计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(9):149−155 [doi: 10.15888/j.cnki.csa.007452]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

感受野特征增强的 SSD 目标检测算法①

谭 龙, 高 昂

(黑龙江大学 计算机科学与技术学院, 哈尔滨 150080)
通讯作者: 谭 龙, E-mail: tanlong@hlju.edu.cn

摘 要: SSD (Single Shot multi-box Detector) 算法是在不同层的特征图上, 进行多尺度对象的检测, 具有速度快和精度高的特点. 但是, 传统 SSD 算法的特征金字塔检测方法很难融合不同尺度的特征, 并且由于底层的卷积神经网络层具有较弱的语义信息, 也不利于小物体的识别, 因此本论文提出了以 SSD 算法的网络结构为基础的一种新颖的目标检测算法 RF_SSD, 该算法将不同层及不同尺度的特征图以轻量级的方式相融合, 下采样层生成新的特征图, 通过引入感受野模块, 提高网络的特征提取能力, 增强特征的表征能力和鲁棒性. 和传统 SSD 算法相比, 本文算法在精度上有明显提升, 同时充分保证了目标检测的实时性. 实验结果表明, 在 PASCAL VOC 测试集上测试, 准确率为 80.2%, 检测速度为 44.5 FPS.

关键词: SSD 算法; 目标检测; 卷积神经网络; 感受野; 计算机视觉

引用格式: 谭龙,高昂.感受野特征增强的 SSD 目标检测算法.计算机系统应用,2020,29(9):149–155. http://www.c-s-a.org.cn/1003-3254/7452.html

SSD Object Detection Algorithm with Feature Enhancement of Receptive Field

TAN Long, GAO Ang

(School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China)

Abstract: SSD (Single Shot multi-box Detector) algorithm is used to detect multi-scale objects on feature maps of different layers, which has the characteristics of fast speed and high accuracy. However, the feature pyramid detection method of traditional SSD algorithm is difficult to fuse the features of different scales, and because the convolutional neural network layer at the bottom has weak semantic information and is not conducive to the recognition of small objects, so this paper proposes a novel object detection algorithm RF_SSD based on the network structure of SSD algorithm. In this algorithm, feature maps of different layers and scales are fused in a lightweight way, and new feature maps are generated in the lower sampling layer. By introducing the receptive field module, the feature extraction ability of the network is improved, and the characterization ability and robustness of the feature are enhanced. Compared with the traditional SSD algorithm, the accuracy of the proposed algorithm is significantly improved, and the real-time performance of object detection is fully guaranteed. The experimental results show that the accuracy is 80.2% and the detection speed is 44.5 FPS on the PASCAL VOC test set.

Key words: SSD algorithm; object detection; convolutional neural network; receptive field; computer vision

目标检测是计算机视觉领域的一项重要任务, 是生活中如实例分割[1], 面部分析[2], 汽车自动驾驶[3]、视频分析[4] 等各种视觉应用的先决条件.

近些年, 伴随着深度卷积神经网络的充分发展[5]以及良好的数据集注释先验工作的积累[6], 物体检测器的性能得到了显著提高. 但是, 物体检测过程中的尺度

① 基金项目: 国家自然科学基金面上项目 (81373537); 黑龙江省自然科学基金面上项目 (F201434)
Foundation item: General Program of National Natural Science Foundation of China (81373537); General Program of Natural Science Foundation of Helongjiang Province, China (F201434)
收稿时间: 2019-11-11; 修改时间: 2019-12-09; 采用时间: 2019-12-20; csa 在线出版时间: 2020-09-04

变化仍然是所有检测器的关键挑战, 为了识别不同尺度的物体, 早期大多数的检测器都是基于手工制作的特征[7], 并且利用图像金字塔. 考虑到内存和检测时间, 这些工作无论在计算还是花费上都是昂贵的. 得益于卷积神经网络的发展, 手工设计的特征已逐渐被卷积神经网络计算的特征所取代. 最近的检测系统[8,9] 利用卷积神经网络 (ConvNets) 在单个输入尺度图像依次进行运算, 获得不同尺度的特征图, 然后用最顶层特征图来预测具有不同尺度和纵横比的候选边界框. 然而, 最顶部的特征图具有固定的感受野, 与自然图像中的不同尺度的物体冲突. 特别是小物体在最顶层上几乎没有信息, 因此可能会损害物体检测性能, 尤其是小物体.

在解决多尺度问题方面, SSD 利用从下到上的特征金字塔来适应各种尺寸的物体, 然而, SSD 算法的特征金字塔形式未能利用深层特征图中强大的语义信息, 这对于小物体检测至关重要. 因为语义信息对于检测视觉上困难的物体 (例如小的, 遮挡的物体) 是决定性的, 为了克服 SSD 的缺点并使网络对对象尺度更加稳健, 最近的工作 (例如 FPN[9], DSSD[10], RON[11]) 建议将低分辨率带有强语义信息的特征图同具有高分辨率但带有弱语义弱信息的特征图通过自上而下的通道横向连接. 与 SSD 中的自下而上的方式相比, 横向连接将语义信息一个接一个地传递到浅层, 从而增强了浅层特征的检测能力. 与传统检测器相比, 这些网络在精度方面有着显著的提高. 但是我们注意到这些在最顶层特征图中使用反卷积层的方法完全丢失了小物体的精细细节.

本文致力于提高小物体的检测性能, 缓解 SSD 算法的尺度变化问题, 同时又不失实时检测速度. 通常, 较深层中的深层特征对于分类子任务更具辨别性, 而较浅层中的浅层特征则对于物体位置回归子任务更有利. 此外, 浅层特征更适合于具有简单外观的特征对象, 而深层特征适用于具有复杂外观的对象. 基于此, 本文通过特征融合模块将具有语义信息的深层特征添加到浅层特征中, 以获得具有丰富信息的特征图, 将来自不同层次的不同尺度的特征图投影并连接在一起, 然后用 BN[12] 层进行归一化处理, 最后附加下采样层以生成新的特征金字塔, 此外, 添加了感受野模块 (RFM), 以加强从轻量级 CNN 模型中学到的深层特征, 使它们有助于检测器快速准确. 与传统 SSD 相比, 本文算法 RF_SSD 主要选择 VGG16 作为骨干网络, 而不是更深层次的 ConvNets (例如 ResNet[13] 或 DenseNet[14]), 原因是深层卷积神经网络 (ConvNets) 虽然对特征提取有利但会加大计算量同时降低检测速度, 实验表明本文所提出的结构在精度上比 SSD 算法有所提升. 本文的贡献主要表现为以下几点:

(1) 提出了新颖的、轻量级的特征融合方式, 主要是将不同层的特征图合并, 并生成特征金字塔, 降低了重复检测一个对象的多个部分或者多个对象合并到一个对象的检测概率, 同时小物体检测表现更好.

(2) 借鉴混合空洞卷积和 Inception 结构, 设计并添加感受野模块来增强网络的特征提取能力, 同时在不增加卷积参数的前提下增大卷积感受野, 加强轻量级卷积神经网络学到的深层特征, 保证检测器的实时性.

(3) 在 PASCAL VOC 数据集上进行了定性与定量的实验, 结果表明, 同传统 SSD 算法相比, 本文所提出的算法在目标检测性能上有显著的提升, 同时以相对低的速度损耗提高了小物体的准确率.

1 相关工作

在目标检测算法研究中, 无论是在单阶段检测器还是两阶段检测器中, 相关研究者都投入了大量的工作来改善目标检测中的尺度变化问题, 大致可分为两种策略. 一种是图像金字塔, 通过图像的尺度变化来产生具有语义代表性的多尺度特征, 然后用来自不同尺度的图像的特征分别产生预测, 最后将这些预测放在一起进行评估以给出最终预测. 在识别精度和定位精度方面, 来自多尺寸图像的特征确实超越仅基于单尺度图像的特征. 诸如 OHEM[15] 和 SNIP[16] 之类的方法都采用了这种策略. 虽然性能得到了提升, 但这种策略在时间和内存方面花销很大, 所以在实时任务中很难得到应用. 另一种是利用网络内的特征金字塔以较低的计算成本来模拟图像金字塔. 该策略比第一个策略需要的内存和计算成本要少得多, 从而可以在实时网络的训练和测试阶段中进行使用. 此外, 特征金字塔构建模块可以很容易地修改, 并应用在最先进的基于深度神经网络的探测器. MS-CNN[17], SSD[8], DSSD[10], FPN[9], YOLOv3[18], RetinaNet[19] 和 RefineDet[20] 以不同的方式采用了这种策略.

此外, MS-CNN[17] 提出了两个子网络, 并首先将多尺度特征结合到用于物体检测的深度卷积神经网络中. 提议子网利用几种分辨率的特征图来检测图像中的多

尺度物体. SSD 利用 VGG16 网络的后几层的特征图和额外特征层进行多尺度预测. FPN 将高层特征与低层特征相结合, 由最近邻居上采样和横向连接实现. DSSD 实现了反卷积层, 用于聚合上下文和增强浅层特征的高级语义信息, RefineDet[20] 采用了两步级联回归, 在保持 SSD 效率的同时, 在准确性方面取得了显著提高.

2 RF_SSD 算法

本节将在 SSD 框架基础上, 分析算法涉及到的特征融合处理、感受野模块的设计以及算法的具体处理过程.

SSD 采用不同尺度的特征图来检测物体, 以 VGG16[21] 作为骨干网络, 采用级联卷积的方式生成不同尺度的特征图, 结合 YOLO 的回归思想和 Faster-RCNN 的 Anchor 机制, 使用全图各个位置的多尺度区域特征进行回归, 既保证检测速度又保持了精度. 同时在对特征图预测时, 采用卷积核来预测一系列 Default Bounding Boxes 的类别和坐标偏移.

由于小物体不会在浅层中丢失太多的位置信息, 并且大物体也可以在较深层中很好地定位和识别, 所以 SSD 算法使用浅层特征图检测小物体, 深层特征图检测大物体这种策略是合理的, 但问题是由浅层产生的小物体的特征缺乏足够的信息, 这将导致小物体检测性能的不良. 此外, 小物体也严重依赖于上下文信息, SSD 网络结构如图 1 所示.

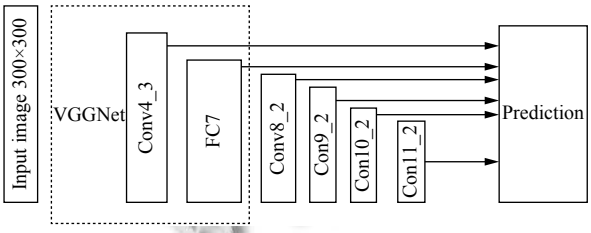

图 1 SSD 算法结构图

2.1 特征融合 (Feature Fusion)

针对传统 SSD 算法缺点, FPN 和 DSSD 利用顶层特征的反卷积层, 经过验证, 这种方法可以大大提高传统探测器的性能, 但却需要多个功能合并过程. 而且右侧的新特征只能融合相应的左侧和更高层级的特征[9,10]. 此外, 潜在特征和大量特征的 element-wise process 过程也会消耗大量时间. 基于此, 本文提出了一种轻量级和高效的特征融合模块来处理这项任务. 本文的动机是以适当的方式一次融合不同级别的特征, 并从融合特征生成特征金字塔.

传统的 SSD300 是基于 VGG16 的, 作者选择 Conv4_3, FC7 和新添加的 Conv8_2, Conv9_2, Conv10_2, Conv11_2 层特征图进行检测. 相应的特征图的大小为 38×38, 19×19,10×10,5×5,3×3 和 1×1. 本文认为大小小于 10×10 的特征图太小而几乎没有要合并的信息, 所以本文先将 Conv8_2 的 stride 设为 1, 这样 Conv9_2 的大小为 10×10, 然后本文选择 Conv4_3, Conv9_2 融合为新的特征图, 增强了浅层特征的语义信息, 同时也有很强的几何细节信息表征能力.

在传统的处理方法中, 主要有两种方法合并不同的特征图: concatenation 及 element-wise summation. Element-wise summation 要求特征图的通道相同, 这意味着我们必须将特征图转换为相同的通道. 由于此要求限制了融合特征图的灵活性, 所以我们选择用 concatenation 方式. 为了使 Conv4_3, Conv9_2 融合为新的特征图, 需对 Conv9_2 进行上采样处理. 如图 2 所示, 首先使用大小为 2×2, 通道数为 256 的反卷积核进行上采样, 将输出通过 3×3 的卷积核映射至 BN 层, 然后再到下一个反卷积核. Con4_3 通过 1×1 的卷积核直接映射输出至 BN 层. 最终将 Conv4_3 通过 1×1 卷积层的输出与 Con9_2 经过两层反卷积层的输出进行 concat 操作, 之后传入至 ReLU 层, 再通过 L2 Normalization 层做归一化处理, 同时增加模型的鲁棒性.

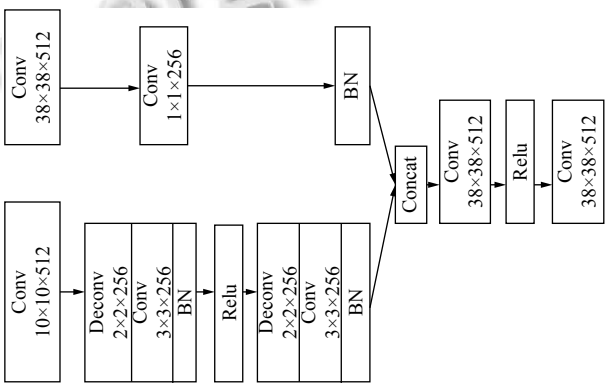

图 2 特征融合模块

2.2 感受野模块

本模块采用多支路卷积形式, 其内部结构可以分为两个部分: 多支路卷积层和空洞卷积层. 多支路卷积层的结构和 Inception 相同, 模拟不同尺寸的感受野, 空洞卷积层利用空洞卷积模拟不同尺寸感受野之间的关

系[22]. 在卷积神经网络中, 卷积核的感受野大小和卷积核的尺寸成正相关, 通过改变卷积核的尺寸可以获得不同大小的感受野, 进而更加有效的利用特征信息. 本文的设计借鉴了 Inception-V4 和 Inception -ResNet[23], 结构如图 3 所示, 首先在每个分支结构中使用 1×1 的卷积层, 减少特征图中通道数量, 用 2 个连续的 3×3 Conv 替代 Inception 模块中的 5×5 Conv, 从而实现网络深度的增加, 之后将原有 3×3 的卷积核分解成两个一维的卷积核 (1×3 和 3×1), 目的是加速计算, 同时网络宽度增加, 增加了网络的非线性. 除此之外, 为了保留更多的原始特征信息, 增加了一条剪接支路.

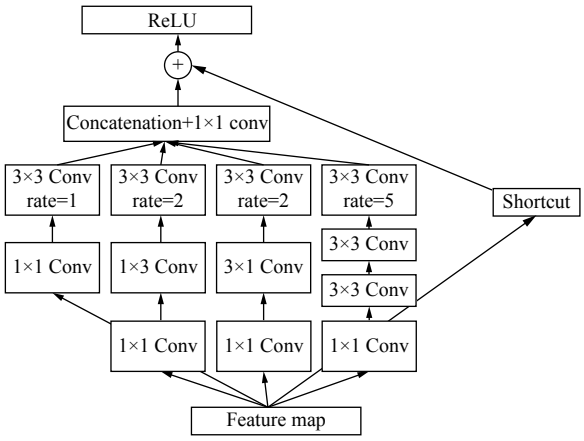

图 3 RFM 模块

本算法感受野模块结构上借鉴了混合空洞卷积和 Inception, 混合空洞卷积 (hybrid dilated convolution) 由文献 [24] 提出, 通过叠加多个不同空洞率的空洞卷积来避免网格效应和平衡不同尺寸感受野之间的关系, 解决了传统卷积神经网络采用池化层所造成的内部数据结构遗失和小物体信息无法重建等问题, 同时协调多支路卷积, 在很好的结合多支路卷积的同时提高了算法的检测效率.

2.3 算法结构

本文算法是以 SSD 算法框架为基础构建的, 提出新的特征融合模块来充分利用深层的特征信息以此提高算法的检测精度, 同时改善小物体检测的效果, 另外, 通过在特征提取网络上添加感受野模块来提高特征的提取能力. 无论特征融合模块还是感受野模块都比较简单, 所以在极大程度上保留了 SSD 原有的网络结构, 保证了检测速度. 整体的算法结构如图 4 所示, 骨干网络采用 VGGNet, 先对 Con9_2 层特征图进行尺寸调整, 后将调整尺寸后的 Con9_2 层特征图与 Con4_3 层特征图传入 Feature Fusion 模块产生新特征图, 经 BN 层后, 通过一系列下采样形成特征金字塔, 同时加入感受野模块. 具体描述如下: 第一是将 Conv8_2 的 stride 设为 1, 这样 Conv9_2 的大小为 10×10, 然后使用两层反卷积核对 Conv9_2 进行上采样处理. 之后将输出和经过 BN 层的 Conv4_3 进行 concat 操作, 之后传入至 ReLU 层, 再通过 L2Normalization 层做归一化处理. 第二, 对新得到的特征图进行下采样 (包含一些 1×1 和 stride 为 2 的 3×3 的卷积层来改变通道数和特征图的大小), 形成新的特征金字塔, 同时利用新添加的感受野模块对新的特征信息进行检测. 第三, 用 RFM 替换掉中间两层卷积层, 考虑到最后两个卷积层的尺寸, 将最后两层保持不变.

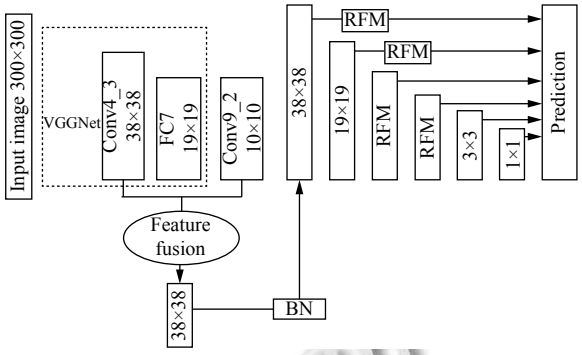

图 4 本文的算法结构

本文的损失函数采用了传统的 SSD 算法的处理方式, 回归函数输出物体的位置坐标, Softmax 函数进行预测分类. 总的损失函数为位置误差 (locatization loss, loc) 与置信度误差 (confidence loss, conf) 的加权和:

$$L(x,c,l,g)=\frac{1}{N}\left(L_{\text{conf}}(x,c)+\alpha L_{\text{loc}}(x,l,g)\right) \quad (1)$$

其中, N 是先验框的正样本数量. c 为类别置信度预测值. l 为先验框所对应边界框的位置预测值, x 为预测框的类别匹配信息, 而 g 是 ground truth 的位置参数. 权重系数 α 通过交叉验证设置为 1.

位置误差定义为:

$$L_{\text{loc}}(x,l,g)=\sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}^{k}\,smooth_{l1}\left(l_i^m-\hat{g}_j^m\right) \quad (2)$$

$$\hat{g}_j^{cx}=\left(g_j^{cx}-d_i^{cx}\right)/d_i^w \quad (3)$$

$$\hat{g}_j^{cy} = \left(g_j^{cy} - d_i^{cy}\right)/d_i^h \tag{4}$$

$$\hat{g}_j^{w} = \log\left(\frac{g_j^{w}}{d_i^{w}}\right) \tag{5}$$

$$\hat{g}_j^{h} = \log\left(\frac{g_j^{h}}{d_i^{h}}\right) \tag{6}$$

对于位置误差, 其采用 Smooth L1 loss, 定义如下:

$$smooth_{l1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{7}$$

置信度误差, 其采用 Softmax loss, 定义如下:

$$\begin{cases} L_{\text{conf}}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^p \log\left(\hat{c}_i^p\right) - \sum_{i\in Neg} \log\left(\hat{c}_i^0\right) \\ \text{where} \quad \hat{c}_i^p = \dfrac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)} \end{cases} \tag{8}$$

其中, $x_{ij}^p \in \{1,0\}$为一个指示参数, 若$x_{ij}^p = 1$则表明第 i 个先验框与第 j 个 ground truth 匹配, 并且 ground truth 的类别为 p. 参数 Pos 代表正样本, Neg 代表负样本. m 包含了边框的中心坐标 (cx, cy) 和边框的宽 (w)、高 (h); x 表示先验框和 ground truth 匹配上的指示参数; g 为 ground truth 的位置参数.

3 实验分析

3.1 数据增强

在进行训练之前, 可先通过数据增强的方式对数据进行预处理, 以此提高数据集的多样性, 使模型有更高的鲁棒性. 常用的数据增强方式如随机翻转、缩放、颜色变化和裁剪等. 通过将数据增强方式应用到训练当中, 可使模型学到旋转不变性和对称不变性.

3.2 网络训练策略

本文算法采用与 SSD 算法相似的训练策略, 都是使用训练好 VGGNet 网络, 本文使用 PASCAL VOC 2007 和 PASCAL VOC 2012 数据集, 同时把与真实框 (ground truth) 的交并比 (IOU) 大于 0.5 的预测框认为是正样本. 采用平均精度 (mAP) 作为评测算法性能的度量, 帧速 (Frame Per Second, FPS) 作为目标检测速度的评价指标. 训练阶段将输入图像的大小设为 300×300 像素. 训练时, 我们用 VOC 2007 trainvaland VOC2012 trainval (VOC07+12) 的联合数据集训练, 在 VOC 2007 test 测试集上测试. 本文的硬件环境为深度学习框架 Caffe, ubuntu16.04 系统, GPU 显卡型号为 NVIDIA 1080Ti, Batch size=16, 初始学习率设定为 0.001, max_epoch 设置为 180 K, 然后在步骤 100 K, 140 K 和 180 K 除以 10. 将权重衰减设置为 0.0005. 和 SSD 算法一样采用动量为 0.9 的 SGD 来优化本文算法.

3.3 PASCAL VOC2007 测试结果分析

PASCAl VOC 是一个用于物体分类识别和检测的标准数据集, 该数据集包括 20 个类别, 表 1 为 PASCAl VOC 具体类别.

表 1 PASCAL VOC 数据集类别

种类	类别
人	人
动物	猫, 狗, 马, 羊, 牛, 鸟
交通工具	自行车, 火车, 摩托车, 轿车, 公共汽车, 飞机, 船
其他	盆栽, 电视, 沙发, 水瓶, 椅子, 餐桌

本文算法模型与主流目标检测算法在 VOC2007 数据集的实验结果如表 2 所示, 本文所提出的算法准确率达到了 80.2%, 比传统的 SSD 算法有 2.7% 的提升, 比 DSSD 算法提高了 1.2%, 但比 R-FCN 低了 0.3%, 原因在于 R-FCN 算法使用 ResNet-101 作为基础网络, 相比于 VGG-16, 算法网络结构更深, 提取特征的能力更强, 但同时也降低了网络的检测速度. 本文算法的检测速度为 44.5 FPS, 虽然相比于传统的 SSD 算法, 速度也有所下降, 但满足实时检测需求.

表 2 不同目标检测算法在 PASCAL VOC 2007 上的检测结果

方法	数据	backbone	GPU	FPS	mAP
Faster-RCNN	07+12	VGGNet	Titan X	7	73.2
R-FCN	07+12	ResNet-101	Titan X	9	80.5
YOLOv2	07+12	Darknet-19	Titan X	81	73.7
SSD300	07+12	VGGNet	1080Ti	85	77.5
DSOD300	07+12	DS/64-192-48-1	Titan X	17.4	77.7
RSSD300	07+12	VGGNet	Titan X	35	78.5
DSSD321	07+12	ResNet-101	Titan X	9.5	78.6
Proposed	07+12	VGGNet	1080Ti	44.5	80.2

不同目标检测算法在精度和速度上的分布如图 5 所示, Faster-RCNN, R-FCN, YOLOv2, DSOD, RSSD, DSSD 算法是在 Titan X GPU 上进行测试的, 而 SSD 和本文提出的算法是在 1080 Ti GPU 上测试的. 从图 5 中也可看出本文的算法在检测速度和精度上有着一定的优势.

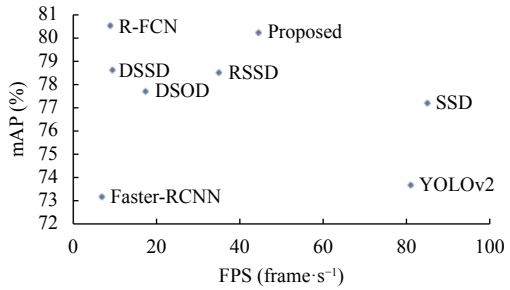

图 5 不同的检测算法在检测速度和精度上的分布

本文将传统的 SSD 算法和 RF_SSD 算法在每一类目标检测的精度上进行比较, 结果如表 3 所示. 从表中可知, 飞机, 自行车, 鸟, 船, 瓶子, 公交车等类别都有显著的提升, 其中, 瓶子, 盆栽的检测精度较低, 虽然得益于本文提出的网络结构, 相比于传统的 SSD 算法, 精确度有所提升, 但因物体相比于其他类别太小, 特征提取较少, 导致相应检测精度不高. 但总体来说, 本文算法相比于 SSD 算法 mAP 提高了 2.7%, 基本满足实际需求, 同时也论证了本文算法思想的可行性.

同时本文对比了感受野模块对算法检测结果的影响 (参见表 4), 通过实验可知添加感受野模块可提高算法的准确率, 说明感受野对算法性能有一定的提升, 同时由于感受野模块采用多个支路卷积, 提高了模型的复杂度, 所以导致检测速度降低.

表 3 本文算法在 PASACAL VOC2007 测试集上单个类别的测试结果

方法	network	mAP	aero	bike	bird	boad	bottle	bus	car	cat	chair	train
SSD300	VGG-Net	77.5	79.5	83.9	76.0	69.6	50.5	87.0	85.7	88.1	60.3	87.6
RF_SSD	VGG-Net	80.2	87.2	87.4	83.2	73.2	57.6	88.3	88.2	88.4	62.2	85.8
方法	network	mAP	cow	table	dog	horse	mbike	person	plant	sheep	sofa	tv
SSD300	VGG-Net	77.5	81.5	77.0	86.1	87.5	83.97	79.4	52.3	77.9	79.5	76.8
RF_SSD	VGG-Net	80.2	83.6	75.4	84.8	89.4	87.0	82.3	56.4	82.4	83.5	77.7

表 4 感受野模块对算法准确率的影响

Add RFM	FPS	mAP
×	65.6	78.8
√	44.5	80.2

最后本文分析了不同卷积层融合后的结果, 结果如表 5, 若融合 Conv3_3, Conv4_3, 和 Conv9_2, 则在 VOC2007 的 mAP 为 79.8%, 若去掉 Conv3_3, 则 mAP 为 80.2%, 表明 Conv3_3 对检测器的结果并没有太大的影响, 原因在于 Conv3_3 卷积层提取的特征图包含较多的背景噪声. 此外, 本文从 COCO 数据集中随机挑选了几张照片, 测试结果如图 6 所示.

表 5 不同层融合的测试结果

融合层	mAP
Conv3-Conv9	79.8
Conv4-Conv9	80.2

4 结论

本文基于 SSD 算法, 提出了一种新颖高效的目标检测算法, 通过将不同层的特征图以轻量级的方式融合在一起, 使新的特征图既有深层特征的语义信息, 同时又有高分辨率, 然后采用下采样层生成特征金字塔, 之后设计添加感受野模块, 提高网络的特征提取能力, 提高了算法的整体精度, 也改善了小目标的检测效果. 本文的算法在精度上超越了传统 SSD 算法以及一系列其他目标检测算法, 由于添加感受野模块, 增加了网络特征的提取能力, 增加了精度, 但加深了深度和模型复杂度, 导致检测速度降低, 虽以速度换取精度, 但基本满足实时检测要求. 和大多数单阶段目标检测结构一样, 本算法类别不平衡问题依旧未能得到解决. 未来, 将继续改进该算法, 使用 anchor-free 模型方法或进一步修改目标损失函数改善类别不平衡问题, 设计轻量型的特征提取和融合网络结构, 在不降低精度的同时提高速度.

图 6 COCO 2017 上的实例检测结果

参考文献

1 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy. 2017. 2980–2988.

2 Zheng YT, Pal DK, Savvides M. Ring loss: Convex feature normalization for face recognition. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 5089–5097.

3 Dollár P, Wojek C, Schiele B, *et al*. Pedestrian detection: A benchmark. Proceedings of 2009 IEEE Computer Vision and Pattern Recognition. Miami, FL, USA. 2009. 304–311.

4 Wang XL, Gupta A. Videos as space-time region graphs. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2018. 413–431.

5 Xie SN, Girshick R, Dollár P, *et al*. Aggregated residual transformations for deep neural networks. Proceedings of 2017 Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. 2017. 5987–5990.

6 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 740–755.

7 Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

8 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multiBox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 21–37.

9 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 2117–2125.

10 Fu CY, Liu W, Ranga A, *et al*. DSSD: Deconvolutional single shot detector. arXiv: 1701.06659, 2017.

11 Kong T, Sun FC, Yao AB, *et al*. Ron: Reverse connection with objectness prior networks for object detection. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 5936–5944.

12 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France. 2015. 448–456.

13 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.

14 Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. 2017. 4700–4708.

15 Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. 2016. 761–769.

16 Singh B, Davis LS. An analysis of scale invariance in object detection–SNIP. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 3578–3587.

17 Cai ZW, Fan QF, Feris RS, *et al*. A unified multi-scale deep convolutional neural network for fast object detection. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 354–370.

18 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

19 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318–327. [doi: 10.1109/TPAMI.2018.2858826]

20 Zhang SF, Wen LY, Bian X, *et al*. Single-shot refinement neural network for object detection. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 4203–4212.

21 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.

22 王伟锋, 金杰, 陈景明. 基于感受野的快速小目标检测算法. 激光与光电子学进展, 2020, 57(2): 021501.

23 Szegedy C, Ioffe S, Vanhoucke V, *et al*. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, CA, USA. 2017. 4278–4284.

24 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017.