YOLOv3: An Incremental Improvement

Joseph Redmon Ali Farhadi

摘要

图像目标检测是图像处理以及工业领域重点研究的方向之一，其主要任务是从图像中定位目标并对目标种类进行判断。YOLO 是近几年一直处于目标检测领域的领先地位的算法。YOLO 将目标检测概括为一个回归问题，实现端到端的训练和检测，具有良好的速度精度平衡。文章作者对其在 2016 年提出的 YOLO 目标检测方法做了一系列改进。首先，作者借鉴 ResNet 的残差结构，实现了 Darknet-53 主干网络。其次，增加了多尺度检测，参考特征金字塔 FPN，通过多尺度特征融合，实现了更好的多尺寸目标预测效果。作者在 ImageNet 和 COCO 数据集上测试了该方法的性能。实验结果中，YOLOv3 在不损失过多检测精度的同时，具有很好的实时性能。

关键词：Darknet-53；多尺度目标检测

1 引言

在 YOLO 被提出之前，就已经有较成熟的 R-CNN 系列算法，这些算法都是基于区域提议和位置回归两个步骤完成的，虽然 Faster R-CNN 提出使用 RPN 进行区域提议后，检测的速度提升了不少（可达到 5 fps），但仍然无法应用到视频实时检测中。为了提升速度，减少检测模型的推理开销，YOLO 使用单阶段检测方法，抛弃了区域提议步骤，将目标检测问题看作是一个回归问题，直接从图像像素信息得到边界框的位置和类别概率。因为没有了区域提议，所以 YOLO 只需要将原始图片或中间的特征层处理一次即可，从而实现实时目标检测。

2 相关工作

2.1 传统目标检测算法

传统目标检测算法主要使用窗口滑动来对候选框进行定位、使用手工方式对特征进行提取和选择。其中特征提取方式对检测结果影响较大，通常利用方向梯度直方图（HOG）对候选框窗口进行判断，整个检测过程效率与精度都较低，所以传统方法逐渐被淘汰。计算能力的大幅度提高可以加快深度学习在图像领域的应用，优秀的深度学习的方法逐渐替代了传统的目标检测算法。

2.2 基于深度学习的两阶段目标检测算法 Faster R-CNN

Faster R-CNN[1]是著名的两阶段算法。Faster R-CNN 主要由特征提取网络、RPN 网络、感兴趣区域池化层（ROI Pooling）、坐标回归以及类别分类构成。Faster R-CNN 的主干特征提取网络的选取比较灵活，根据不同的用途，可以使用 VGG，Resnet、MobileNet 等网络对图像特征进行提取。该算法需要将输入图像调整成合适的大小再输入主干网络中，在主干网络提取特征后会得到一个共享特征层（Feature Map）, 该特征层一部分需要进行候选框的提取，另一部分需要对感兴趣区域进行池化。候选框提取后会映射到 Feature Map 上，然后对映射的部分进行 MaxPooling，得到 ROI Pooling 层的输出。最后，再通过 Softmax 函数进行分类和边框的回归得到最终的检测框。

3 本文方法

3.1 YOLOv3 概述

YOLO 的全称是 You only look once.YOLOv1 第一次使用回归的思想进行目标检测。使用深度学习，让模型在输入的图片与输出的目标的坐标位置以及类别之间进行回归。YOLOv2 在 v1 的基础上将骨干网络替换为 DarkNet。YOLOv3[2]进一步改进了网络架构，引入了残差连接，然后还使用不同尺度的特征图进行预测，极大提高了模型进行多尺度目标检测的能力。YOLOv3 没有太多的创新，主要是借鉴一些好的方案融合到 YOLO 里面。在保持速度优势的前提下，提升了预测精度，尤其是加强了对小物体的识别能力。

3.2 多尺度预测

YOlOv3 提取多特征层进行目标检测，一共提取三个特征层。三个特征层位于主干部分 Darknet53 的不同位置，分别位于中间层，中下层，底层，三个特征层的 shape 分别为 (52,52,256)、(26,26,512)、(13,13,1024)。获得三个有效特征层后，YOLOv3 利用这三个有效特征层进行 FPN 层的构建。13x13x1024 的特征层进行 5 次卷积处理，处理完后利用 YoloHead 获得预测结果，一部分用于进行上采样 UmSampling2d 后与 26x26x512 特征层进行结合，结合特征层的 shape 为 (26,26,768)。结合特征层再次进行 5 次卷积处理，处理完后利用 YoloHead 获得预测结果，一部分用于进行上采样 UmSampling2d 后与 52x52x256 特征层进行结合，结合特征层的 shape 为 (52,52,384)。结合特征层再次进行 5 次卷积处理，处理完后利用 YoloHead 获得预测结果。利用这三个 shape 的特征层传入 YOlO Head 获得 3 个尺度的预测结果。此外，作者沿用了 k-means 方法聚类得到 3 个尺度的 9 个先验框。

3.3 Darknet-53

YOLOv3 使用了一个新的网络 Darknet-53 作为特征提取器。该网络融合了 v2 中的 Darknet-19 和残差网络模块。YOLOv3 在 Darknet-53 中使用了连续的 3x3 和 1x1 卷积层。同时使用残差网络的方法进行连接。在 ImageNet 上，Darknet-53 相较于 Darknet-19 提高了 3%。与 ResNet-101、ResNet-152 的准确率相当，但是推理效率更高。

4 复现细节

4.1 与已有开源代码对比

复现代码使用了 pytorch 提供的 nn.module 类，实现了 Darknet-53 主干、用于多尺度特征融合的颈部以及损失函数。模型训练代码复现部分参考了 GitHub 用户的实现[3]。参考了其先冻结主干训练 50 个 epoch，再解冻主干训练整体网络的做法。相较于其他实现，复现代码提供了修改网络结构的参数。对原实现进行了精简，保留骨干功能。

4.2 实验环境

硬件环境：CPU：Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz、GPU：Quadro P5000 * 2
软件环境：Python 3.9.12、pytorch 1.13.0、conda 22.9.0
测试数据集：VOC2017

4.3 创新点

参考 Tolga AKSOY 等人的实现[4]，尝试在网络中加入注意力机制。相较于 YOLOv3，他们提出的 YOLOv3-Reasoner2 在 COCO 数据集上的 mAP 提高了 2.5%。如图 1所示，在 FPN 与 YOLO head 之间加入 ViT 模块处理 FPN 输出的不同大小的三维张量。该方法相当于给模型加入了一个具有更大感受野的层。

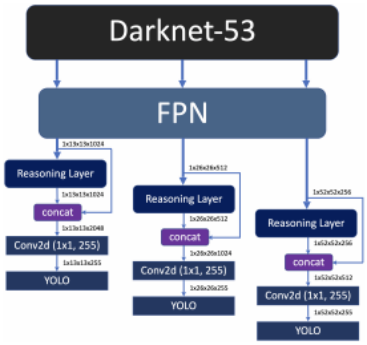

图 1: YOLOv3 结合 ViT 网络示意图

5 实验结果分析

如表 1所示，在 VOC2017 数据集上，结合 ViT 的 YOLOv3 的检测效果与原始方法区别不大。考虑到 VOC2017 数据集中的目标以大目标为主，而 COCO 数据集以小目标居多，该方法没有发挥其独特的注意力机制能力。

Model	**AP**$_{50}$	**AP**$_{75}$
YOLOv3	0.815	0.617
YOLOv3+ViT	0.819	0.598

表 1: YOLOv3 与 YOLOv3+ViT 在 VOC2017 数据集上的精度

6 总结与展望

本次复现实现了 YOLOv3 的核心模块，包括 Darknet-53、多尺度特征融合网络和损失函数。本次复现尝试加入注意力机制，让网络关注目标之间的联系。但是没有取得更好的效果。

参考文献

[1] GIRSHICK R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[2] REDMON J, FARHADI A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[3] Bubbliiiing. yolo3-pytorch[Z]. https://github.com/bubbliiiing/yolo3-pytorch. 2020.

[4] AKSOY T, HALICI U. Analysis of visual reasoning on one-stage object detection[J]. arXiv preprint arXiv:2202.13115, 2022.