计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(9):288−294 [doi: 10.15888/j.cnki.csa.008077]

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

基于改进 Faster-RCNN 的 IT 设备图像定位与识别①

张　晓[1], 丁云峰[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所 系统与软件事业部, 沈阳 110168)
通讯作者: 张　晓, E-mail: 1033610689@qq.com

摘　要: 本文根据国家电网 IT 设备识别的具体应用场景的特点, 通过改进 Faster-RCNN 实现设备的精确识别定位, 进而提高了电网数据中心管理的效率. 文章主要在注意力机制、初始锚框调整以及锚框融合等方面进行改进. 通过与常见图像算法的横向比较发现改进后的模型在收敛速度上提高了 30%, 精度上提高了 1%.

关键词: 图像识别; 注意力机制; 卷积网络; 图像定位

引用格式: 张晓,丁云峰.基于改进 Faster-RCNN 的 IT 设备图像定位与识别.计算机系统应用,2021,30(9):288–294. http://www.c-s-a.org.cn/1003-3254/8077.html

Identification and Location of IT Equipment Based on Improved Faster-RCNN

ZHANG Xiao[1], DING Yun-Feng[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](System and Software Division, Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shengyang 110168, China)

Abstract: In this study, according to the characteristics of specific application scenarios for the IT equipment identification of State Grid, accurate identification and positioning of the equipment is realized with improved Faster-RCNN, thereby improving the management efficiency of the grid data center. The algorithm is improved mainly in terms of the attention mechanism, the initial anchor box adjustment and the anchor box fusion. The comparison with common image algorithms shows that the improved model has the convergence speed and the accuracy increased by 30% and 1%, respectively.

Key words: image identification; attention mechanism; convolutional network; image positioning

国家电网数据中心拥有数量庞大的设备, 提高管理这些设备的效率和降低相关管理支出具有重要意义. 数据中心的 IT 设备 (以下简称设备) 可分为服务器、交换机、路由器等, 且这些设备安装于机柜中. 设备管理的内容有, 详细记录机柜中设备的安装位置和设备型号. 传统的管理方法是人工抄录和定期人工巡视复核, 这种方式不但费时费力而且容易出错, 因此找到一种简单高效的管理方法具有重要的现实意义.

近年来, 人们在设备识别上进行了许多尝试. 传统的方法主要基于设备的电学特性, 不同设备都有自己特定的阻抗、感抗、对地电阻等电学特性, 一方面可以通过电学指标计算一个特征码, 然后通过实时采集相关特征计算实时特征码, 通过比对得到设备的种类; 另一方面可以在把采集的电学特性数据当作特征结合传统机器学习的算法进行分类. 虽然这两种算法在一定程度上可以识别出设备类型, 但是需要安装电学特

① 收稿时间: 2020-12-08; 修改时间: 2021-01-08; 采用时间: 2021-01-20; csa 在线出版时间: 2021-09-02

征的采集装置, 并且类内设备的型号识别也比较困难. 任晓欣等人[1] 提出了图像与支持向量机结合的算法, 但图像与支持向量机结合时重要的一步是图像的预处理, 图像预处理的结果与最终的分类效果息息相关. 并且该方案仅支持图像中单一设备的识别, 对类内不同型号的识别效果比较差. 王祎瑞[2] 提出了基于 YOLOv2 的 TSENet 模型. 该模型主要对 YOLOv2 的输出层进行调整, 通过 fine-turning 的方法对模型输出层进行训练. 该模型可以进行多设备检测, 且算法执行速度快, 适合视频流的处理. 但是设备类型识别准确率比较低, 达不到进行设备识别要求的精度.

通过图像直接识别设备的类型和位置在逻辑上是比较自然的方案. 但是在实际中通过图像进行设备识别的主要困难有: 类间差别大类内差别小, 要区分同一类设备的不同的型号比较困难; 设备颜色与机柜颜色相近, 设备轮廓识别困难; 机柜中设备安放不规则, 定位实际位置困难[3]. 当下的研究中, 深度神经网络在设备管理领域的应用比较少, 结合其在其它相近领域的研究成果, 文本将深度卷积神经网络应用到设备上, 基于文献 [4] 中的提出 Faster-RCNN 和文献 [5] 提出的通道注意块提出了基于注意力机制的 Faster-RCNN 网络. 利用国家电网某数据中心的机柜图像作为数据集进行模型的训练和测试. 并在内部的不同网络结构和外部的几种最有效的算法间进行了比较与分析.

1 神经网络结构设计

1.1 改进模型分析

当前在图像多物体识别领域已经取得了许多成果, 相关研究可以分为 RCNN 体系和 YOLO 体系. RCNN 体系中主要有 RCNN, Fast-RCNN, Mask-RCNN, SSD 以及 Faster-RCNN. 从字面意思上就可以得到 Faster-RCNN 是从 Fast-RCNN 改进来的, 而 Mask-RCNN 则是在 Faster-RCNN 基础上添加 FCN 层进行像素级别上的卷积, 该算法在语义分割上取得了非常好的成绩, SSD 算法[6] 提出一种新的锚框生成机制, 不再需要 RPN 结构来提取生成锚框, 简化了网络的模型. YOLO 体系最先有谷歌团队提出来后续经过改进产生了 YOLOv1, YOLOv2 等算法模型. 本文选用 Faster-RCNN 作为基础模型是因为虽然 Faster-RCNN 不如 YOLOv2 的运算速度快, 但是模型的精确度高[7]. 而 Mask-RCNN 在训练时需要进行像素级标注构建训练样本困难, 并且该结构分块明确在模型改进上相对容易.

综合以上考虑使用 Faster-RCNN 作为网络的基础模型, 该模型首先经过预训练的 VGG-16 网络提取特征图, 然后分两个分支一个是用来初步生成锚框的 RPN 结构, 另一个与 RPN 的结果通过 ROIPooling 作用生成新的特征图然后进行物体分类和锚框的重新调整. 虽然该结构在很多图像分类任务上取得了比较优秀的成绩, 但是在 IT 设备识别中, 设备的型号一般标注在左上角或者右上角的一个小区域中, 我们希望在进行分类时除了关注设备的轮廓外, 还要更多的关注于设备类型标签区域. 除此之外, 设备在机柜中的位置相比自然图片比较规则, 可以根据实际情况减少锚框的初始生成数量, 可以在一定程度上减少 RPN 模块的训练时间. 且该模型中 ROIPooling 中的线性插值法进行融合时比较暴力, 与原图的映射关系也不明确, 机柜中的设备之间排布比较紧密所以我们采用了更好的双线性插值来做模型融合.

本文提出了基于注意力机制 Faster-RCNN 网络, 模型结构如图 1 所示, 图中加粗的框为改动的地方. 图中加粗的框主要有 3 处, 分别对应标号①~③. 接下来详细阐述改进后的网络结构.

1.2 注意力机制

图 1 中标号①的部分, 主要为神经网络加入注意力机制, 即为特征图增加注意力特征.

注意力机制现在已经被广泛应用在图像领域中. 其在图像的领域的应用主要分为 3 个方面. 一种是 Mnih 等人提出的基于空间域[8] 的注意力模型, 通过一个空间转换器将图片中的空间域信息做对应的空间变换, 然后将图片的关键信息提取出来. 但是这种方式是对所有的通道信息做统一变换, 但是在卷积之后每一通道表达的特征信息就被弱化了. 于是文献 [9] 提出了基于通道域的注意力模型, 通过挤压、激励和注意 3 个模块生成每个通道的注意力分布. 第 3 种方案就是综合空间域和时间域的混合域了, 由 Wang 等人提出[10]. 作者受残差网络的设计思想的启发提出了该注意力机制, 创新点就在于不仅把转换之前的特征向量作为下层的输入, 也把转换之前的特征向量作为下一层的输入, 可以获得更好的注意力特征. 前文提到识别的难点之一是 IT 设备与机柜的颜色相近, 也就是前景色和背景色相近. 并且设备标签一般占整个图的比例较小, 通过注意力机制可以提高标签的识别效果. 因此我们通过对通道设置不同的权重来进一步提高通道的表达能

力. 本文使用了基于通道的注意力机制.

具体的设计细节如图 2 所示. 通过实验, 综合运行时间和精度分别在 VGG-16 的 Pool2、Pool4 和 Pool5 之后添加注意力提取模块.

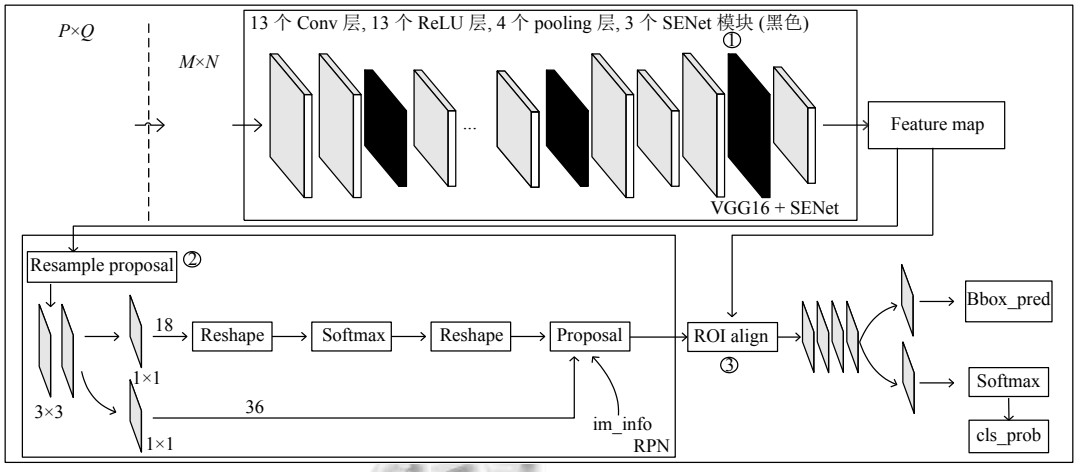

图 1 改进的 Faster-RCNN 结构

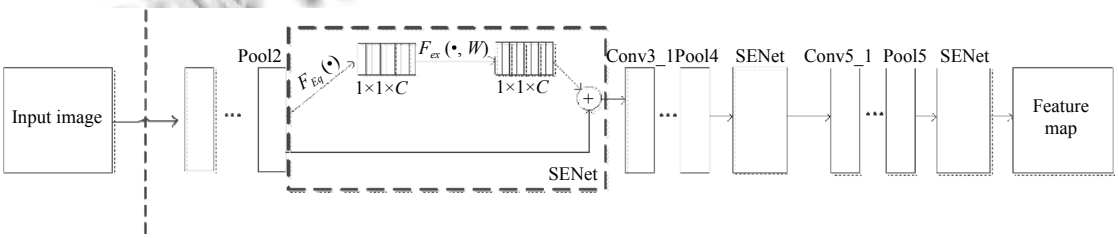

图 2 基于通道域的注意力特征结构

图 2 中虚线框的部分就是一个注意力提取模块, 我们希望通过注意力提取模块学习每个通道的权重, 从而产生通道的注意力. 注意力机制模块分为 3 个部分: 挤压、激励和注意. 具体来说, 首先将经过卷积和池化生成的特征图输入如式 (1) 所示的挤压函数:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{1}$$

其中, $u_c(i,j)$表示输入注意模块的特征图u的第c个通道的(i,j)位置的像素值. 很明显挤压函数的作用是把每个通道内所有的特征值相加再取平均, 也就是计算了一个全局平均值. 然后将计算得到的z_c输入激励函数$F_{ex}(z,W)$:

$$s = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma(W_2\delta(W_1 z)) \tag{2}$$

其中, δ表示 ReLU 函数, σ表示 Sigmoid 函数, 其中 $W_1 \in \Re^{\frac{C}{r}\times C}$, $W_2 \in \Re^{\frac{C}{r}\times C}$, 参数$r$为维度缩放比率, 为了在不同的网络中限制训练的复杂度. 通过激励函数可以充分使用挤压函数生成的信息, 尽可能捕捉通道间的依赖信息[11,12]. 通过学习这两个权重得到s, 并将其输入尺度函数F_{scale}中:

$$\tilde{x}_c = F_{\text{scale}}(u_c, s_c) = s_c \cdot u_c \tag{3}$$

其中, $u_c \in \Re^{H\times W}$, $\tilde{x}_c$表示通道c的注意力权重.

综上分析, 通道注意力模块的详细设计如图 3 所示.

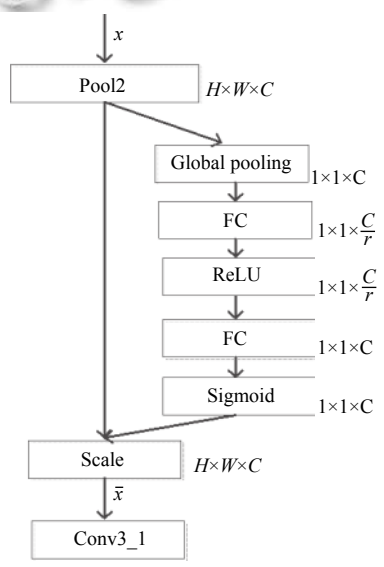

图 3 通道域注意力机制详细结构设计

1.3 初始锚框调整

Faster-RCNN 的 RPN 中每个特征点生成 3 不组不同大小的初始锚框[13,14], 每组有长宽比例不同的 3 个锚框. 但本文中 IT 设备的形状比较规则, 所以本文对每个特征点生成 2 组锚框, 每组 3 个锚框比例分别是, 16:1、4:1 和 1:4. 6 个锚框虽然不能将整张图片完全覆盖但是可以覆盖掉所有的放置在机柜上的 IT 设备. 本文的锚框生成伪代码为:

算法 1. 锚框生成算法

```
//vgg-16
let x_base1 ≜ 1, y_base1 ≜ 1, x_base2 ≜ 16, y_base2 ≜ 16
for scale in scales :
    w=x_base2−x_base1+1
    h=y_base2−y_base1+1
    x_center=x_base1+0.5∗(w−1)
    y_center=y_base1+0.5∗(h−1)
    for ratio in ratios:
        ws_mid=sqrt(size/ratio)
        hs_mid=ws_mid∗ratio
        ws=ws_mid∗scale
        hs=hs_mid∗scale
return ws,hs,x_center,y_center
```

锚框的生成图示如图 4 所示.

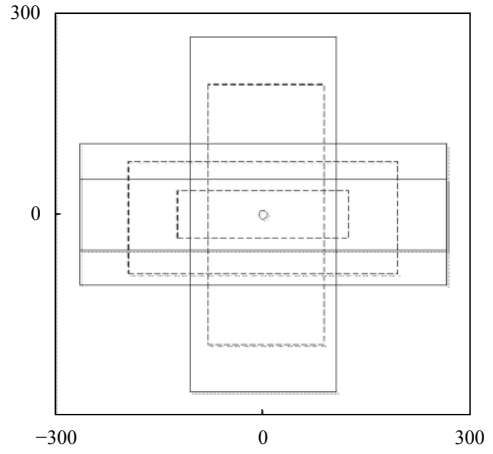

图 4 锚框生成示例

1.4 ROIAlign

Faster-RCNN 中使用 ROIPooling 来融合 RPN 输出和特征图[15–19], 使用最近邻插值直接对特征图进行 pooling, 在恢复尺寸阶段对于缩放后坐标不能刚好为整数的情况, 采用了粗暴的舍去小数, 相当于选取离目标点最近的点, 损失一定的空间精度.

而 ROIAlign 将最近邻插值换为双线性插值, 使得即使缩放后坐标不能刚好为整数, 也能通过插值得到浮点数处的值处理得到 pooling 后的值, 保证了空间精度.

2 模型训练

2.1 VGG-16+SENet 训练

本文加入了 SENet 模块, 我们重新训练了 VGG-16 的特征提取网络. 结合文献 [9] 中提出的 SENet 训练的方法, 我们在 ImageNet2012 数据集上进行了 4 轮测试, 目的是找到在 VGG-16 添加 SENet 的合理方案. VGG16 网络虽然结构简单但在图像分类任务中取得了很好的成绩, 因此现在的许多深度网络将 VGG16 作为复杂图像任务中的特征提取的预训练模型. 该网络的主要特点是通过 13 层卷积的分布进行特征的有效提取, 以每一个池化层的结束为一个特征提取模块. 将 SENet 模块放置在池化层之后是因为每经过池化后一次高层次特征提取才算完成, 这样就在保留了原有特征提取的优良特性基础上发挥 SENet 模块对关键特征增强的优势. VGG16 中共有 5 层池化, 使用 3 个 SENet 模块进行注意力生成. 具体放置方案如表 1 所示.

表 1 SENet 模块放置方案对比

方案	Pool1	Pool2	Pool3	Pool4	Pool5	训练错误率(%)	测试错误率(%)
1	√	—	√	—	√	22.43	23.54
2	√	—	—	√	√	23.59	24.69
3	—	√	√	√	—	24.71	24.77
4	—	—	√	√	√	24.04	24.96

使用 Image2012 数据集的部分数据对该模型进行预训练和注意力放置方案的评估. 从数据集中随机抽取 100 万张有标记的数据, 其中 30 万张作为测试数据, 另外 70 万作为训练数据. 共进行 15 个 epoch, 每个 batch 为 500 张图片. 训练的过程采用 fine turning 的思想, 通过冻结某些层可以加快训练速度, 比如方案 4 中分别在 Pool3、Pool4 和 Pool5 中加入了注意力模块, 在训练时冻结 Pool4 之前的所有层, 只更新 Pool4 之后的层的权重.

模型训练过程中的主要困难是经过修改后的 VGG16 添加了注意力模块, 打破了原来神经网络反向传播算法的递归求解特性. 所以在反向传播到注意力模块时要单独来求解该模块的参数更新方法, 综合式 (1)~式 (3) 得到注意力模块的反向传播的权重公式为:

$$\begin{cases} w_2^l = w_2^l + \eta \cdot \Delta^{l+1} \cdot \sigma' \cdot \delta(w_1^l) \\ w_1^l = w_1^l + \eta \cdot \Delta^{l+1} \cdot \sigma' \cdot z_c \end{cases} \tag{4}$$

其中, Δ^{l+1}表示通过上一层, 即卷积层传递过来的误差. σ'表示式 (2) 中的 Sigmoid 函数的导数, δ表示式 (2) 中的 ReLU 函数, η表示学习率, 训练时学习率设为10^{-4}. 通过式 (4) 来更新权重w_1, w_2, 并且更新时先更新w_1再更新w_2.

分析表 1 在测试和训练阶段的错误率发现, 第一种方案的效果是比较好的. 因此本文采用第 1 种方案, 然后将训练好的网络最后两层移除, 由此得到了特征图生成模块的网络模型.

2.2 其它模块的训练

本文的图片数据来自于国家电网某大数据中心, 数据中心中一共有 33 中不同型号的设备. 所以根据实际情况, 将图 1 中的 RPN 模块的分类层和最终的分类层修改为 33 类输出. 网络中除了与分类层相连的全连接层其它层都处在冻结状态不进行参数的更新. 模型的输入首先将图片输入特征提取网络, 经过特征提取后将结果输入 RPN 网络中进行训练.

3 实验分析

本文选用来国家电网某数据中心的图片数据, 一共 2 万张. 其中 1.7 万张作为训练集, 0.3 万张作为测试集, 数据集使用 LabelImg 工具自行标注, 标注后的数据集如图 5 所示.

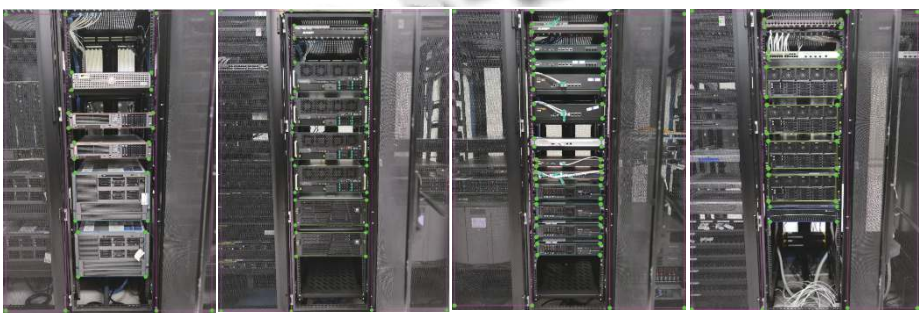

图 5 标注后部分数据集样例

3.1 模型比较

模型训练时一共进行 1000 次迭代, 一次放入图片 150 张. 首先经过预训练特征提取模块得到特征提取图. 将提取到的特征输入 RPN 模块和 ROIAlign 模块进行后续的训练工作. 为了更好的检验修改后的网络模型的性能, 首先进行了相同数据规模和输入图片的两次训练. 分别是原始的 Faster-RCNN 网络和经过修改后带有注意力机制的改进 Faster-RCNN 网络, 并得到如图 5 的训练结果图. 为了更清楚的展示模型性能使用了平方损失作为考察项, 平方损失是个简单高效的损失函数横轴表示进行迭代的轮次. 公式如下:

$$L = \frac{1}{2}\sum_{i-1}^{N}(\tilde{y}_i - y_i)^2 \tag{5}$$

观察图片发现, 改进的 Faster-RCNN 网络模型虽然在前 200 次迭代时波动较大, 但是在 200 轮之后该模型的收敛速度要比原模型快 30%, 并且最终收敛的损失要比原模型低 10%.

为了更好的分析模型性能, 本文选用在图片物体检测领域最常用的 3 个算法 SSD, Mask-RCNN, YOLOv2, 这 3 个算法在前文进行 1.1 节进行了概括介绍, 这里不再阐述. 因为这些算法的输出与本文不相符所以使用前文的 1.7 万张图片对它们进行输出结构修改后的训练. 并选取了 4 张在训练集和测试集都没有的图片的对这些模型进行测试, 保存预测输出的损失值如表 2 所示和输出的预测结果图, 如图 6 所示. 结合表 2 和图 6 所示, 无论在视觉上还是具体的数值指标上, 本算法均取得了优于其他算法的结果.

表 2 不同算法损失比较

验证图像	SSD	Mask-RCNN	YOLOv2	本文算法
Test_1	0.053	0.059	0.061	**0.053**
Test_2	0.074	0.077	0.072	**0.071**
Test_3	0.068	0.066	0.069	**0.065**
Test_4	0.066	0.070	0.074	**0.068**
平均损失	0.065	0.068	0.069	**0.064**

3.2 RPN 网络作用比较

在修改的模型中减少了锚框的生成数量, 在改进的 Faster-RCNN 的训练过程中进行修改和未修改情况

的对比, 对比结果如图 7 所示.

图 7 中横坐标表示模块的收敛所用的时间, 纵坐标表示该模块的平方损失, 该模块的损失代表锚框的位置损失, 红色的带五角星的曲线表示修改后的模块, 另一条则代表未修改的原来的模块. 观察图像发现修改后模块的收敛时间相比原模块缩短 30%, 特别发现通过这个操作, 使得模型的损失降低了大约 1%, 这意味着可以以更高的精度输出 IT 设备的位置. 分析之后发现, 修改之后的锚框在比例和尺寸上更加符合 IT 设备的轮廓比例和大小. 并且少但是更精确的锚框数量, 减少了 RPN 模块中回归模型的运算量和错误率.

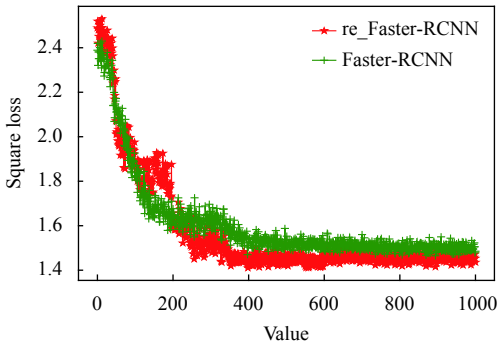

图 6 原始模型与改进模型的训练平方损失

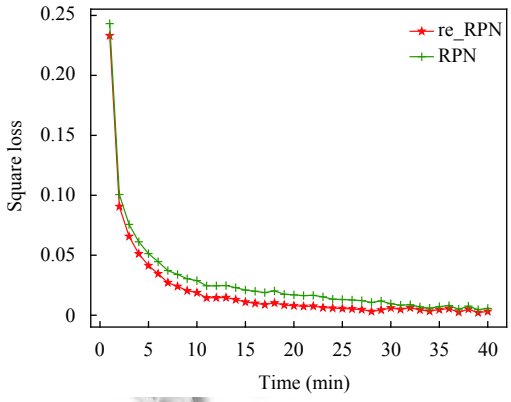

图 7 原始 RPN 与改进 RPN 模块的时间对比

4 结论与展望

本文结合实际应用场景, 综合考量了当前存在的比较优秀的算法. 并针对这些算法的特点进行实际测试. 最终选定了在图像物体检测领域取得了优秀成果的 Faster-RCNN 作为最终基础算法模型, 并根据实际情况对结构进行调整. 首先加入了基于通道的注意力机制, 得到了不同通道权重分布特征图. 通过调整锚框提升了时间效率和初步预测的精度. 特别是 RoiAglin 机制的加入, 可以准确的将 RPN 输出的锚框的坐标与特征图对应, 为后面的精确分类提供了高质量的特征.

但是本文的网络结构比较复杂, 虽然精确度高但是模型的耗时也高. 还需要在模型的结构等方面不断的探索调整, 寻找一个复杂度和耗时比较均衡的模型.

参考文献

1 任晓欣, 胡姗, 燕达, 等. 基于实测的家用电器用电模型研究. 建筑科学, 2012, 28(S2): 223–231.
2 王祎瑞. 基于深度卷积神经网络的变电站设备识别 [硕士学位论文]. 沈阳: 沈阳农业大学, 2018.
3 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6517–6525.
4 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]
5 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 7132–7141.
6 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 2980–2988.
7 Jaderberg M, Simonyan K, Zisserman A, *et al*. Spatial transformer networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2015. 2017–2025.
8 Mnih V, Heess N, Graves A, *et al*. Recurrent models of visual attention. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2014. 2204–2212.
9 Hu J, Shen L, Albanie S, *et al*.. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]
10 Wang F, Jiang MQ, Qian C, *et al*. Residual attention network for image classification. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6450–6458.
11 Krizhevsky A, Sutskever I, Hinton GE. ImageNet

classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, CA, USA. 2012. 1097–1105.

12 Zhang QS, Wu YN, Zhu SC. Interpretable convolutional neural networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 8827–8836.

13 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA. 2015.

14 Trinh DH, Luong M, Rocchisani JM, *et al*. An optimal weight model for single image super-resolution. Proceedings of 2012 International Conference on Digital Image Computing Techniques and Applications. Fremantle, WA, Australia. 2012. 1–8.

15 Zitnick CL, Dollár P. Edge boxes: Locating object proposals from edges. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 391–405.

16 Wang WG, Shen JB. Deep visual attention prediction. IEEE Transactions on Image Processing, 2018, 27(5): 2368–2378. [doi: 10.1109/TIP.2017.2787612]

17 Xu K, Ba JL, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France. 2015. 2048–2057.

18 Smith LN, Topin N. Deep convolutional neural network design patterns. arXiv: 1611.00847. 2017.

19 Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]