 CNN 特征时，在硬盘上保留了 2k 个建议框的 Pool5 特征，虽然这样做只需要一次 CNN 前向网络运算，但是耗费大量磁盘空间；

3.训练时间长，虽然文中没有明确指出具体训练时间，但由于采用 RoI-centric sampling【从所有图片的所有建议框中均匀取样】进行训练，那么每次都需要计算不同图片中不同建议框 CNN 特征，无法共享同一张图的 CNN 特征，训练速度很慢；

4.整个测试过程很复杂，要先提取建议框，之后提取每个建议框 CNN 特征，再用 SVM 分类，做非极大值抑制，最后做 bounding-box 回归才能得到图片中物体的种类以及位置信息；同样训练过程也很复杂，ILSVRC 2012 上预训练 CNN，PASCAL VOC 2007 上微调 CNN，做 20 类 SVM 分类器的训练和 20 类 bounding-box 回归器的训练；这些不连续过程必然涉及到特征存储、浪费磁盘空间等问题。