Fast R-CNN

Ross Girshick
Microsoft Research

摘要

这篇文章针对物体检测提出了一种快速的基于区域的卷积网络方法(Fast R-CNN)。Fast R-CNN 建立在以前工作的基础上，使用深度卷积网络来高效分类物体候选。与之前的工作相比，Fast R-CNN 应用了一些创新，在提高检测准确率的同时，提高了训练和测试速度。在训练非常深的 VGG16 时，Fast R-CNN 比 R-CNN 快 9 倍，在测试时快 213 倍，并在 PASCAL VOC 上达到了更高的 mAP。与 SPPnet 相比，Fast R-CNN 在训练时快 3 倍，在测试时快 10 倍，同时更加准确。我们使用 Python 和 C++（使用 Caffe）实现了 Fast R-CNN，同时代码在开源 MIT 执照下可用，见https://github.com/rbgirshick/fast-rcnn。

1 简介

近来，深度卷积网络显著提高了图片分类和物体检测的准确率。与图片分类相比，物体检测是一个更具挑战性的任务，求解它需要更加复杂的方法。由于它的复杂性，近来的方法在缓慢且不优雅的多阶段管道中训练模型。

由于检测对于精确的物体定位的要求导致的复杂性，产生了两个主要的挑战。首先，必须处理大量物体候选位置（经常被称为“候选位置”）。其次，这些候选位置仅仅提供了粗略的位置，必须通过再次调整来得到精细的位置。解决这些问题的办法通常需要在速度、准确率或简便性上折中。

在这篇文章中，我们优化了基于卷积网络的 sota 物体检测器的训练过程。我们提出了一种可以同时学习分类物体候选并调整它们的空间位置的单一阶段训练算法。

最终的方法在训练深度检测网络（VGG16[**vgg**]）时比 R-CNN[**rcnn**] 快 9 倍，比 SPPnet[**spp**] 快 3 倍。在运行时，检测器网络处理每张图片的时间

为 0.3 秒（不包括候选位置生成时间），同时在 PASCAL VOV 2012 上达到了 66% 的准确率（R-CNN 的准确率为 62%）。

1.1 R-CNN 和 SPPnet

基于区域的卷积网络方法（R-CNN）[**rcnn**] 使用深度卷积网络为物体候选分类，达到了很好的物体检测准确率。然而，R-CNN 有显著缺点：

1. **训练是一个多阶段管道**。R-CNN 首先使用对数损失在物体候选上微调卷积网络。之后，它使用支持向量机（SVM）拟合卷积特征。那些 SVM 作为物体检测器，取代了通过微调习得的 softmax 分类器。在第三个训练阶段，学习得到了边界框回归器。

2. **训练在空间和时间上十分昂贵**。为了训练 SVM 和边界框回归器，需要从每张图片的每个物体候选提取特征并写入磁盘。在深度网络中，例如 VGG16，为了 VOC07 trainval 集合的 5k 张图片，这个过程需要 2.5 个 GPU 日。存储这些特征要求数百 G 的磁盘空间。

3. **检测物体慢**。在测试时，需要从每张测试图片的每个物体候选提取特征。使用 VGG16 进行检测，在 GPU 上处理一张图片的时间为 47 秒。

R-CNN 之所以慢是因为它没有共享计算，而是为每一个物体候选进行了一次卷积网络的前向传播。空间金字塔池化网络（SPPnets）[**spp**] 通过共享计算提高了 R-CNN 的速度。SPPnet 方法为整个输入图片计算了卷积特征图，之后通过从共享特征图中提取的特征向量来为每个物体候选分类。通过对特征图中候选位置对应的部分进行最大池化并得到一个固定大小的输出（例如，6×6）来为候选区域提取特征。池化得到多个输出尺寸之后像空间金字塔池化 [**sp**] 一样将它们连接。SPPnet 在测试阶段为 R-CNN 加速了 10 到 100 倍。训练时间也由于更快的候选位置特征提取减少了 3 倍时间。

然而 SPPnet 也有明显的缺陷。类似 R-CNN，SPPnet 的训练过程也是一个多阶段管道，包括特征提取、使用对数损失微调网络、训练 SVM 和最后的拟合边界框回归器。特征也被写入磁盘。但是不同于 R-CNN，[**spp**] 中提出的微调算法不能更新在空间金字塔池化前的卷积层。毫不意外地，这个局限（那些固定的卷积层）限制了深度网络的准确率。