计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(6):97−103 [doi: 10.15888/j.cnki.csa.007422]

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

基于嵌入式深度学习的电力设备红外热成像故障识别①

王彦博[1], 陈培峰[1], 徐　亮[2], 张合宝[3], 房　凯[4]

[1](国网山东省电力公司检修公司, 济南 250100)
[2](北京科技大学 计算机与通信工程学院, 北京 100081)
[3](积成电子股份有限公司, 济南 250100)
[4](中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 房　凯, E-mail: k2322358928@163.com

摘　要: 随着大型图像集的出现以及计算机硬件尤其是 GPU 的快速发展, 在有限计算资源的嵌入式设备上部署卷积神经网络 (CNN) 模型成为具有挑战性的问题. 电力设备过热故障可以通过采集的红外热成像进行识别. 由于红外辐射在空气中传播衰落, 红外测温结果低于实际温度值. 本文提出一种基于嵌入式设备的高效卷积神经网络用于电力设备热故障检测, 将 SSD 算法中的骨干网络替换为 MobileNet, 同时 Batch Normalization 与前一卷积层合并, 以减少模型参数、提升推理速度、使之能够在轻量级计算平台上运行. 针对红外辐射在空气中传播损失的问题, 提出一种基于 BP 神经网络的红外测温修正单元. 基于上述创新设计了一种电力设备热故障检测系统, 实验以及现场应用表明, 该方法具有较高的准确性以及推理速度.

关键词: 深度学习; 红外热成像; 轻量级; 故障检测; 电力设备

引用格式: 王彦博,陈培峰,徐亮,张合宝,房凯.基于嵌入式深度学习的电力设备红外热成像故障识别.计算机系统应用,2020,29(6):97–103. http://www.c-s-a.org.cn/1003-3254/7422.html

Fault Recognition of Power Equipment in Infrared Thermal Images Based on Deep Learning with Embedded Devices

WANG Yan-Bo[1], CHEN Pei-Feng[1], XU Liang[2], ZHANG He-Bao[3], FANG Kai[4]

[1](Overhaul Company, State Grid Shandong Electric Power Company, Jinan 250100, China)
[2](School of Computer and Communication Engineering, Beijing University of Science and Technology, Beijing 100081, China)
[3](Jicheng Electronics Co. Ltd., Jinan 250100, China)
[4](School of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: With the emerging large image sets and the rapid development of computer hardware, especially GPU, the deployment of Convolutional Neural Network (CNN) model on embedded devices with limited computing resources becomes a challenging problem. Overheating of power equipment can be identified from infrared thermal images. Because of the fading of infrared radiation in the air, the result of infrared temperature measurement is lower than the actual value. In this study, an efficient CNN based on embedded devices is proposed for thermal fault detection of power equipment. The backbone network of SSD algorithm is replaced by MobileNet. At the same time, batch normalization is combined with the previous volume to reduce model parameters, improve reasoning speed, and make it run on a lightweight computing platform. To solve the problem of infrared radiation loss in the air, an infrared temperature correction unit based on BP neural network is proposed. Based on the above innovation, a thermal fault detection system for power equipment is designed. Experiments and field applications show that the proposed method has high accuracy and reasoning speed.

Key words: deep learning; infrared thermal imaging; lightweight; fault detection; electric power equipment

① 收稿时间: 2019-10-19; 修改时间: 2019-11-15; 采用时间: 2019-11-21; csa 在线出版时间: 2020-06-10

1 引言

电力系统中的变电站与输电线路是联系发电厂与用户的重要枢纽, 其安全稳定运行至关重要, 一旦电力设备发生故障, 电力系统的安全性以及供电的稳定性将受到极大影响[1]. 电力设备长期受气候因素等外部环境影响极易发生故障, 因此需要对电力设备进行定期巡检维护来保证供电系统正常运行. 据相关资料统计, 高达 90% 的电力系统事故是由电力设备故障引起的, 其超过 50% 的故障设备会在早期阶段出现异常的发热症状[2]. 红外测温的原理是探测器探测并接收被测目标发射出的红外辐射能量, 将接收到的红外辐射能量转换成与之相对应的电信号, 再经过专门的电信号处理系统获得物体表面的温度分布状态. 电力设备的热故障由电力设备类别、发热部位以及发热程度等因素决定, 其温度分布形式也不同. 因此, 红外技术十分适用于电力设备的热故障检测. 分析电力设备表面的温度分布信息, 能够对发现电力设备中潜在的隐患和故障, 并对故障的严重程度做出定量的判断[3,4].

目前, 电力系统中最主要的巡检形式是人工巡检, 人工现场诊断或者采集信息供后续分析. 人工巡检工作量大, 管理成本高, 需要对技术人员进行技能培训, 信息采集与故障分析都需要人工来完成. 而国内电力系统分布广泛且部分地区所处环境恶劣, 增加了巡检成本与难度, 人工巡检变得异常复杂. 如果巡检不及时, 一旦电力设备出现故障, 将会造成严重事故. 如今, 卷积神经网络与传感器于信息技术的融合得到广泛的研究, 并应用于电力系统巡检, 在一定程度上减少了巡检成本和难度, 为及时发现安全隐患并且排除故障和处理突发状况奠定了基础[5].

卷积操作具备强大的特征提取能力, 与全连接层相比具有更少的参数优势, 应用在图像数据中具有得天独厚的优势, 卷积神经网络在计算机视觉领域占据了毋庸置疑的地位. 然而, 嵌入式设备的硬件资源限制使得卷积神经网络难以在小型设备上进行部署, 所以卷积神经网络结构优化是本文的重要目标. 卷积神经网络大部分计算都集中在卷积操作上, 因此减少网络模型的复杂度、减少模型的计算量和参数量的关键是设计高效的卷积结构, 在精度损失可接受的范围内, 大幅提高网络速度.

由于被测目标表面发射的红外辐射强度在空气传播中将会衰减, 被测目标的红外测温结果往往低于被测目标的实际温度. 因此, 本文采用 BP 神经网络, 对被测目标的红外测温结果进行修正.

2 相关工作

电力设备热故障检测的首要步骤就是对采集到的红外热成像进行分析, 检测出图像中电力设备. 电力设备检测主要分为两个任务: 一是定位任务, 从图像检测出电力设备并给出该设备的位置信息; 二是分类任务, 判断出每个电力设备的具体类别.

由于视角、遮挡等因素使目标物体发生形变, 使得电力设备检测成为具有一定难度的任务[6]. 文献[7]提出了一种基于 SIFT (Scale-Invariant Feature Transform)[8] 和 OTSU[9]特征匹配的变电站视频电力塔倾角检测方法, 通过匹配特征点来计算电力塔的倾角;文献[10]为了在模板图像和变电站的监控图像之间进行 SIFT 进行特征匹配, 采用 RANSAC[11]用来消除图像中的不匹配状况, 其中模板图像的边缘通过 OTSU 分割得到. 传统的电力设备检测算法依赖于人工提取每个电力设备特征, 虽然获得了较好的发展, 但由于人工特征对不同图像数据适用性较差, 且提取的特征表达能力不足, 容易导致最终分类错误和漏检.

深度学习能够克服传统方法的缺陷, 通过逐层的特征变换, 将蕴含于数据中的信息映射到新的特征空间上, 让计算机自动的学习到高层的特征[12], 随着数据量的增加而增强学习的效果, 获得更高的匹配精度. R-CNN (Region CNN)[13]可以说最先使用深度学习进行目标检测的算法, 利用海量数据来训练卷积神经网络模型来提取特征. SSP-Net (Spatial Pyramid Pooling Network)[14]设计空间金字塔采样层 (Spatial Pyramid Pooling, SSP) 实现网络模型输入任意图像大小. 针对 R-CNN 对每一个候选区域都要独自使用卷积神经网络提取特征的问题, Fast R-CNN[15]设计共享卷积层提高了 R-CNN 速度. Faster R-CNN[16]第一次提出 RPN (Region Proposal Network) 网络, 直接使用卷积神经网络产生候选区域. RPN 网络提取少量的高质量预选区域, 具有很高的召回率. YOLO (You Only Look Once)[17] 对目标检测使用回归方式, 采用单个卷积神经网络模型来实现端对端 (End-to-End) 的目标检测.

深度神经网络复杂、训练大、计算量、参数量大, 模型的部署需要非常强大的硬件设备, 嵌入式设备由于硬件限制只能部署浅层神经网络, 目标检测精度受

到限制. 要实现工程化, 需要研究深度网络模型的压缩、降低内存占用、降低功耗、减少计算量和参数量. 剪枝 (pruning) 应用于神经网络中来移除一些不重要的权重, 能够有效加快网络的速度, 提升网络泛化性能[18–24]. 参数量化 (quantization) 就是从权重中归纳出若干个能代表某一类权重的具体数值, 这一类代表被存储在码本 (codebook) 中, 原本的权重矩阵只需要记录各自代表的索引即可, 极大地降低了存储开销[25,26]. 卷积神经网络计算大部分集中在卷积操作, 因此压缩模型、提高效率需要设计新型的网络结构, 如 MobileNet[27]、ShuffleNet[28].

3 电力设备检测

目标检测是深度学习的一个重要研究领域, 通过获取目标信息, 提取目标特征来进行训练学习、特征分类等. 本文将在 SSD 的基础上实现电力设备检测, 采用谷歌提出的适用于嵌入式以及移动端的高效轻量级卷积神经网络 MobileNet 替代 SSD 中的 VGG-16 网络, 与 SSD 算法相比, 具有更好的环境适应性、鲁棒性以及更高的精度.

3.1 基于嵌入式平台的电力设备检测

SSD 是典型的基于深度学习的目标检测算法, 与 R-CNN 系列目标检测算法相比, SSD 取消中间的候选框和像素特征的重采样过程, 保证速度的同时保证了检测精度, SSD 输出一系列离散化的候选框, 候选框生成在不同层上的特征图且长宽比不同, 经过卷积神经网络的前馈操作, SSD 生成一系列固定大小的候选框, 使用小卷积 Filter 来预测候选框位置中的目标类别和偏移即候选框中包含目标种类的概率, 最后通过极大值抑制方法得到最终的预测结果.

但是 SSD[29]以 VGG-16[30]作为特征提取网络, 需要消耗大量的计算资源, 这些网络模型通常部署在 GPU 上, 对于硬件要求极高, 在嵌入式平台上难以运行如此巨大的网络模型, 将极大影响电力设备检测效率. 为了使 SSD 适用于嵌入式设备, 本文对 SSD 进行优化改进. 我们使用 MobileNet 替代 SSD 中的 VGG-16 网络. MobileNet 采用深度可分离卷积 (depthwise separable convolutions) 来替代常规的卷积层, 深度可分离卷积将标准卷积分解成为深度卷积和逐点卷积, 当输入 feature map 为 $m\times n\times 16$, 想输出 32 通道, 那么卷积核应为 $16\times 3\times 3\times 32$, 则可以分解为深度卷积: $16\times 3\times 3$, 得到的是 16 通道的特征图谱, 点卷积: $16\times 1\times 1\times 32$. 如果用标准卷积, 计算量为: $m\times n\times 16\times 3\times 3\times 32=m\times n\times 4608$, 用深度可分解卷积之后的计算量为: $m\times n\times 16\times 3\times 3+m\times n\times 16\times 1\times 1\times 32=m\times n\times 656$, 减少了卷积神经网络的计算量以及参数量, 提高网络运行效率. MobileNet-SSD 与 SSD 网络结构如图 1 所示.

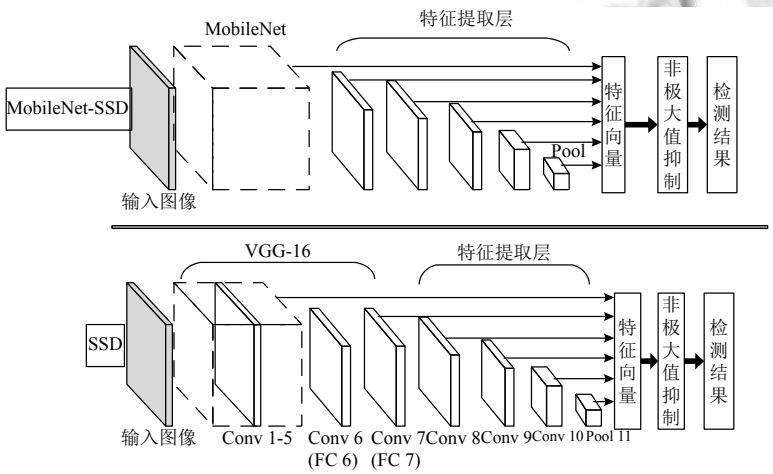

图 1 MobileNet-SSD 与 SSD 网络结构对比

3.2 合并 Batch Normalization 层

在训练深层卷积神经网络时, 使用 Batch Normalization 层能够加速训练速度, 提升网络收敛速度. 在卷积神经网络中, Batch Normalization 层一般放置在卷积层或者全连接层之后, Batch Normalization 层将数据进行归一化后, 能够有效地解决梯度消失和梯度爆炸问

题, 并加速训练拟合速度. Batch Normalization 层在深度卷积神经网络的训练阶段起到一定的积极作用, 但在网络模型的部署阶段, 在模型预测时多了一层计算, 将影响模型整体的运算速度, 并且增加显存与内存的占用空间. 因此, 在网络模型的部署阶段, 需要将 Batch Normalization 层合并到卷积层中, 以提高网络模型的速度.

假设, 每一层的输入均表示为 X, W 为卷积权重, b 为卷积偏置, 先进行卷积运算, 卷积层的运算公式为:

$$W\times X+b \tag{1}$$

在卷积运算后, 进行 Batch Normalization 层运算. Batch Normalization 层进行两个操作, 第一个是归一化, 归一化运算公式如下:

$$\frac{X-\mu}{\sqrt{\sigma^2+\varepsilon}} \tag{2}$$

其中, μ 为均值, σ 为方差, ε 为一个较小数, 防止分母为零.

Batch Normalization 层第二个操作是缩放:

$$\gamma X+\beta \tag{3}$$

其中, γ 为缩放因子, β 为偏置.

卷积层和 Batch Normalization 层合并后, 得到:

$$\gamma\times\frac{(W_{\text{old}}\times X+b_{\text{old}})-\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta \tag{4}$$

得到新的卷积权重如下:

$$W_{\text{new}}=\frac{\gamma}{\sqrt{\sigma^2+\varepsilon}}\times W_{\text{old}} \tag{5}$$

得到新的偏置如下:

$$b_{\text{new}}=\frac{\gamma}{\sqrt{\sigma^2+\varepsilon}}(b_{\text{old}}-\mu)+\beta \tag{6}$$

4 红外测温结果修正

一般, 物体表面发射出来的红外辐射强度在空气传播中会衰减, 因此, 被测目标的红外测温结果往往低于被测目标的实际温度, 距离越远, 实际差值越大. 所以需要修正被测目标的红外测温结果. BP 神经网络算法广泛应用在各种场景, 而在探究红外测温的影响参数时, BP 神经网络算法相比线性插值法和多元线性回归法, 非线性映射能力强适应性强, 精确度高.

温度修正模块将采用 BP 神经网络来对红外测温结果进行温度修正. 如图 2 所示, 输入层的输入为被测目标的红外测温结果和该目标的测量距离, 将这两类数据输入到 BP 神经网络中, 得到最终修正过的温度.

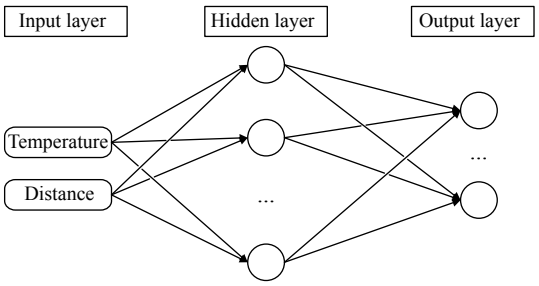

图 2 神经网络结构

温度修正模块设计的 BP 神经网络的隐藏层含有 α 个节点, 输入层含有 β 个节点, 输出层含有 γ 个节点. 输出层与隐藏层间神经元的神经元的传递函数为线性传递函数. 假设 x_i 为神经元的状态, y_i 为输出, 神经元的输入状态 x_i 和输出 y_i 的关系为线性变化, 如式 (1) 所示:

$$y_i=f(x_i) \tag{7}$$

假设 $D_{1,2,\cdots,\gamma}$ 分别为温度修正模块 BP 神经网络的输入, 则神经网络对应的输出:

$$E_i=D_i,\ i=1,2,\cdots,\gamma \tag{8}$$

假设, L_{1j}^2, L_{ji}, 为连接权重, M^2, M_i 为偏置值, 则隐藏层中第 j 个节点的输入:

$$F_i=L_{j1}\times E_1+L_{j2}\times E_2+\cdots+L_{j\gamma}\times E_\gamma+M_j \tag{9}$$

于是, 得到隐藏层中的第 j 个节点的输出, 见式 (10):

$$H_j=f(x_j),j=1,2,\cdots,\alpha \tag{10}$$

由此, 将得到输出层中第 j 个节点的输入:

$$k=\sum_{j=1}^{\alpha}L_{1j}^2\times H_j+M^2 \tag{11}$$

最后, 得到输出层中第 j 个节点的输出:

$$Y_j=f(x_j)=f(k)=k \tag{12}$$

5 电力设备热故障检测系统

本文为山东电力设计了电力设备热故障检测方法, 将红外技术与深度学习相结合, 把视频流的读取、深度学习的电力设备检测、BP 神经网络的温度修正以及数据的可视化融合. 基于嵌入式深度学习的电力设备热故障检测架构如图 3, 该架构主要有 3 个层次.

底层为数据的读取, 红外热像仪通过以太网输出 MPEG-4 格式的视频流, 并将红外热像仪视频流解码成帧并传送到下一层.

中间层是数据处理层, 主要从上一层获取的红外

热像仪视频流进行热故障诊断. 为数据处理层中电力设备检测算法的识别搭建了相对应的深度学习框架, 部署的相关网络模型是已经训练完成的模型, 这一层只进行检测任务, 不进行网络模型的训练任务. 这一层主要负责实时电力设备检测、设备定位; 部署 BP 神经网络对红外测温结果进行温度修正; 修正后的温度将通过先验知识库来判断该设备是否出现异常发热症状.

顶层是数据服务层, 将中间层数据处理的结果进行可视化显示, 最终的检测结果能够以一种更为直观的表现方式呈现出.

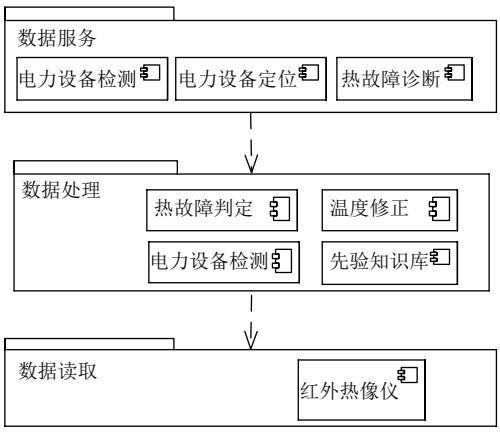

图 3 电力设备热故障检测框架

基于嵌入式深度学习的电力设备热故障诊断方法主要分为 3 项任务: 电力设备的检测与定位, 目标设备的温度提取, 目标设备热故障诊断, 类图如图 4 所示.

电力设备热故障检测的检测流程并对各个阶段进行了简单介绍. 其检测流程如图 5 所示.

(1) 从红外热像仪读取红外热成像视频流, 将其解码成帧;

(2) 电力设备检测算法检测每帧图像中是否含有电力设备, 并将其定位;

(3) 根据上一个步骤所得到的定位信息, 从红外热像仪中获取红外测温与激光测距数据;

(4) 根据红外测温与激光测距数据, 通过温度修正模块得到修正后的温度;

(5) 最后利用先验知识库, 对修正后的温度进行热故障诊断, 得到热故障检测结果.

6 实验

6.1 硬件环境

本课题涉及到的神经网络模型将分为两个阶段处理: 训练阶段和部署阶段. 两个阶段涉及到的硬件环境不一样. 由于嵌入式设备的计算资源限制, 嵌入式设备主要用于部署本课题所涉及到的算法模型, 不进行训练任务. 本文所涉及到的训练任务将由一台装有英伟达 TITAN X 显卡的机器来执行. Jetson TX1 是英伟达第二代嵌入式平台开发套件, 拥有先进的嵌入式视觉计算系统. Jetson TX1 核心仅有信用卡大小, 但 Jetson TX1 GPU 模块的浮点运算能力达到 1 Teraflops, 显然 Jetson TX1 是理想的嵌入式解决方案.

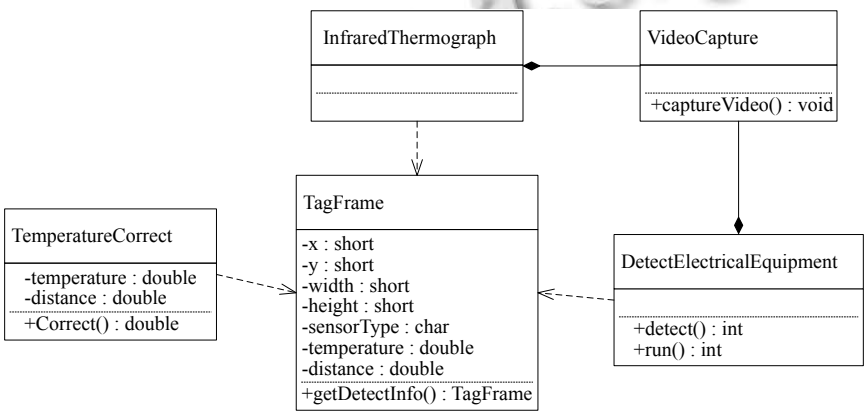

图 4 电力设备热故障检测类图

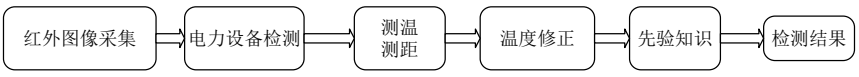

图 5 检测流程

6.2 效果展示

表 1 为系统配置表, 表 2 为部分设备实际测量温度与修正后温度的比较. 从表中可以看出, 经过 BP 神经网络修正后的最大绝对误差为 1.05 ℃, 平均绝对误差为 0.99 ℃.

表 1 Jetson TX1 配置

硬件	配置
GPU	NVIDIA Maxwell 256 CUDA cores
CPU	64-bit A57 CPUs
Memory	4 GB 64-bit LPDDDR4
Storage	16 GB eMMC
Video Encode	4K_2K 30 Hz
Video Decode	4K_2K 60 Hz

表 2 部分测量数据与 BP 修正结果对比 (单位: ℃)

环境温度	实测温度	修正温度	绝对误差
10.05	25.01	24.03	0.98
15.13	25.03	26.05	1.02
20.32	25.10	24.17	0.93
25.60	25.08	24.03	1.05
30.36	25.36	26.33	0.97

图 6 是电力设备热故障检测效果展示图, 检测结果标出所检测出来的电力设备在图像中的位置信息以及修正后的温度信息, 其中绿色标注中的电力设备表示正常, 红色标注中的电力设备表示出现过热故障.

6.3 性能与准确率测试

如表 3 所示, 我们在 Jetson TX1 嵌入式开发平台上运行相关目标检测算法, 结果表明, 与原始的 SSD (VGG16) 相比, MobileNet-SSD 在精度可接受范围内大幅提升运行速度, 处理一张图像只需 58 ms, 也就是约为 17 帧/s, 对于嵌入式设备上的应用具有重要意义.

7 总结

针对电力设备的故障识别问题, 本文提出了一种基于嵌入式深度学习的电力设备热故障检测方法. 该方法首先基于 MobileNet-SSD 算法实现电力设备检测, 让计算机自主地学习电力设备特征信息并准确检测到电力设备, 并在温度修正模块中采用 BP 神经网络来对红外测温结果进行温度修正, 最终实现无人自动化电力设备热故障检测. 实验结果表明, 该方法能够较对电力设备故障实现精准的检测. 同时, 本文提出的 Mobilenet-SSD 算法在精度可接受范围内实现了推理速度的大幅提升, 在轻量级计算平台上满足时间性能的前提下, 跟其他目标检测算法相比 Mobilenet-SSD 具有更高的准确性. 由于检测到的不同设备故障具有不同的图像特征, 其深层特征支持故障原因的进一步推理, 因此在下一阶段的中, 我们将继续研究利用深度学习分析设备故障的深层原因.

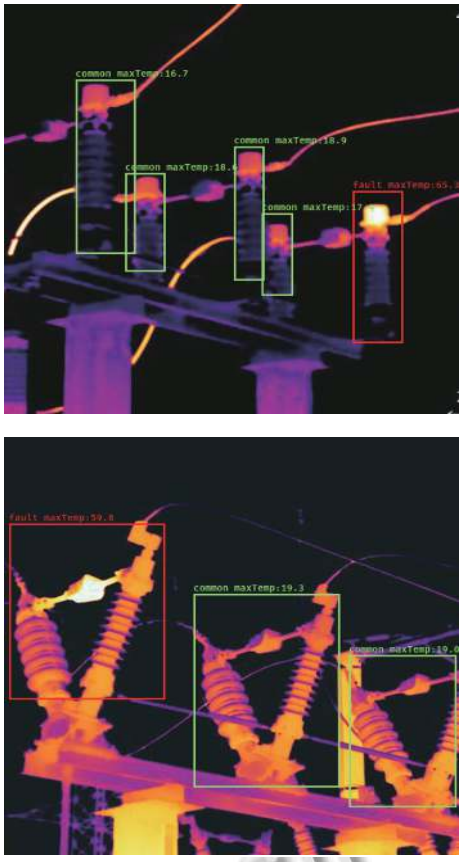

图 6 检测结果

表 3 电力设备检测结果对比

目标检测算法	运行速度 (ms)	准确率 (%)
MobileNet -SSD	58	86.7
SSD(VGG16)	597	92.0
YOLO-tiny	90	80.6
YOLOv3	322	96.1

参考文献

1 王家林, 夏立, 吴正国, 等. 电力系统故障诊断研究现状与展望. 电力系统保护与控制, 2018, 38(18): 210–216.
2 Liu J, Lin LH, Liu GQ, *et al*. A substation monitoring and warning system based on infrared technology and image separating. Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering. Xiamen, China. 2008. 66–70.
3 李杨. 电力设备状态检修及故障诊断中红外技术的应用分析. 世界有色金属, 2016, (12): 164, 167.
4 李孟兴. 电力设备故障红外诊断系统的研究与实现. 电力信息化, 2013, 11(2): 36–39.

5 Amantea R, Goodman LA, Pantuso FP, *et al*. Progress toward an uncooled IR imager with 5-mK NETD. Proceedings of SPIE Infrared Technology and Applications XXIV, vol.3436. San Diego, CA, USA. 1998. 647–659.

6 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

7 Yu P, Dong BG, Xue YJ. Electric power tower inclination angle detection method based on SIFT feature matching. Applied Mechanics and Materials, 2012, 236–237: 759–764. [doi: 10.4028/www.scientific.net/AMM.236-237.759]

8 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]

9 Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62–66. [doi: 10.1109/TSMC.1979.4310076]

10 Zhou QQ, Zhao ZB. Substation equipment image recognition based on SIFT feature matching. Proceedings of 2012 5th International Congress on Image and Signal Processing. Chongqing, China. 2012. 1344–1347.

11 Fischler MA, Bolles RC. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 1981, 24(6): 381–395. [doi: 10.1145/358669.358692]

12 Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554. [doi: 10.1162/neco.2006.18.7.1527]

13 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587.

14 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]

15 Girshick R. Fast R-CNN. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1440–1448.

16 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]

17 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.

18 Cun YL, Denker JS, Solla SA. Optimal brain damage. Proceedings of the 2nd International Conference on Neural Information Processing Systems. Cambridge, UK. 1989. 598–605.

19 Hassibi B, Stork DG. Second order derivatives for network pruning: Optimal brain surgeon. Advances in Neural Information Processing Systems. San Francisco, CA, USA. 1993. 164–171.

20 Han S, Pool J, Tran J, *et al*. Learning both weights and connections for efficient neural networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, UK. 2015. 1135–1143.

21 Hu HY, Peng R, Tai YW, *et al*. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. arXiv: 1607.03250, 2016.

22 Luo JH, Wu JX. An entropy-based pruning method for CNN compression. arXiv: 1706.05791, 2017.

23 Yang TJ, Chen YH, Sze V. Designing energy-efficient convolutional neural networks using energy-aware pruning. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6071–6079.

24 Molchanov P, Tyree S, Karras T, *et al*. Pruning convolutional neural networks for resource efficient inference. arXiv: 1611.06440, 2016.

25 Han S, Mao HZ, Dally WJ. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv: 1510.00149, 2015.

26 Chen WL, Wilson JT, Tyree S, *et al*. Compressing neural networks with the hashing trick. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France. 2015. 2285–2294.

27 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861, 2017.

28 Zhang XY, Zhou XY, Lin MX, *et al*. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 6848–6856.

29 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.

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