中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

基于轻量化卷积神经网络的金属断口图像识别

闫涵, 卢伟, 吴玉虎

引用本文:
闫涵,卢伟,吴玉虎. 基于轻量化卷积神经网络的金属断口图像识别[J]. 控制与决策, 2024, 39(9): 2858−2866.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.1424

基于轻量化卷积神经网络的金属断口图像识别

闫 涵, 卢 伟†, 吴玉虎

(大连理工大学 控制科学与工程学院，辽宁 大连 116024)

摘 要: 工业环境下金属断口图像识别是金属失效分析的重要一环, 具有重要的研究意义. 卷积神经网络(convolutional neural networks, CNN)已被证实在图像识别任务中是有效的, 但是在工业环境下的金属断口图像识别仍然面临以下问题: 1) 金属断口图像具有较强的类内复杂性和类间相似性; 2) 现有CNN网络结构复杂, 参数较多, 很难部署在嵌入式设备上. 针对上述问题, 提出一种基于轻量化CNN的金属断口图像识别方法. 首先, 设计一种多特征融合的CNN模型结构来提升网络的特征提取能力, 并给出一种混合剪枝算法对网络进行轻量化处理来降低算法复杂度; 然后, 将重要超参数搜索视为优化问题, 利用贝叶斯优化(Bayesian optimization, BO)算法进行求解, 实现模型设计和剪枝过程的自动化; 接着, 以金属断口图像数据集为例进行实验分析, 实验结果表明所提出模型仅需3.82 M的参数量即可实现97.56 %的识别精度; 最后, 将训练好的模型部署到Jetson Nano嵌入式平台上, 验证了所提出算法在实际应用中的可行性.

关键词: 深度学习；图像识别；金属断口；轻量化网络；贝叶斯优化；Jetson Nano

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.1424

引用格式: 闫涵, 卢伟, 吴玉虎. 基于轻量化卷积神经网络的金属断口图像识别[J]. 控制与决策, 2024, 39(9): 2858-2866.

Metal fracture recognition based on lightweight convolutional neural network

YAN Han, LU Wei†, WU Yu-hu

(College of Control Science and Engineering，Dalian University of Technology，Dalian 116024，China)

Abstract: The recognition of metal fracture images in an industrial environment plays a pivotal role in the analysis of metal failures and carries substantial research significance. While convolutional neural networks have been proven effective in image recognition tasks, the recognition of metal fracture images in an industrial environment still encounters the following challenges. 1) Metal fracture images exhibit strong intra-class complexity and inter-class similarity. 2) Existing CNN structures are complex, with a large number of parameters, which makes deployment on embedded devices challenging. To address the aforementioned problems, this paper proposes a metal fracture image recognition method based on the lightweight CNN. First, a CNN model structure with multi-feature fusion is designed to enhance the network's feature extraction capability. Second, a hybrid pruning algorithm is proposed to slim the network and reduce the complexity of the algorithm. Simultaneously, the search process for important hyperparameters is treated as an optimization problem, and the Bayesian optimization (BO) algorithm is utilized to solve it, thereby automating the model design and pruning process. The experimental results show that the proposed method requires only 3.82 million parameters to achieve 97.56 % recognition accuracy. The deployment on the Jetson Nano embedded platform verifies the practical feasibility of the proposed method.

Keywords: deep learning；image recognition；metal fracture；lightweight structure；Bayesian optimization；Jetson Nano

0 引 言

金属材料是航空航天、交通运输、冶金制造等领域的重大工程、大型设备的主要用材. 在复杂环境作用下, 金属材料会发生腐蚀、疲劳、断裂等各种模式的失效事故, 进而造成重大经济损失和人员伤亡. 因此, 对金属断口的高效智能识别成为保障金属

收稿日期: 2023-10-12；**录用日期:** 2024-01-16.

基金项目: 国家自然科学基金项目(62073056, 61876029)；辽宁省应用基础研究计划项目(2023JH2/101300207)；大连市重点领域创新团队项目(2021RT14)；新疆维吾尔族自治区科技重大专项项目(2022A01001).

责任编委: 魏秀琨.

†通讯作者. E-mail: luwei@dlut.edu.cn.

生产以及安全使用的关键环节. 目前,绝大多数金属断口识别依赖人工经验完成,存在如下问题: 1) 一些金属生产以及使用环境较为恶劣,难以实现人工实时检测; 2) 依赖人工经验易存在主观误判、检测效率较低等问题. 近年来,随着机器学习技术在计算机视觉领域的发展,基于机器视觉的金属断口图像识别方法取得了重大突破[1]. 然而,在实际工业应用场景下的金属断口图像识别仍然存在以下问题.

1) 特征提取问题. 现有的金属断口图像识别方法大多基于传统的机器学习方式,即人工设计特征提取算法对金属断口的轮廓、纹理、颜色等特征进行提取,再通过设计分类器对金属断口图像进行识别. 如马曼曼[2]提出了一种二维经验曲波变换的特征提取方法,将金属断口图像的能量、熵和峭度作为3种特征参量,并设计最小二乘孪生支持向量机分类器对金属断口进行识别,该方法虽然在断口图像识别任务中具有一定效果,但是需要提取特征类型较多,因此计算复杂度较高; 黎明等[3]为了增加所提取金属断口特征的多样性,提出通过利用Trace变换提取图像全局纹理特征和局部二值模式提取图像局部纹理特征的混合特征提取方法,并利用支持向量机识别自适应加权后的特征来提升断口的识别性能,该方法有效选取了对断口识别任务贡献较大的特征,并减少了分类所需特征数,但是缺少对提取特征的高级语义分析; Naik等[4]提出了通过局部二值特征以及线性判别分析分类器的金属断口识别方法,该方法进一步减少了分类所用特征数,降低了算法计算开销,但是算法仅针对两种断裂类型,因此适用度有限. 此外,金属断口图像具有较强的类内复杂性和类间相似性. 如相同类型的解理断口间会有河流状的解理花纹或突起的山脊形特征,沿晶断口中会有不同类型的韧窝断口特征. 上述现象会导致基于人工设计特征提取算法的金属断口分类方法难以达到所需性能. 近年来,基于CNN的金属断口图像识别技术也在研究中. Liu等[5]提出了一种用于铝板损伤检测的深度学习方法,该方法首先将断口图像转换为lamb波形信号,再将其作为特征训练CNN模型,通过特征转换的方式虽然能够降低深度学习模型训练时间,但是在特征提取部分需要额外步骤,增加了算法复杂度,且难以保证提取特征的有效性; Thomas等[6]使用U-Net作为特征提取算法来解决标注数据有限的问题,并将该方法应用于钢和铜的表面检测; Yang等[7]开发了一个基于堆叠自编码器的深度学习模型,用来学习铁砧的鲁棒性特征表示; Croom等[8]提出一种改进U-Net算法,用来识别具有融合缺陷金属断口的表面应力场. 虽然上述方法取得了一定效果,但是由于金属断口数据集仍然属于小样本数据量,直接训练CNN模型进行金属断口识别易发生过拟合,导致模型泛化性能较差.

2) 模型冗余问题. 随着CNN模型特征提取能力以及识别性能的提升,模型的结构也随之复杂,训练和推理所需硬件资源以及消耗时间也随之增多. 在工业场景下,模型的使用大多在嵌入式系统或边缘服务器端,因此很难运行复杂的深度学习模型. 为了解决模型冗余的问题,模型剪枝技术作为有效的模型轻量化方法获得了广泛研究[9-11]. Liu等[12]通过使用批量归一化层的比例缩放因子对模型通道进行剪枝,缩小了20倍模型尺寸并减少5倍计算操作,该方法具有简单易实现的优势,但是需要提前设定剪枝比例等参数; Yu等[13]提出了AutoSlim框架实现了快速的单步剪枝,但是该方法仅适用于对神经元的裁剪; Liu等[14]通过提出AutoCompress框架实现了自动结构化剪枝以及压缩策略选择,该方法通过启发式学习代替强化学习来实现压缩策略选择,虽然能够克服深度强化学习潜在的问题,但是算法的计算开销较大; Zheng等[15]提出了一种可微网络通道剪枝方法,在训练过程中,利用卷积核参数生成保留不同结构的概率,对该参数使用梯度下降算法进行训练,最后根据训练结果进行剪枝,该算法虽然具有一定的压缩效果,但是仅能用于调节通道的数量. 此外,上述单一剪枝方法对于深度学习模型的压缩比率有限[16],压缩后的模型仍然会存在一定程度冗余,影响模型的运算和部署效率.

3) 模型部署问题. 上述无论是基于传统机器学习还是深度学习的金属断口识别方法研究,大多停留于算法设计阶段[17],通过在高性能计算机上进行仿真实验来验证模型的有效性. 但是在工业场景下,难以直接配置高性能计算机,算法大多需要在嵌入式设备或边缘服务器上实现. 在其他类型图像识别任务中,虽然有一些在嵌入式设备部署的方案,但是部署方式较为单一,难以迎合复杂的现场环境需求[18].

针对上述问题,本文设计一种基于轻量化CNN的金属断口图像识别方法. 具体包括: 1) 针对特征复杂易错分的问题,使用基于迁移学习的VGGNet-16和ResNet-50作为特征提取基模型,通过设计自适应加权特征拼接方式来增强特征提取能力; 2) 针对模型冗余效率低的问题,设计一种混合剪枝算法,对模型进行高效压缩,降低参数量; 3) 针对模型部署问题,

给出两种基于嵌入式系统的金属断口图像识别模型部署方案,以适应不同现场的需求.

1 基于轻量化CNN的金属断口识别方法

1.1 算法总体结构

算法的整体结构如图1所示. 首先,在特征提取网络部分,选取目前在图像识别任务中应用最为广泛且性能较好的VGGNet-16模型和ResNet-50模型作为特征提取的基模型. 考虑到金属断口图像的数据量,首先,基模型在其他域中进行预训练,并通过迁移学习获得低阶特征提取器. 为了增强特征提取性能,结合集成学习的思想,设计了BO模块对不同基模型分配权重,使得模型对正确类别更加敏感. 然后,在混合剪枝算法部分:在第1阶段剪枝中利用模型BN层中的缩放因子对通道重要度进行排序,并去除低重要度通道;在第2阶段剪枝中,通过K-Means聚类算法对不同卷积核输出的特征图聚类,去除产生相似度较高特征图的卷积核. 在剪枝阶段同样使用BO模块对剪枝过程中的重要超参数进行自动搜索. 最后,各压缩模型产生的特征通过特征拼接的方式融合为新特征,并经分类层获得最终分类结果.

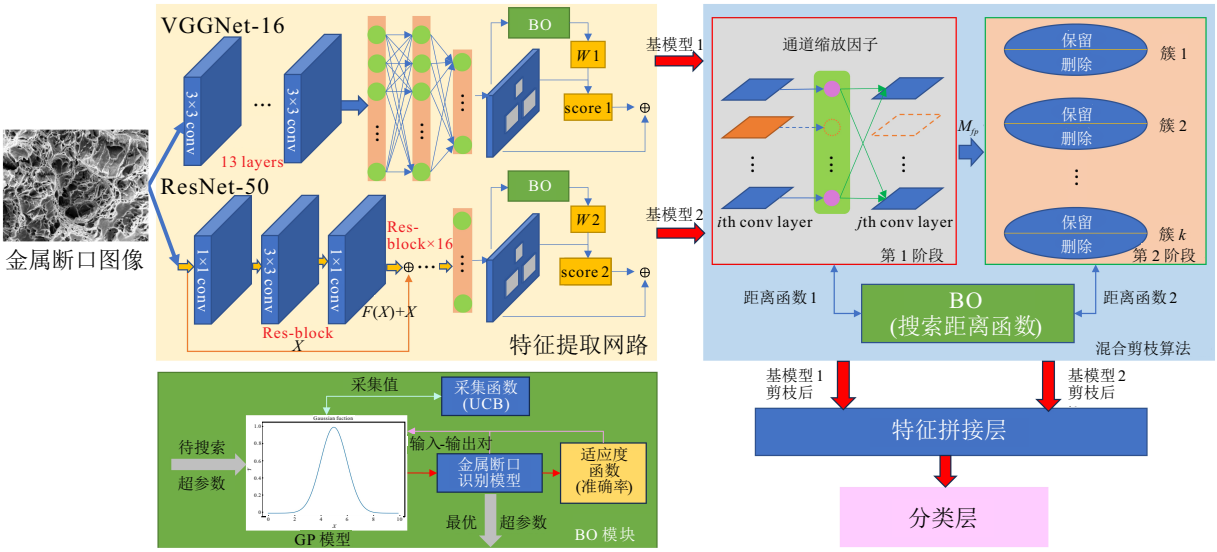
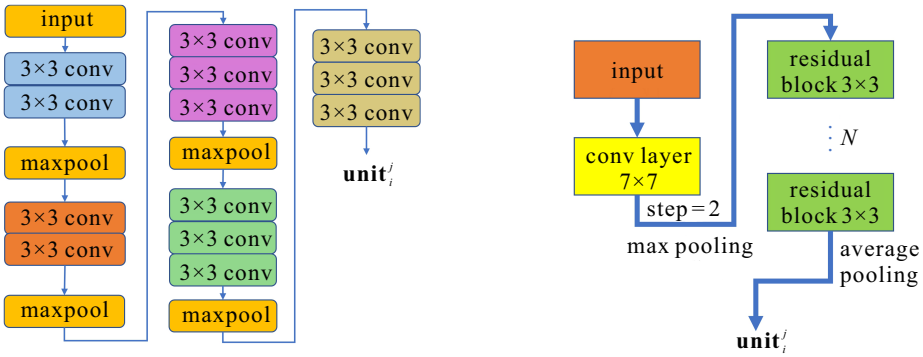


图 1 算法总体结构

1.2 特征提取网络

目前,在图像识别领域有两种常用经典网络结构VGGNet和ResNet. 二者因图像识别准确率高、速度快,被广泛用于图像识别任务的特征提取骨干网络. 针对工业环境下小样本、复杂特征的金属断口图像识别,本文选择VGGNet-16和ResNet-50作为特征提取基模型. 其中:VGGNet通过使用连续的3×3小卷积核来模拟大卷积核的感受野,进而提升算法的特征提取能力;ResNet在直连式深度学习模型结构的基础上设计跳跃连接以及拟合残差结构来提升模型效率. 在本文特征提取模块中:首先,将VGGNet和ResNet基模型在其他通用大型数据集(ImageNet)上进行预训练;然后,通过迁移学习将低阶特征提取器(边缘、纹理、颜色等)迁移至金属断口图像识别任务中;最后,将迁移模型在金属断口图像数据集上进行训练,获得最终的特征提取模型. VGGNet和ResNet模型的基础结构如图2所示.

经过训练后的基模型最终提取到的特征可简单



(a) VGGNet网络结构 (b) ResNet网络结构

图 2 VGGNet和ResNet网络结构

表示为

$$\mathbf{unit}_i^j = \begin{cases} \boldsymbol{x}_i^j \boldsymbol{c} + \boldsymbol{b}, \ i = 1; \\ (\boldsymbol{x}_i^j \boldsymbol{c} + \boldsymbol{b}) + \boldsymbol{x}_i^j, \ i = 2. \end{cases} \tag{1}$$

其中: $\mathbf{unit}_i^j$ 为不同特征提取基模型经卷积操作后获得的特征; i 为模型类型; j 为获得的特征图索引; $\boldsymbol{x} \in \mathbb{R}^{c\times w\times h}$ 为模型中最后一层卷积层的输入, c 为输入通道数, w 和 h 分别为输入图像的长和宽; $\boldsymbol{c} \in \mathbb{R}^{n\times k\times k}$ 为卷积核, n 为卷积核个数, k 为卷积核尺寸; $\boldsymbol{b} \in \mathbb{R}^{c\times 1\times 1}$ 为偏置项.

经过迁移学习后, 模型在金属断口图像识别任务中具有一定的特征提取能力, 但是由于金属断口图像间复杂的类内复杂性以及类间相似性, 直接应用所提取特征易产生错分等情况, 进而导致识别性能无法满足需求. 本节在上述特征提取的基础上设计了一个特征融合加权机制, 该机制通过对各基模型产生的特征计算加权分数, 进而融合为新的加权特征, 以进一步提升模型的特征提取能力. 具体而言, 当一张金属断口图像同时输入各基模型时, 每个基模型产生的特征图为 $\mathbf{unit}_i^j$. 基模型产生的特征图分数可计算为

$$\boldsymbol{S}_i^j = \mathrm{mul}(\theta_i, \mathbf{unit}_i^j). \tag{2}$$

其中: θ_i 为各基模型的权重, 其具体分配过程将在后文第 1.4 节中说明; $i = 1, 2$; $j = 1, 2, \ldots, n_i$; mul 表示张量逐元素相乘. 为了避免分数过小时造成的特征丢失, 将获得的分数 $\boldsymbol{S}_i^j$ 与原始特征图相加获得加权特征, 加权特征表示为

$$\boldsymbol{O}_{ij} = \mathrm{add}(\boldsymbol{S}_{ij}, \mathbf{unit}_i^j). \tag{3}$$

最后, 通过特征拼接层将不同基模型产生的加权特征进行拼接, 所获得融合加权特征表示为

$$\boldsymbol{O}_f = \mathrm{concat}([\mathbf{unit}_1^j, \mathbf{unit}_2^j]), \tag{4}$$

其中 concat 表示将特征按照通道维度进行拼接操作.

1.3　混合剪枝算法

第 1.2 节通过对不同基模型提取到的特征进行加权融合, 可增强模型的特征提取能力, 使得模型性能符合金属断口识别需求. 但是, 由于不同基模型的融合进一步增大了模型参数量, 给模型的部署和推理效率带来影响. 针对该问题, 设计一种混合剪枝算法, 对特征提取模块中的基模型进行剪枝.

首先, 在第 1 阶段剪枝过程中, 通过模型的通道重要性对卷积层进行剪枝, 如图 3 所示. 假设原模型是具有 L 层的卷积神经网路, 则对于其中某一卷积层 l 的输入输出关系可表示为

$$\boldsymbol{a}^{(l)} = f(\boldsymbol{z}^{(l)}) = f(\boldsymbol{W}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}). \tag{5}$$

其中: $\boldsymbol{z}^{(l)}$ 和 $\boldsymbol{a}^{(l)}$ 分别为第 l 层的输入和输出, $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别为权重和偏置项, f 为激活函数. 为了衡量模型中通道的冗余程度, 引入度量因子 γ 对通道的重要性进行判定. 目前, 现代 CNN 网络结构会在卷积层以及全连接层后接 BN 层对其进行归一化处理, BN 层的原理可表示为

$$\hat{\boldsymbol{z}}^{(l)} = \frac{\boldsymbol{z}^{(l)} - \boldsymbol{\mu}_{\mathcal{B}}}{\sqrt{\boldsymbol{\sigma}_{\mathcal{B}}^2 + \varepsilon}} \odot \boldsymbol{\alpha} + \boldsymbol{\beta}. \tag{6}$$

这里: $\boldsymbol{z}$ 和 $\hat{\boldsymbol{z}}$ 分别为 BN 层的输入和输出; $\boldsymbol{\mu}_{\mathcal{B}}$ 和 $\boldsymbol{\sigma}_{\mathcal{B}}$ 分别为输入在批次 $\mathcal{B}$ 上的均值和方差; 可训练参数 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 分别为缩放和平移因子, 用来将线性变换转换到其他尺度. BN 层中的可学习参数 α 具有缩放通道的性质, 因此本文选用该参数作为衡量通道重要性的度量因子, 即 $\gamma = \alpha$. 引入该因子后, 模型在第 1 阶段剪枝过程中的训练损失函数可表示为

$$L_{fp} = \sum_{(\boldsymbol{x},y)} l(f(\boldsymbol{x}, \boldsymbol{W}), y) + \lambda \sum_{\alpha\in\Gamma} g(\alpha). \tag{7}$$

其中: $(\boldsymbol{x}, y)$ 分别为训练集的输入和标签; $\boldsymbol{W}$ 为权重参数; $l(\cdot)$ 为原始 CNN 模型的损失函数; $g(\cdot)$ 为用来将缩放因子引入损失函数中的距离函数, 由下文第 1.4 节中 BO 模块自动搜索.

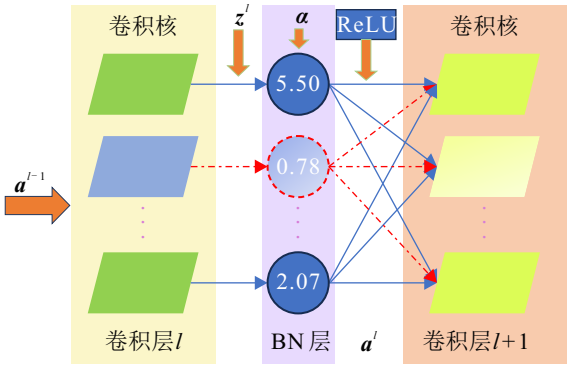


图 3　第 1 阶段剪枝

模型在经过第 1 阶段剪枝后仍然会存在冗余的卷积核, 导致模型参数量较大, 部署和推理效率有限. 因此, 在第 2 阶段剪枝过程中, 通过衡量特征图的相似性对卷积核进行修剪, 进一步压缩模型大小. 第 2 阶段剪枝的整体过程如图 4 所示. 为了去除冗余卷积核, 首先需计算各卷积核的相似性. 与其他方法中直接计算卷积核内部参数相似度不同, 所提出方法通过计算卷积核产生特征图的相似性对卷积核进行剪枝. 从生成特征图的角度进行判定可使得剪枝过程更加准确. 本节采用 K-Means 聚类算法对卷积核生成的特征图进行聚类. 由于直接聚类高纬度特征图 (如 16×16、8×8 等) 的聚类效果较差, 需先对特征图进行降维. 通过分别计算特征图的 $L1$ 范数和 $L2$ 范数将高

维特征图降为二维. 对于特征图 $\boldsymbol{f}_i^j$,降维后的特征为

$$\boldsymbol{r}_{ij}=[r_{ij}^1,r_{ij}^2]=\left[\sum_{\boldsymbol{x_n}\in\boldsymbol{f}_i^j}|\boldsymbol{x}_n|,\left(\sum_{\boldsymbol{x}_n\in\boldsymbol{f}_i^j}|\boldsymbol{x}_n|^2\right)^{\frac{1}{2}}\right]. \quad (8)$$

然后,对降维后的特征图进行 K-Means 聚类,具体过程如下.

step 1: 输入第 i 个卷积层 l_i 生成的 n 个特征图: $[\boldsymbol{f}_i^1,\boldsymbol{f}_i^2,\dots,\boldsymbol{f}_i^n]$.

step 2: 由式 (8) 计算降维后的特征集合: $[[r_{i1}^1,r_{i1}^2],[r_{i2}^1,r_{i2}^2],\dots,[r_{in}^1,r_{in}^2]]$.

step 3: 设定聚类中心数 k,根据 step 2 中的特征计算聚类中心的值 $([\boldsymbol{c}_1,\boldsymbol{c}_2,\dots,\boldsymbol{c}_k])$.

step 4: 由下式计算各样本点 $\boldsymbol{f}_i^j$ 与中心点 $\boldsymbol{c}_k$ 的距离,并根据计算结果将样本分配到不同簇内:

$$\arg\min\sum_{j=1}^{n}\sum_{m=1}^{k}|\boldsymbol{r}_{ij}-\boldsymbol{c}_k|. \quad (9)$$

step 5: 根据新的簇更新中心点值.

step 6: 迭代 step 2 ~ step 4 直至各簇中心点和样本点不变.

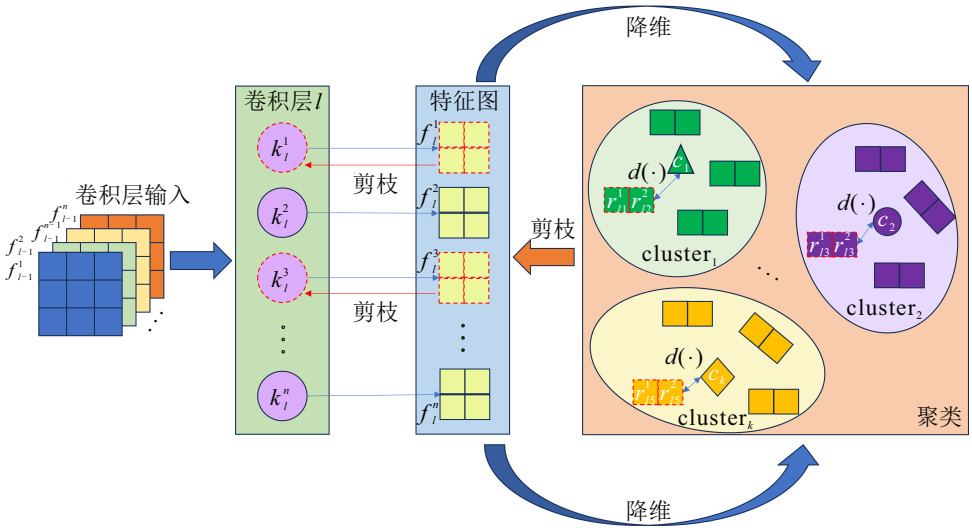


图 4 第 2 阶段剪枝

通过 K-Means 算法,相似特征图被分到相同簇,计算各特征图到中心点的距离,距离较小的特征图以及对应的卷积核被剪除,剪枝公式为

$$T_k=\text{rank}_{p\times x}[d(\boldsymbol{f}_i^m,\boldsymbol{c}_k),\dots,d(\boldsymbol{f}_i^n,\boldsymbol{c}_k)],$$
$$\text{s.t. }[\boldsymbol{f}_i^m,\dots,\boldsymbol{f}_i^n]\in\text{cluster}_k. \quad (10)$$

其中: T_k 为每个簇中的剪枝阈值; p 为剪枝比率; x 为 cluster_k 中的特征图数量; $d(\cdot)$ 用来衡量 $\boldsymbol{f}_i^j$ 与 $\boldsymbol{c}_k$ 间的距离,函数类型由下文第 1.4 节中的 BO 模块自动搜索,且 $\boldsymbol{f}_i^j\in\text{cluster}_k$.

1.4 BO 模块

在基于 CNN 的轻量化金属断口识别模型中有 4 个重要参数会影响模型的性能和效率,分别为第 1.2 节特征提取网路中两个基模型生成特征所占权重 θ_i 以及第 1.3 节混合剪枝算法中两个重要距离函数 $g(\cdot)$ 和 $d(\cdot)$. 待搜索变量的类型和搜索值对应的搜索结果如表 1 所示.

表 1 待搜索变量类型和搜索值对应结果

变量类型	初始空间	搜索值	对应结果
θ_i	$(0,1)$	$(0,1)$	$(0,1)$
$g(\cdot)$	$(0,2]$	$g(\cdot)\in(0,1]$ $g(\cdot)\in(1,2]$	$L1$ 范数 $L2$ 范数
$d(\cdot)$	$(0,3]$	$d(\cdot)\in(0,1]$ $d(\cdot)\in(1,2]$ $d(\cdot)\in(2,3]$	余弦距离 欧氏距离 曼哈顿距离

在传统的深度学习算法中,参数的设定大多基于技术人员的相关经验或通过试错的方式进行确定,不仅效率较低,且获得的结果可能无法使得模型性能达到最优[19]. 针对上述问题,本文将参数确定过程定义为一个优化问题,并通过贝叶斯优化算法 (BO) 实现自动寻优. BO 模块的结构如图 1 中 BO 模块部分所示,其优化目标为

$$\arg\max_{\theta_i,g(\cdot),d(\cdot)}\ \text{Acc}(\theta_i,g(\cdot),d(\cdot)), \quad (11)$$

其中 $\text{Acc}(\cdot)$ 为按照权重 θ_i 融合特征且根据距离函数 $g(\cdot)$、$d(\cdot)$ 进行剪枝后的模型识别准确率. 为了解决该优化问题, BO 算法实现过程如下: 首先,随机初始化 N 个不同参数组合的模型,并通过训练获得相应的识别准确率. 然后,将不同参数组合及其对应的识别准确率组成输入-输出对. 接着,优化过程通过迭代以下 4 步进行.

step 1: 使用高斯过程作为代理辅助模型拟合输入-输出对,获得模型参数和识别准确率的后验分布;

step 2: 使用上置信度 (upper confidence bound, UCB) 函数作为采集函数, 并基于后验分布计算采集函数值, 预测采集值最高的候选参数组合;

step 3: 根据预测结果在训练集上训练候选参数组合对应的模型, 并在测试集上计算识别准确率;

step 4: 适应度值最高的候选参数组合与对应的模型分类精度组成新的输入-输出对, 并将其添加到输入输出数据集中.

当达到设定的迭代次数后, 上述循环过程停止. 最后, 将具有最高分类精度模型对应的输入-输出对作为最终搜索结果.

1.5　模型部署方案

为了在工业场景下实现金属断口图像的识别, 需对基于轻量化CNN的金属断口图像识别模型在嵌入式系统上进行部署. 如图 5 所示: 针对工业背景下不同使用场景的需求, 本文给出两种基于嵌入式系统 Jetson Nano b01 的模型部署方式, 分别为模型离线部署方式和模型在线部署方式.

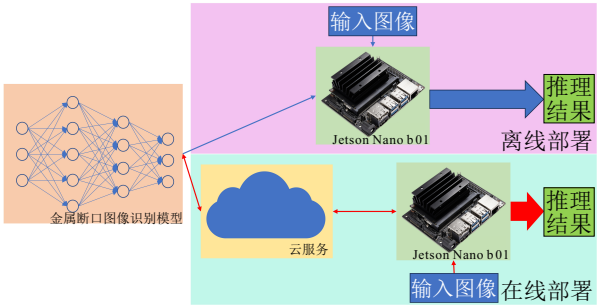


图 5　模型部署方式

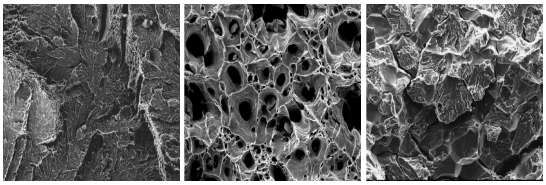
在模型离线部署方式中, 直接将嵌入式系统作为模型结构以及参数存储的载体. 首先, 将训练好的模型通过 TensorRT 转换为在嵌入式系统中运行较快的推理引擎. TensorRT 在转换过程中不会改变模型的底层运算和参数量, 通过 GPU 对模型中水平以及垂直结构进行整合从而达到加速的目的. 对于所提出算法, TensoRT 首先将模型中包含的输入张量、卷积层、BN 层、池化层以及激活函数整合为单步计算图, 将张量合并的拼接操作融入下一层网络结构, 以减少运算步骤和数据传输时间. 在生成推理引擎后, 将该模块直接下载到 Jetson Nano b01 开发板中. 推理过程中, 将金属断口图像输入到嵌入式系统中, 在该系统上使用转换后的推理引擎完成推理并输出识别结果. 这种部署方式主要针对工业现场中网络信号较差或工作环境恶劣 (高温、野外等) 的情况, 其优点是无需借助网络, 可在任意地点或场景下使用.

在模型在线部署方式中, 将嵌入式系统作为数据传输的载体, 而具体的推理过程在服务器端完成. 首先, 将训练好的模型结构和相关参数存储在服务器端 (工控机、云服务器等), 根据服务器端存储的数据搭建模型结构并加载模型参数, 对推理模型进行部署. 云服务器端算力较大, 因此模型构建过程中无需进行额外转换. 模型部署好后, 将 Jetson Nano b01 开发板通过以太网与服务器相连, 实现嵌入式系统与服务器的通信. 在推理过程中, 嵌入式系统主要负责数据的读入、传输以及最终推理结果的输出. 具体推理过程在服务器端实现. 其优点是推理速度对比直接使用嵌入式系统更快; 缺点是过度依赖网络, 最终推理的效率会受到网络延时等因素的影响, 且对工作环境有所要求.

2　实验分析

2.1　实验数据集

本文实验使用金属断口图像数据集验证所提出算法的有效性. 数据集由两部分构成: 1) 采集不同温度下 13 Cr 不锈钢以及不同压力下 X 70 管线钢的断口电镜图像; 2) 通过网络搜集的金属断口电镜图像. 初始数据集中各类均衡共 1 500 张. 金属断口数据集中包含的 3 种常见金属断口类型: 解理断口、韧窝断口以及沿晶断口典型图像如图 6 所示.



解理断口　　韧窝断口　　沿晶断口

图 6　金属断口数据集典型图像

实验前, 使用深度学习研究中常用的数据集扩充方法对初始数据集进行增强, 以避免因数据量较少带来的模型过拟合问题, 提高模型的泛化性能. 数据集的扩充利用 Keras 库实现, 通过随机旋转、缩放、裁剪等方式 (如表 2 所示), 将初始数据集扩增至 7 500 张.

表 2　数据集增强参数配置

参数	配置	描述
图像尺寸	128×128	输入图像尺寸
水平翻转	是	水平翻转输入图像
垂直翻转	是	垂直翻转输入图像
ZCA 白化	1×10^{-6}	对输入数据使用 ZCA 白化
旋转范围	50	输入图像旋转角
缩放	1/255	将像素值缩放至 [0, 1]
垂直移动范围	0.1	垂直移动百分比
水平移动范围	0.1	水平移动百分比
测试集比例	0.2	测试集与训练集比值

2.2　实验平台和参数设置

实验平台分为模型训练平台和模型部署平台. 在模型训练平台中: CPU 处理器为频率 2.3 GHz 的英

特尔酷睿i7-12700 H,用于深度学习的图形处理单元(GPU)为16 GB内存版本的RTX 3080 ti,网卡型号为英特尔AX 211,编程环境为基于Python语言的PyTorch和Keras深度学习库. 在模型部署平台中,所使用的嵌入式设备为Jetson Nano b 01,该设备CPU为具有4核的Cortex-A 57,GPU为具有128核的Maxwell,内存大小为4 GB. 配备网卡型号为英特尔8265 AC. 嵌入式平台设备如图7所示.

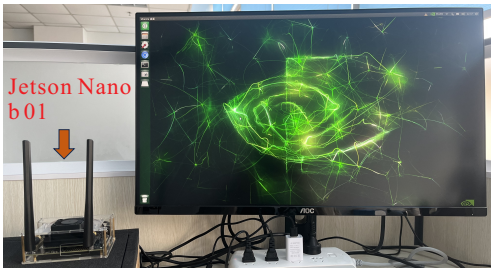


图 7 嵌入式设备实验平台

实验参数设置如下:模型训练时,输入图像的批量大小为32,训练轮次为80. 训练损失函数为交叉熵损失函数,优化器为Adam优化器,学习率为1×10^{-4}并采用指数衰减策略. 基模型在ImageNet数据集上进行预训练来增强特征提取能力,并通过迁移学习的方式避免过拟合、提升模型的泛化性能. 模型离线部署时,需对训练好的模型用TensorRT进行加速. 嵌入式设备Jetson Nano b 01采用10 w高性能运行模式. 模型在线部署时,采用模型训练时的PC作为云端推理服务器,网络通讯通过websocket协议实现.

2.3 消融实验

在所提出算法中,设计BO模块对算法中的重要超参数进行搜索. 然而,在BO模块中需通过初始点构建GP模型作为代理辅助模型,因此初始点的个数会影响所提出算法的性能. 为了研究初始点个数对所提出算法的影响,在金属断口数据集上进行消融实验,考虑到算法的时间复杂度,将BO模块初始点个数范围设置为10～30,采样间隔为5,消融实验结果如图8所示. 由图8可见:当初始点个数从10增加至20时,所提出方法的性能有所提升;当初始点个数大于20时,性能没有明显提升. 该现象表明,初始点个数会影响BO模块的性能,但是所提出算法对其并不敏感. 此外,随着初始点个数提升,BO模块的搜索时间几乎呈线性增长. 因此,为了平衡所提出算法的性能与效率,最终将BO模块的初始点个数设置为20.

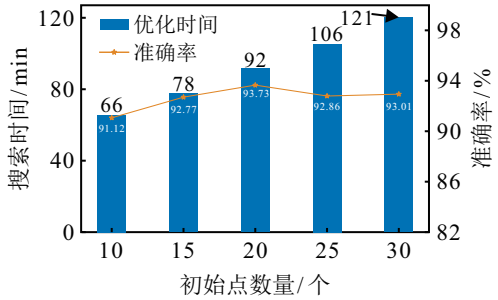


图 8 初始点个数消融实验结果

为了分析所提出方法中BO模块和混合剪枝算法对模型性能以及效率的影响,设计消融实验进行验证. 实验中将仅包含特征提取模块和分类器的模型定义为M_1模型,M_1与其不同变体的消融实验结果如表3所示. 由表3可见,手动调参设计的M_1模型效率是最低的. 加入混合剪枝算法后,模型的效率显著提升,但是所提出算法性能随之降低. 在加入BO模块后,模型在保证所提出算法轻量化的同时,自动调节了特征提取以及剪枝过程的重要超参数,使得模型性能达到最优.

表 3 不同模块消融实验结果

对比方法	Acc/%	参数量/M
M_1:基模型	96.82	38.63
M_2: M_1+混合剪枝算法	93.10	3.82
M_3: M_2+BO模块	97.56	3.82

2.4 对比结果

为了验证所提出基于轻量化CNN金属断口图像识别方法的有效性,在相同实验条件下,将该方法与先进的深度学习图像识别方法以及模型轻量化方法进行对比. 为了衡量实验结果,采用算法在测试集上的识别准确率(Acc)作为性能评价指标. 采用模型参数量、浮点运算量(floating point operations, FLOPs)以及模型在嵌入式系统上的推理速度作为效率评价指标. 实验结果如表4所示.

表 4 实验对比结果

对比方法	Acc/%	FLOPs/G	参数量/M	推理速度/s
VGGNet-16 [20]	95.13	0.732	14.23	8.34
ResNet-50 [21]	94.25	0.158	25.59	7.94
DenseNet-121 [22]	93.18	0.916	7.04	6.32
EfficientNet [23]	96.02	0.391	15.34	11.24
MobileNet [24]	92.52	0.288	4.20	7.78
FCS-MLFT [25]	90.44	0.197	3.94	6.84
HRank [26]	91.06	0.177	3.56	7.05
MRKP [27]	92.69	0.162	4.13	6.53
本文方法+在线部署	97.56	0.131	3.82	1.20
本文方法+离线部署	97.56	0.131	3.82	5.87

注: G: Giga, 1×10^9; M: Million, 1×10^6; s: seconds, 秒.

由表4可见,相比于基模型、高性能深度学习图像识别方法以及先进的模型轻量化方法,所提出方法在模型性能和效率指标上均有较大提升. 首先,对比基模型VGGNet-16和ResNet-50模型,所提出方法在识别准确率上分别提升了2.55 %和3.51 %,模型参数量分别减少了73.16 %和85.07 %,且在嵌入式系统上的推理速度提升了29.62 %和26.07 %;然后,与高性能

深度学习图像识别方法 DenseNet-121 和 EfficientNet 进行对比, 所提出方法将金属断口的识别准确率提升了 4.70 % 和 1.60 %, 模型参数量降低了 45.74 % 和 17.85 %, 且推理速度提升了 7.12 % 和 47.78 %; 接着, 与轻量化网络 MobileNet 进行对比, 所提出方法仍然能够在参数量减少 9.05 % 的同时将识别精度和推理速度提升 5.44 % 和 24.55 %; 最后, 与近年来先进的模型轻量化方法进行对比, 所提出方法在模型性能上有较大提升, 对比 FCS-MLFT、HRank 与 MRKP 方法分别提升了 7.84 %、7.14 % 和 5.25 %. 在模型效率方面, 对比 FCS-MLFT、HRank 与 MRKP 方法, 所提出方法在参数量上略高于 HRank 方法, 在 FLOPs 值上优于其他方法, 且在嵌入式系统上推理速度最快. 上述结果表明了所提出方法在特征提取和模型轻量化方面的有效性.

为了验证模型不同部署方法的性能, 表 4 中第 9 行与第 10 行对比了不同部署方式下模型的推理速度. 第 9 行为模型的在线部署模式, 由表 4 可见, 由于推理过程是在装备了 GPU 的高性能计算机上完成, 整个推理时间仅需 1.20 s. 对比第 10 行中直接在嵌入式设备上对模型进行部署所需的 5.87 s, 在线部署方式将推理时间缩短了 79.56 %. 然而, 在工业场景下, 有时不具备良好的网络环境, 且高性能计算机造价较高, 往往是嵌入式系统的 20 倍左右. 因此, 鉴于离线部署方式的灵活度和低成本, 本文认为该方式是更符合大多数工业实际需求的部署方式.

为了分析所提出方法对不同类型金属断口的分类性能, 图 9 给出了金属断口识别任务上的混淆矩阵. 由图 9 可见: 所提出方法对解理断口和沿晶断口分类性能较好; 错分情况主要集中在韧窝断口中, 且在韧窝断口的错分图像中, 主要将韧窝断口分类为解理断口, 其原因是韧窝断口中常常会掺杂少量解理花纹, 所提出方法虽然可提升算法识别性能, 但是仍然会出现部分错分现象.

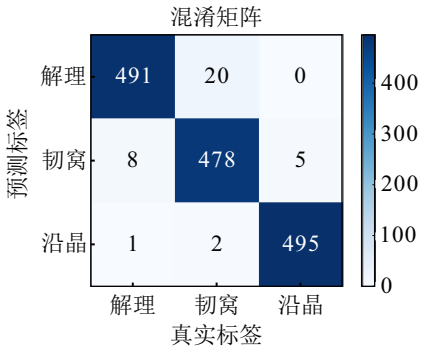


图 9　断口识别结果混淆矩阵

从金属断口图像测试数据集中随机选择 3 类典型断口图像各一张, 在嵌入式系统上进行断口类型识别实验. 图 10 为金属断口图像的识别结果, 由图 10 可见, 3 种断口类型均被正确识别, 无错分发生.

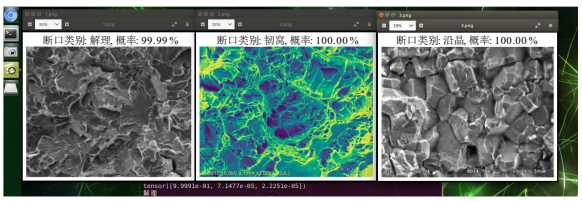


图 10　断口识别可视化结果

3　结　论

目前, 工业环境下的金属断口图像识别大多基于传统机器学习方法, 识别精度不高. 少数基于深度学习的方法, 因模型参数量和运算量过大, 无法应用于真实工业场景中. 为了提高工业场景下金属断口的识别性能, 本文提出了一种基于轻量化 CNN 的金属断口图像识别方法, 该方法以 VGGNet 和 ResNet 作为特征提取基模型, 设计了自适应加权特征拼接方法来增强特征提取能力, 利用混合剪枝算法降低了模型的参数量, 并通过 BO 模块对模型的重要参数进行自动搜索. 通过在金属断口图像数据集上的实验验证了所提出算法的有效性. 同时, 通过两种不同的部署方式将该模型部署到嵌入式系统中, 以验证所提出算法在实际应用的可行性. 未来研究可放在如下两方面: 1) 如何在数据量有限的情况下, 进一步提升深度学习对复杂断口特征的提取能力, 并对该能力进行可解释性研究; 2) 如何在嵌入式系统中实现金属断口图像的实时采集、读取以及检测.

参考文献 (References)

[1] 雷明. 机器学习与应用 [M]. 北京: 清华大学出版社, 2019: 26-33.
(Lei M. Machine learning and application[M]. Beijing: Tsinghua University Press, 2019: 26-33.)

[2] 马曼曼. 基于经验曲波变换的断口图像处理方法研究 [D]. 南昌: 南昌航空大学, 2018.
(Ma M M. Research on fracture image processing method based on empirical curvelet transform[D]. Nanchang: Nanchang Hangkong University, 2018.)

[3] 黎明, 邢冬冬, 汪宇玲, 等. 多特征融合的金属断口图像分类 [J]. 模式识别与人工智能, 2018, 31(5): 453-461.
(Li M, Xing D D, Wang Y L, et al. Metal fracture image classification based on adaptive fusion of multiple features[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(5): 453-461.)

[4] Naik D L, Kiran R. Identification and characterization of fracture in metals using machine learning based texture recognition algorithms[J]. Engineering Fracture Mechanics, 2019, 219: 106618.

[5] Liu H, Zhang Y F. Deep learning based crack damage detection technique for thin plate structures using guided lamb wave signals[J]. Smart Materials and Structures, 2020, 29(1): 015032.

[6] Thomas A, Durmaz A R, Straub T, et al. Automated quantitative analyses of fatigue-induced surface damage by deep learning[J]. Materials, 2020, 13(15): 3298.

[7] Yang J, Chen B, Wang Y N, et al. Crack detection in carbide anvil using acoustic signal and deep learning with particle swarm optimisation[J]. Measurement, 2021, 173: 108668.

[8] Croom B P, Berkson M, Mueller R K, et al. Deep learning prediction of stress fields in additively manufactured metals with intricate defect networks[J]. Mechanics of Materials, 2022, 165: 104191.

[9] Su T T, Zhang J S, Yu Z Y, et al. STKD: Distilling knowledge from synchronous teaching for efficient model compression[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(12): 10051-10064.

[10] 刘相良, 张林丛, 朱宏博, 等. 基于多级表征线索注意模型的轻量化抠图方法[J]. 控制与决策, 2024, 39(1): 87-94.
(Liu X L, Zhang L C, Zhu H B, et al. A lightweight image matting method based on attentive model for multi-level appearance cues[J]. Control and Decision, 2024, 39(1): 87-94.)

[11] 程旗, 李捷, 高晓利, 等. 基于深度稀疏低秩分解的深度神经网络轻量化方法[J]. 控制与决策, 2023, 38(3): 751-758.
(Cheng Q, Li J, Gao X L, et al. Lightweight method of deep neural network based on deep sparse low rank decomposition[J]. Control and Decision, 2023, 38(3): 751-758.)

[12] Liu Z, Li J G, Shen Z Q, et al. Learning efficient convolutional networks through network slimming[C]. IEEE International Conference on Computer Vision. Venice, 2017: 2755-2763.

[13] Yu J H, Huang T. AutoSlim: Towards one-shot architecture search for channel numbers[J/OL]. 2019, arXiv: 1903.11728.

[14] Liu N, Ma X L, Xu Z Y, et al. AutoCompress: An automatic DNN structured pruning framework for ultra-high compression rates[C]. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, 2020: 4876-4883.

[15] Zheng Y J, Chen S B, Ding C H Q, et al. Model compression based on differentiable network channel pruning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(12): 10203-10212.

[16] Yang W, Xiao Y C. Structured pruning via feature channels similarity and mutual learning for convolutional neural network compression[J]. Applied Intelligence, 2022, 52(12): 14560-14570.

[17] 叶卓勋, 刘妹琴, 张森林. 基于轻量化深度学习网络的工业环境小目标缺陷检测[J]. 控制与决策, 2023, 38(5): 1231-1238.
(Ye Z X, Liu M Q, Zhang S L. Small-scale defect detection in industrial environment based on lightweight deep learning network[J]. Control and Decision, 2023, 38(5): 1231-1238.)

[18] 顾德英, 罗聿伦, 李文超. 基于改进YOLOv5算法的复杂场景交通目标检测[J]. 东北大学学报: 自然科学版, 2022, 43(8): 1073-1079.
(Gu D Y, Luo Y L, Li W C. Traffic target detection in complex scenes based on improved YOLOv5 algorithm[J]. Journal of Northeastern University: Natural Science, 2022, 43(8): 1073-1079.)

[19] Li J Y, Zhan Z H, Xu J, et al. Surrogate-assisted hybrid-model estimation of distribution algorithm for mixed-variable hyperparameters optimization in convolutional neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(5): 2338-2352.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. 2014, arXiv: 1409.1556.

[21] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[22] Huang G, Liu Z, van Der M L, et al. Densely connected convolutional networks[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 4700-4708.

[23] Tan M X, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks[J/OL]. 2019, arXiv: 1905.11946v5.

[24] Howard A G, Zhu M L, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[J/OL]. 2017, arXiv: 1704.04861.

[25] Yang W, Xiao Y C. Structured pruning via feature channels similarity and mutual learning for convolutional neural network compression[J]. Applied Intelligence, 2022, 52(12): 14560-14570.

[26] Lin M B, Ji R R, Wang Y, et al. HRank: Filter pruning using high-rank feature map[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 1526-1535.

[27] Zhu J H, Pei J H. Progressive kernel pruning CNN compression method with an adjustable input channel[J]. Applied Intelligence, 2022, 52(9): 10519-10540.

作者简介

闫涵 (1994–), 男, 博士生, 从事深度学习、图像处理与模式识别等研究, E-mail: yan_han@mail.dlut.edu.cn;

卢伟 (1976–), 男, 教授, 博士, 博士生导师, 从事粒计算、计算智能、知识发现和表示等研究, E-mail: luwei@dlut.edu.cn;

吴玉虎 (1980–), 男, 教授, 博士, 博士生导师, 从事非线性系统、博弈决策和控制理论方法等研究, E-mail: wuyuhu@dlut.edu.cn.