

DOI: 10.3969/j.issn.0254-0150.2024.10.013

文献引用: 刘信良,陈国宁,苏化,等.一个用于磨粒图像快速分类的轻量化 CNN 模型[J].润滑与密封,2024,49(10):102-107.

Cite as: LIU Xinliang, CHEN Guoning, SU Hua, et al. A lightweight CNN model for fast classification of wear particle images[J]. Lubrication Engineering, 2024, 49(10): 102-107.

一个用于磨粒图像快速分类的轻量化 CNN 模型

刘信良, 陈国宁, 苏化, 王静秋, 王晓雷

(南京航空航天大学, 直升机动力学全国重点实验室, 江苏南京 210016)

摘要: 针对磨粒分析 CNN 模型存在参数多、运算速度慢、难以实际应用等问题, 开展磨粒图像分类 CNN 模型的轻量化研究。通过分析模型各层的参数量、运算量和剪枝敏感度, 确定卷积层 4 和卷积层 5 为滤波器剪枝的目标; 对卷积层 4 和 5 所有滤波器重要性进行计算并排序, 以 75% 的剪枝率去除重要性低的滤波器并重新训练, 获得轻量化模型。实验结果表明, 轻量化后的模型在保证准确率几乎不降低的情况下实现了磨粒图像的快速分类, 其理论参数量和内存占用量均减少 50% 以上, 运算速度提高 20% 以上。研究结果为 CNN 模型在便携式、移动式铁谱分析设备上的应用提供参考。

关键词: 铁谱分析; 卷积神经网络; 磨粒图像分类; 轻量化

中图分类号: TH117.1; V263.6

A Lightweight CNN Model for Fast Classification of Wear Particle Images

LIU Xinliang, CHEN Guoning, SU Hua, WANG Jingqiu, WANG Xiaolei

(National Key Laboratory of Helicopter Aeromechanics, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu 210016, China)

Abstract: Aiming at the disadvantages of CNN model for wear particle analysis, such as too many parameters, slow operation speed, and difficulty in practical application, research on lightweight of the CNN model for wear particle image classification was carried out. The parameters, computation and pruning sensitivity of each layer in the CNN model were analyzed, which determined the convolution layer 4 and 5 as the filter pruning targets. The importance of all filters in convolution layer 4 and 5 was calculated, and the filters were sorted by their importance. A lightweight model was obtained after removing the filters with low importance at 75% pruning rate and retraining. The experimental results show that after lightweight processing, the number of theoretical parameters and amount of memory usage are reduced by more than 50%, and the operation speed is increased by more than 20% while the accuracy barely drops. This study provides an idea for the application of CNN model on the portable and mobile ferrography equipment.

Keywords: ferrography; convolutional neural network; wear particle classification; lightweight processing

机械设备运转过程中, 各零部件间的相对运动会
导致摩擦、磨损发生, 而磨损加剧会导致零件失效,
最终引发机械故障^[1-5]。铁谱分析是机械设备润滑系
统磨损检测的常用手段, 通过观察铁谱片上沉积的铁
磁性磨粒, 分析磨粒的尺寸、形貌、分布等参数, 可

判断设备磨损状态^[6-9]。

传统铁谱图像分析技术结合数字图像处理及模式
识别方法, 可将图像中的磨粒与背景分离, 识别出每
个磨粒的类别, 进而依据所有磨粒信息判断设备的磨
损情况。该方法提升了铁谱分析的自动化程度, 对一

基金项目: 直升机传动技术重点实验室基金项目 (HTL-A-21G03)

收稿日期: 2023-06-07; 修回日期: 2023-06-29

作者简介: 刘信良 (1993—), 男, 博士研究生, 研究方向为摩擦学, 状态监测与故障诊断。E-mail: liuxinliang@nuaa.edu.cn。

通信作者: 王静秋 (1972—), 女, 博士, 教授, 研究方向为摩擦学、计算机图像处理、分子动力学模拟。E-mail: mееjqwang@nuaa.edu.cn。

些典型磨粒分析取得较好的结果。然而，该方法在磨粒图像采集时存在光照不均、噪声等问题，影响磨粒分割的准确性；对于磨粒识别的准确性依赖人工设计的特征；其分析流程是多步骤单向线性过程，容易导致误差累积与传递。

近年来，卷积神经网络（CNN）因具有端到端处理、自动提取特征、自学习等优点被引入到铁谱分析技术之中^[10]。WANG 等^[11]构建了一个包含 4 个卷积层、2 个全连接层的 CNN 模型，实现了对 7 种磨粒的分类；PENG 等^[12]以 AlexNet^[13]模型为特征提取器，用 4 个支持向量机（SVM）^[14]模型作为分类器，实现了对 4 种磨粒的分类；LIU 等^[15]基于 DeepLabv2^[16]构建了“Encoder-ASPP-Decoder”结构的语义分割模型，实现了对 5 种磨粒按照类别的分割和识别；SUN 等^[17]基于 Mask RCNN^[18]提出了磨粒实例分割模型，实现了对 5 种磨粒的分类、检测和分割。

上述模型可以达到较高精度，提高铁谱分析的自动化和智能化的水平。然而，卷积神经网络参数众多、运算量巨大，要求高性能的硬件平台作为支撑，提升了其应用成本，限制了其在便携式、移动式铁谱分析设备上的应用。本文作者开展了磨粒图像分类 CNN 模型轻量化研究，利用滤波器剪枝的方法对模型进行剪枝，扩展铁谱分析的应用场景。

1 磨粒图像分类 CNN 模型构建与训练

1.1 磨粒图像分类数据集

实验所用的磨粒图像采集自航空发动机的润滑油系统，在 100 倍光学显微镜下拍摄得到。随着航空发动机运行时间增长，磨损也会逐步发生并变得严重。在制谱时，磨损颗粒会在谱片上堆积起来，以磨粒链的形式呈现在磨粒图像上。链状磨粒的多少，即磨粒含量在一定程度上反映了设备的磨损状态，磨粒含量越高表示磨损状况越严重。如图 1 所示，根据链状磨粒含量从低到高的顺序，可将磨粒图像划分为 4 个类别，分别是类别 0-无磨粒、类别 1-少量磨粒、类别 2-中量磨粒和类别 3-大量磨粒，分别对应轻微、正常、较严重和严重磨损^[19]。

设备运行过程中，相比于正常状态，异常磨损发生的机会较少，异常磨粒的样本远少于正常和轻微磨损。图 2 所示为各类图像数量分布，可以看出数据集存在类别不均衡、数据量少的问题，因此需要进行类别均衡及数据增广等操作。

针对类别不均衡问题，从类别 0 和类别 1 的图像样本中分别随机挑选出 200 个样本，与类别 2 和类别 3 的全部样本组成新数据集；针对类别 3 数据过少问题，用类别 0 的磨粒图像作为背景，类别 1 和 2 图像

中的磨粒链分割出来作为前景库，随机粘贴到背景图中，平滑处理后得到 29 幅类别 3 的图像。

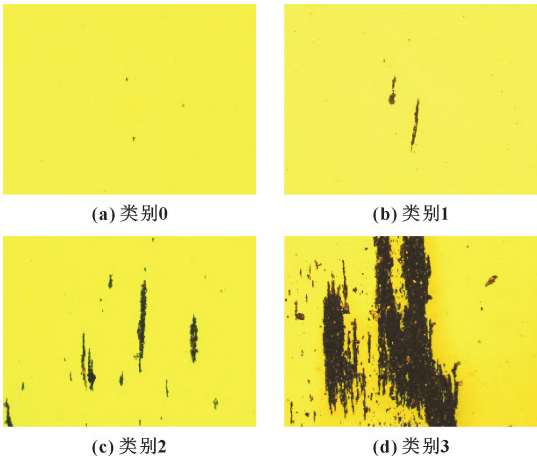


图 1 不同等级的磨粒图像

Fig. 1 Different level of wear particle images: (a) class 0; (b) class 1; (c) class 2; (d) class 3

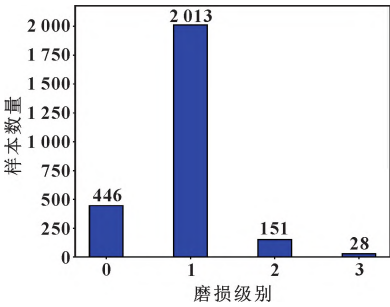


图 2 数据集样本分布

Fig. 2 Distribution of the samples in dataset

表 1 所示为经过类别平衡和数据增广后新数据集，训练时从测试集随机抽取 25% 的样本作为验证集。

表 1 经过类别均衡和数据增广后的数据集

Tab. 1 Dataset after category balance and data augmentation

数据	类别 0	类别 1	类别 2	类别 3
训练集	150	150	101	37 *
测试集	50	50	50	20
总计	200	200	151	57

注：* 处含 29 张扩充样本和 8 张原始样本。

1.2 磨粒图像分类 CNN 模型的选取

利用上述磨粒图像分类数据集对 3 个常用的 CNN 分类模型 AlexNet、ResNet34 和 ResNet50^[20]进行初步训练，其损失函数如图 3 所示。可以看出，在 SGD 和 Adam 优化器下进行迁移训练，AlexNet 的收

收敛速度更快。3 个模型各项参数如表 2 所示, 可见 AlexNet 的 F1 值达到 88.37%, 分类性能最优; 相较 ResNet34 (R34) 和 ResNet50 (R50), AlexNet 理论运算量小、处理速度更快。但 AlexNet 参数量大, 会占用较多的内存和显存资源, 如果采用模型轻量化技术, 有望在不降低模型表现的前提下降低其参数量, 从而减少资源占用。综合考虑准确率和性能, 文中采用 AlexNet 作为磨粒图像分类模型。

磨粒图像分类 CNN 模型的训练过程采用 Adam 优化器, 设置学习率为 0.000 1, 批量 (Batch) 大小为 32, 训练 6 个周期 (Epoch) 共迭代 98 次。训练过程中损失值平稳下降, 训练集与验证集的准确率稳步上升且差距很小, 证明训练完成。计算各类别的 F1 值, 其平均值为 88.37%。

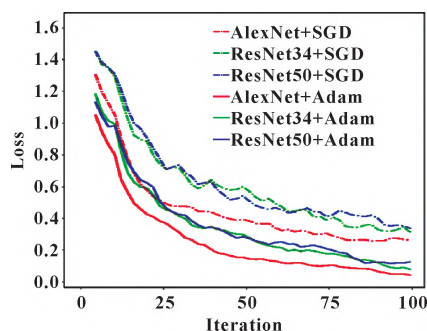


图 3 3 个模型训练过程中的损失值

Fig. 3 The loss values during the training of three models

表 2 3 种模型参数对比

Tab. 2 Parameter comparison of three models

参数	AlexNet	R34	R50
F1 值/%	88.37	86.66	85.27
参数量/ 10^6	61.1	11.7	21.8
乘加运算/ 10^9	0.8	1.8	3.7
CPU 处理速度(毫秒/张)	75.3	165.8	306.5
GPU 处理速度(毫秒/张)	1.9	3.9	7.0

经过训练之后, 磨粒图像分类模型具备了较强的分类能力, 但其参数多、运算量大, 因此需要对其进行轻量化, 提升实际应用的效率。

2 磨粒图像分类 CNN 模型轻量化

2.1 磨粒图像分类 CNN 模型的参数分析

图 4 所示为磨粒图像分类 CNN 模型各层的理论参数和理论运算量。可以看出, 模型参数主要集中在全连接层, 而运算主要集中在卷积层。因此, 在剪枝敏感度相同的情况下, 优先对理论参数量和理论运算

量占比较高的层进行剪枝, 这样会得到较好的压缩和加速效果。

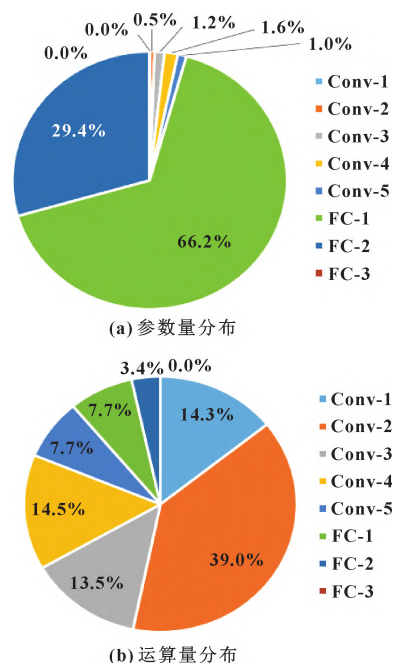


图 4 模型各层的参数量和运算量占比

Fig. 4 The proportion of parameters and calculation for each layer of the model: (a) distribution of parameters; (b) distribution of operation

此外, 还需要考虑各层的剪枝敏感度。随着剪枝率的上升, 模型的精度通常会有不同程度的下降, 下降幅度越大, 说明该层对剪枝操作越敏感。为测定各层的剪枝敏感度, 按递增的剪枝率独立地对所有包含参数的层分别执行剪枝操作 (不包含再训练)。图 5 所示为磨粒图像分类 CNN 模型中卷积层和全连接层的敏感度曲线, 可以看出, 卷积层 4 (Conv-4) 和卷积层 5 (Conv-5) 的敏感度较低。

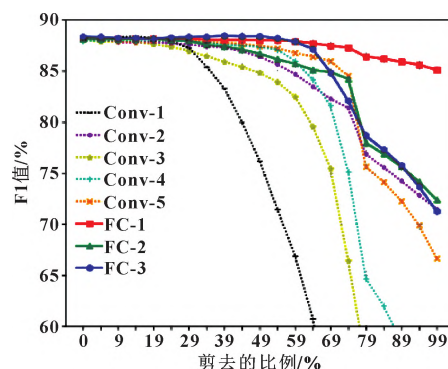


图 5 模型中各层的剪枝敏感度

Fig. 5 Pruning sensitivity of each layer in the model

总体来看,全连接层的敏感度低于卷积层;具体到单个的层,卷积层1的剪枝敏感度最高,全连接层1的敏感度最低。因此,选择低敏感度的层剪枝、避免对高敏感度的层剪枝,可以减少轻量化过程给模型带来的精度损失。

2.2 滤波器剪枝原理与策略

滤波器剪枝是通过去除卷积层/全连接层中不重要的卷积核/神经元,减少模型参数和运算量,达到轻量化效果的方法,具有不依赖底层稀疏矩阵运算库、通用性高的优点。

对于卷积层的 K 个滤波器,分别按式(1)计算其重要性并将其按照增序排列,得到序列 $\{I_1, I_2, \dots, I_K\}$,取该序列对剪枝率 R 的百分位数 P_R 作为滤波器剪枝的阈值 T ,见式(2)。

$$I = \sum_{i=0}^F \sum_{j=0}^F \sum_{k=0}^{C_i} |w_{i,j,k}| \quad (1)$$

$$T = P_R \quad (2)$$

重要性低于该阈值的滤波器将在剪枝过程中被去除。由于特征图与滤波器一一对应,故该层输出特征图的通道会减少。受此影响,卷积层 $l+1$ 中的蓝色部分代表的滤波器被相应地去除。如图6(a)所示。

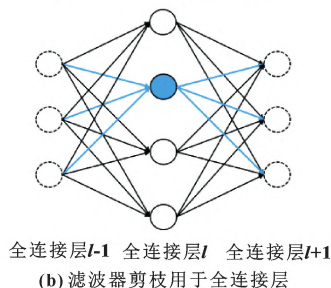
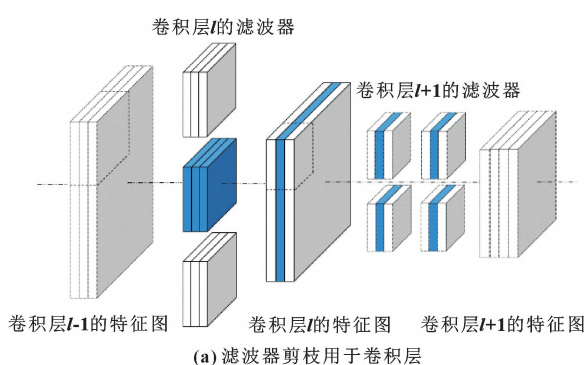


图6 滤波器剪枝原理

Fig. 6 The principle of filter pruning technology: (a) filter pruning applying on convolution layers; (b) filter pruning applying on fully connected layers

全连接层的剪枝以图6(b)为例,剪枝过程中去除的对象是单个的神经元。设全连接层 l 有 N 个神

经元,前一层有 M 个神经元,则全连接层 l 的权重矩阵可视为 $N \times M$ 的二维矩阵,记其中的元素为 w ,则该层第 i 个神经元的重要性 I 可表示为

$$I = \sum_{j=0}^M |w_{i,j}| \quad (3)$$

经过剪枝后当前层的权重被去除,后一层中的部分权重同样被去除,如图6(b)中蓝色部分所代表的权重。

2.3 滤波器剪枝过程及结果分析

根据图5中各层敏感度曲线,以 $R_1 = R_2 = 75\%$ 的剪枝率分别对磨粒图像分类CNN模型的卷积层4(Conv-4)和卷积层5(Conv-5)剪枝。图7所示为滤波器剪枝的过程。为避免单次剪枝可能存在的误差,剪枝操作分两轮进行,每轮采用37.5%的剪枝率,且剪枝后立即进行再训练,重复上述过程2次之后达到预定的剪枝率(75%)。

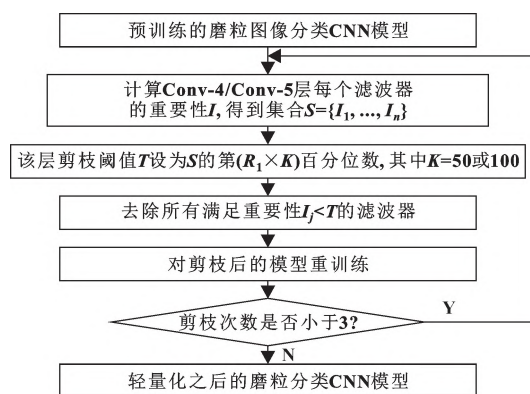


图7 滤波器剪枝过程

Fig. 7 The process of filter pruning technology

图8表示了剪枝及再训练过程对磨粒图像分类CNN模型的影响。

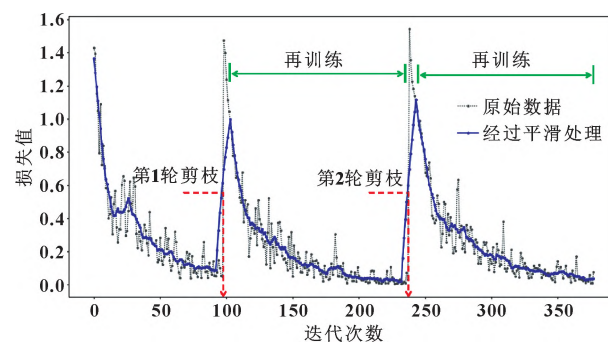


图8 滤波器剪枝及再训练过程的损失值变化

Fig. 8 Loss value changes during filter pruning and retraining

由图 8 可以看出,在两轮剪枝的过程中,损失值总体呈现波浪形状。每经过一轮剪枝,模型的损失值都会突然上升到预训练之前的初始水平,但经过之后的再训练,损失值重新回到较低的水平。该现象证明剪枝后的再训练对于维持模型的精度起到了重要作用。

在磨粒链测试集上的结果表明,轻量化后的模型的平均 F1 值在较小损失的情况下,其理论参数量和理论运算量却得到了大幅的减少。如表 3 所示,轻量化后的模型平均值仅下降 0.74% (由 88.37% 降至 87.63%),其中,类别 0 和 3 下降较小 (0.22% 和 0.25%),类别 1 和 2 下降较大 (0.94% 和 1.57%)。如表 4 所示,经过滤波器剪枝后模型的理论参数量减少 51.8%,理论运算量减少 32.8%,理论性能有

了较大提升。值得注意的是,由于滤波器剪枝会同时影响当前层和后续层,尽管仅对卷积层 4 和卷积层 5 进行了滤波器剪枝,但全连接层 1 仍间接受到了剪枝操作的影响;卷积层 5 同时受到前一层剪枝和本层剪枝的影响,因此理论参数量和运算量的降幅最大,均在 90% 以上。

表 3 轻量化前后模型的分类结果

Tab. 3 Classification results of models before and after lightweight

模型	F1 值/%				
	类别 0	类别 1	类别 2	类别 3	平均值
轻量化前	85.11	81.48	91.67	95.23	88.37
轻量化后	84.89	80.54	90.10	94.98	87.63

表 4 模型滤波器剪枝结果

Tab. 4 The results of filter pruning for the model

卷积层	剪枝前后的 滤波器数量	剪枝率/%	剪枝前后的理论参数量及变化率		剪枝前后的理论运算量及变化率	
			参数量/ 10^6	变化率/%	运算量/ 10^6	变化率/%
Conv-4	64,64	75	0.88 0.22	-75.0	71.66 17.92	0
Conv-5	192,192	75	0.59 0.04	-93.7	37.75 9.36	-93.8
FC-1	-	0	37.75 9.44	-75.0	37.75 9.44	-75.0
总体	-	-	57.02 27.49	-51.8	492.59 375.14	-32.8

表 5 滤波器剪枝对模型的压缩效果

Tab. 5 Effect of filter pruning on model compression

相关指标	剪枝前	剪枝后	变化率/%
理论参数量/ 10^6	57.02	27.49	-51.8
.pth 权重文件大小/MB	222.7	107.34	-51.8
* 仅加载模型时的内存占用/MB	200.0	100.0	-50.0
* CPU 单张推理时内存占用峰值/MB	235.8	122.4	-48.1
* CPU 四张推理时内存占用峰值/MB	287.2	168.8	-41.2
* GPU 单张推理时显存占用峰值/MB	836.6	709.0	-15.3
* GPU 四张推理时显存占用峰值/MB	871.0	743.0	-14.7

注: * 处数据均为单独运行 10 次后取平均值。

表 6 滤波器剪枝对模型的加速效果

Tab. 6 Effect of filter pruning on model acceleration

参数	剪枝前	剪枝后	变化率/%
理论运算量/ 10^6	492.6	375.14	-23.84
* CPU 单张推理时间 t_1 /ms	75.3	54.6	-27.44
* GPU 单张推理时间 t_2 /ms	1.9	1.5	-21.05

注: * 处数据均为单独运行 10 次后取平均值。

为验证模型实际的压缩和加速效果,在 Windows 平台采用如下硬件设备进行实际测试:内存 16G,

CPU 使用 AMD Ryzen 5 3600X, GPU 使用 NVIDIA GeForce RTX 2060 SUPER; 软件使用 Pytorch^[21] 框架。测试结果如表 5 和表 6 所示,无论在 CPU 还是在 GPU 上运行,剪枝后的模型都表现出明显的压缩和加速效果。然而,在 GPU 上运行时的轻量化效果不如在 CPU 上明显,这是由于深度学习框架已经集成了 cuDNN^[22] 等专门用于深度神经网络的加速库。但对于使用存储空间和计算性能都相对有限的便携式设备,滤波器剪枝的模型压缩和加速效果十分明显。

3 结论

(1) 针对磨粒图像分类 CNN 模型的轻量化研究,使用滤波器剪枝的方法,对磨粒图像分类 CNN 模型的 Conv-4、Conv-5 层进行轻量化。理论计算表明,轻量化后模型准确率几乎不降低(仅下降 0.74%)的情况下,其参数数量和运算量显著减少。实际运行结果表明,轻量化后的模型达到较高的压缩率,且在 CPU 和 GPU 的运算速率均提升 20% 以上。

(2) 提出的磨粒图像分类 CNN 模型轻量化方法,在保证准确率的情况下,有效地减少参数数量和运算量,为 CNN 模型在移动式和便携式铁谱分析设备的应用提供思路。

参考文献

- [1] 彭鹏,陈李果,汪久根,等.机械磨损的检测技术综述[J].润滑与密封,2018,43(1):115-124.
PENG P, CHEN L G, WANG J G, et al. A review on detection technology of mechanical wear [J]. Lubrication Engineering, 2018, 43(1): 115-124.
- [2] 陈长征,白秉三,严安.设备故障智能诊断技术研究进展[J].沈阳工业大学学报,2000,22(4):349-352.
CHEN C Z, BAI B S, YAN A. Development of intelligent fault diagnosis [J]. Journal of Shenyang Polytechnic University, 2000, 22(4): 349-352.
- [3] 屈梁生,张海军.机械诊断中的几个基本问题[J].中国机械工程,2000,11(S1):211-216.
QU L S, ZHANG H J. Some basic problems in machinery diagnostics [J]. China Mechanical Engineering, 2000, 11(S1): 211-216.
- [4] 李炜,伦淑娴.设备故障诊断技术的现状及其发展[J].甘肃工业大学学报,1998,24(2):66-69.
LI W, LUN S X. Present situation and its advance of fault diagnostic technique for equipment [J]. Journal of Gansu University of Technology, 1998, 24(2): 66-69.
- [5] 吴今培.智能故障诊断技术的发展和展望[J].振动、测试与诊断,1999,19(2):79-86.
WU J P. Development and prospect of intelligent fault diagnosis [J]. Journal of Vibration, Measurement & Diagnosis, 1999, 19(2): 79-86.
- [6] 马佳丽,陈果,康玉祥,等.航空发动机磨损故障多目标融合诊断[J].航空动力学报,2024,39(10):20220191.
MA J L, CHEN G, KANG Y X, et al. Multi-objective fusion diagnosis of aeroengine wear fault [J]. Journal of Aerospace Power, 2024, 39(10): 20220191.
- [7] 程治升,费逸伟,杨宏伟,等.基于铁谱的航空发动机图谱技术研究[J].润滑与密封,2013,38(1):87-91.
CHENG Z S, FEI Y W, YANG H W, et al. Research on aero-engine atlas technology based on ferrography [J]. Lubrication Engineering, 2013, 38(1): 87-91.
- [8] 王汉功,陈桂明.铁谱图像分析理论与技术[M].北京:科学出版社,2005.
- [9] 张建,李艳军,曹愈远,等.免疫支持向量机用于航空发动机磨损故障诊断[J].北京航空航天大学学报,2017,43(7):1419-1425.
ZHANG J, LI Y J, CAO Y Y, et al. Immune SVM used in wear fault diagnosis of aircraft engine [J]. Journal of Beijing University of Aeronautics and Astronautics, 2017, 43(7): 1419-1425.
- [10] SCHMIDHUBER J. Deep learning in neural networks: an overview [J]. Neural Networks, 2015, 61: 85-117.
- [11] WANG J Q, LIU X L, WU M, et al. Direct detection of wear conditions by classification of ferrograph images [J]. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2020, 42(4): 152.
- [12] PENG Y P, CAI J H, WU T H, et al. A hybrid convolutional neural network for intelligent wear particle classification [J]. Tribology International, 2019, 138: 166-173.
- [13] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.
- [14] SUYKENS J A K. Support vector machines: a nonlinear modeling and control perspective [J]. European Journal of Control, 2001, 7(2/3): 311-327.
- [15] LIU X L, WANG J Q, SUN K, et al. Semantic segmentation of ferrography images for automatic wear particle analysis [J]. Engineering Failure Analysis, 2021, 122: 105268.
- [16] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.
- [17] SUN K, LIU X L, CHEN G N, et al. Wear debris recognition and quantification in ferrography images by instance segmentation [J]. Tribology Transactions, 2022, 65(3): 508-518.
- [18] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2980-2988.
- [19] 卿华,王新军.飞机油液监控技术[M].北京:航空工业出版社,2011.
- [20] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.
- [21] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high-performance deep learning library [C]//Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, CANADA, 2019.
- [22] CHETLUR S, WOOLLEY C, VANDERMERSCH P, et al. cuDNN: efficient primitives for deep learning [J]. Neural and Evolutionary Computing, ArXiv e-Prints, 2014: ArXiv:1410.0759.