SSD: Single Shot MultiBox Detector

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy,
Scott Reed, Cheng-Yang Fu, Alexander C. Berg

摘要

我们提出了一种在图片中使用单个深度神经网络进行物体检测的方法。我们的方法，名为 SSD，将边界框的输出空间离散为存在于特征图每个位置的一组具有不同高宽比和尺度的默认框。在预测时，网络为每个默认框中的每个物体类别的存在生成分数，并对框进行调整以更好的匹配物体形状。此外，该网络结合了来自具有不同分辨率的多个特征图的预测，以自然地处理各种大小的物体。相对于需要物体候选的方法，SSD 很简单，因为它完全消除了候选生成和后续像素或特征重采样阶段，并将所有计算封装在单个网络中。这使得 SSD 易于训练并可以直接明了地集成到需要检测组件的系统中。在 PASCAL VOC、COCO 和 ILSVRC 数据集上的实验结果证实，SSD 与使用额外物体候选步骤的方法相比，它们准确性可以相提并论，同时 SSD 速度要快得多，同时为训练和推理提供统一的框架。对于 300×300 的输入，SSD 在 Nvidia Titan X 上以 59 FPS 的速度在 VOC2007 上测试达到了 74.3%mAP 的准确率，对于 512×512 的输入，SSD 达到了 76.9%mAP 的准确率，优于同类最先进的 Faster R-CNN 模型. 与其他单阶段方法相比，即使输入图像尺寸更小，SSD 依然具有更好的准确性。代码位于：https://github.com/weiliu89/caffe/tree/ssd。

1 简介

当前最先进的物体检测系统是如下方法的变体：假设边界框，为每个框重新采样像素或特征，并应用高质量分类器。自选择性搜索 [1] 以来，该管道一直在检测基准上占上风，直到目前在 PASCAL VOC、COCO 和 ILSVRC 检测上的领先结果，都是基于 Faster R-CNN[2]，尽管有更深的特征，如 [3]。虽然准确，但这些方法对于嵌入式系统来说计算量太大，即使使用高端硬件，对于实时应用程序来说也太慢。这些方法的检测速度通常以每帧秒数 (SPF) 为单位，甚至是最快的高精度检测器 Faster R-CNN 的运行速度仅为

每秒 7 帧 (FPS)。已经有许多尝试通过攻击检测管道的每个阶段来构建更快的检测器（参见第 4 节中的相关工作），但到目前为止，显着提高速度只是以显着降低检测精度为代价。

本文提出了第一个基于深度网络的物体检测器，该检测器不对假设的边界框像素或特征进行重采样，并与重采样的方法一样准确。这使得高准确度的检测速度有了明显的提高（在 VOC2007 测试中，SSD 达到了 59 FPS，mAP 为 74.3%，而 Faster R-CNN 为 7 FPS，mAP 为 73.2% 或 YOLO 为 45 FPS，mAP 63.4%）。速度的根本提高来自于消除了边界框候选和随后的像素或特征重采样阶段。我们不是第一个这样做的（参见 [4,5]），但通过增加一系列的改进，我们方法比以前的尝试显著提高了准确性。我们的改进包括使用小卷积核来预测物体类别和边界框位置的偏移，为不同的高宽比检测使用单独的预测器（核），并将这些核应用于网络后期的多个特征图，来在多个尺度上进行检测。通过这些修改——特别是在不同尺度上使用多层预测——我们可以使用相对较低的分辨率输入实现高精确度，并进一步提高了检测速度。虽然这些贡献独立来看可能很小，但我们注意到所产生的系统将 PASCAL VOC 的实时检测精度从 YOLO 的 63.4% mAP 到我们的 SSD 的 74.3% mAP。这比最近非常引人注目的关于残余网络 [**resnet**] 的工作在检测精度上有者更大的相对改善。此外，高质量检测的速度的显著提高可以扩大计算机视觉的使用范围。

我们将我们的贡献总结如下：

- 我们介绍了 SSD，一种多类别的单次检测器，它比以前的单次检测器（YOLO）更快，而且明显更准确，实际上与进行显示区域候选和集合的较慢技术一样准确（包括快速 R-CNN）。
- SSD 的核心是使用应用于特征图上的小卷积核预测类别分数和固定的默认边界框的偏移量。
- 为了达到较高的检测精度，我们从不同尺度的特征图中产生不同尺度的预测，并显式按高宽比分开预测。
- 这些设计特点导致了简单的端到端训练和高精确度，甚至在低分辨率的输入图像上也是如此，进一步改善了速度与精确度的权衡。
- 实验包括在 PASCAL VOC、COCO 和 ILSVRC 上对不同输入尺寸的模型进行时间和精度分析，并与一系列最新的最先进的方法进行比较。

2 The Single Shot Detector (SSD)

本节将介绍我们提出的 SSD 检测框架（2.1节）和相关的训练方法（2.2节）。之后，第 3 节将介绍了对应数据集的具体模型细节和实验结果。

2.1 模型

SSD 方法基于一个前向传播卷积网络，它产生一个固定大小的边界框集合，并对这些框中存在的物体类别实例进行评分，然后通过一个非最大抑制步骤来产生最终的检测结果。前期的网络层是基于用于高质量图像分类的标准结构（在分类层之前截断），我们将其称之基础网络。然后，我们向网络添加辅助结构，以产生具有以下关键特征的检测结果：

用于检测的多尺度特征图 我们在截断的基础网络的末端添加卷积特征层。这些层的大小逐渐减少，并允许在多个尺度上预测检测。预测检测的卷积模型对于每个特征层都是不同的（参考 Overfeat[4] 和 YOLO[5]，它们在单一尺度的特征图上进行操作）。

用于检测的卷积预测器

2.2 训练