DOI:10.12177/emca.2025.026 文章编号:1673-6540(2025)05-0513-14 中图分类号:TM 351 文献标志码:A

基于多重稀疏 MobileNetV2 的变压器故障诊断方法

刘 航, 史智予, 刘志坚*, 罗灵琳, 李 明, 牛 犇
(昆明理工大学 电力工程学院,云南 昆明 650500)

Transformer Fault Diagnosis Method Based on Multi-Level Sparse MobileNetV2

LIU Hang, SHI Zhiyu, LIU Zhijian*, LUO Linglin, LI Ming, NIU Ben
(Faculty of Electric Power Engineering, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: **[Objective]** Deep learning models are widely used in transformer fault diagnosis due to their ability to learn underlying data patterns and construct hierarchical feature representations. However, their massive number of parameters, complex network topology, and high calculation and storage costs limit their practical application in fault diagnosis of power transformers. **[Methods]** To address the above issues, this study proposed a transformer fault diagnosis method based on multi-level sparse MobileNetV2. First, spindle-shaped and hourglass-shaped blocks were used to compactly improve the inverted residual blocks of the MobileNetV2 model, reducing parameter number and computational complexity from the model structure itself to achieve preliminary model sparsity. Second, a group-level pruning method based on dependency graph model was proposed. The coupled parameters in the model were grouped, and a group-level pruning optimization strategy based on L2 norm was designed to perform sparse training and pruning fine-tuning. This process removed redundant structures and parameters in the model, further reducing parameter number and computational complexity and enhancing model sparsity. Finally, an 8-bit symmetric uniform quantization and quantization-aware training method was proposed. The 32-bit high-resolution floating-point parameters in the model were quantized into 8-bit low-resolution integer parameters. Building on this, model inference was performed to further reduce the computational complexity and achieve multi-level model sparsity. **[Results]** The results of numerical experiments and performance evaluations showed that compared with the original MobileNetV2 model, the improved multi-level sparse model proposed in this study achieved a fault identification accuracy of 95.2%, while reducing the parameter number, computational complexity, and model size by approximately 73.5%, 96.9%, and 68.8%, respectively. Moreover, the inference time for identifying 1 000 images was only 0.66 seconds. **[Conclusion]** The proposed method in this study effectively combines three types of individual sparsity methods: compact model improvement, model pruning, and parameter quantization. It achieves multi-level sparsity of deep learning models while maintaining high accuracy, effectively addressing the issue of over-parameterization caused by limited sample data in power transformer fault diagnosis and eliminating its adverse effects.

Key words: transformer; fault diagnosis; dependency graph; group-level pruning; quantization-aware training

摘 要: **【目的】**深度学习模型具有学习数据潜在规律和构建层次化特征表示的优势,因此被广泛应用于变压器故障诊断。然而,深度学习模型的参数量巨大、网络拓扑结构复杂且计算、存储成本高昂,限制了其在电力变压器故障诊断领域的应用。针对此问题,本文提出了一种多重稀疏 MobileNetV2 的变压器故障诊断方法。**【方法】**首先,使用纺锤块和沙漏块对 MobileNetV2 模型的倒残差块进行紧凑改进,从模型本身降低参数量和计算量,初步稀疏模型。其次,提出了一种基于依赖图模型的组级剪枝方法,对模型中的耦合参数进行分组,并基于 L2 范数设

基金项目: 云南省自然科学基金资助项目(202303AA080002,202401AT070356)
Project funded by Natural Science Foundation of Yunnan Province (202303AA080002, 202401AT070356)

计组级剪枝优化策略,对模型进行稀疏训练与剪枝微调,删除模型中的冗余结构和参数,进一步降低参数量和计算量,实现模型稀疏化。最后,提出 8 位宽对称均匀量化和量化感知训练方法,将模型中的 32 位宽高分辨率浮点参数量化为 8 位宽低分辨率整型参数,并在此基础上进行模型推理,再次降低计算量,实现模型的多重稀疏化。**【结果】**数值试验和性能评估结果表明:与 MobileNetV2 模型相比,本文提出的多重稀疏化 MobileNetV2 模型在将故障识别精度提升至 95.2% 的前提下,将参数量、计算量和模型大小分别降低了约 73.5%、96.9% 和 68.8%,且识别 1 000 张图片的推理时间仅为 0.66 秒。**【结论】**本文所提方法有效结合了紧凑改进、模型剪枝和参数量化三种单一稀疏化方法,在保证模型精度的前提下,实现了深度学习模型的多重稀疏化,较好地解决了电力变压器故障样本数据稀缺导致的模型过参数化问题,并消除了相关不利影响。

关键词: 变压器;故障诊断;依赖图;组级剪枝;量化感知训练

0　引言

电力变压器是实现电网中能量转换和传输的核心设备,具有在运数量大、分布范围广和工作时间长的特点,一旦发生故障,其故障影响与损失会远高于其他设备。因此,开展电力变压器的故障识别工作对电力系统的安全稳定运行至关重要[1]。油中溶解气体是反映变压器内部故障的重要特征量,当变压器内部发生过热或放电故障时,会导致故障点附近的绝缘油分解为低分子量气体(包括氢气、甲烷、乙炔、乙烷和乙烯等),并溶解于变压器油中。由于不同性质的故障产生的气体组分存在差异,可将其作为变压器故障识别的重要依据[2]。根据 IEEE C57.104-2019 中的标准,分为正常、放电兼过热、中低温、高温过热、低能和高能放电等 6 个类别[3]。

自 1950 年以来,基于油中溶解气体分析(Dissolved Gas Analysis, DGA)的变压器故障识别方法得到了快速发展及应用。根据其发展历史,可将故障识别方法归纳为传统方法、机器学习方法和深度学习方法三类。传统方法包括三比值法、编码法和大卫三角形法等[4],其简单高效且适用于故障数据稀缺场景,但由于油中溶解气体与故障类型映射关系复杂,易导致编码不全或故障比例失衡等问题。机器学习方法通过建立气体与故障类型的非线性映射关系来实现变压器的故障识别,目前应用广泛的机器学习方法有随机森林(Random Forest, RF)[5]、极限学习机[6]和支持向量机(Support Vector Machine, SVM)及其变体等[7-8]。例如,文献[9]基于样本集成学习和蛇优化算法优化 SVM 模型,提升了变压器故障诊断的精度。机器学习方法能够对故障样本进行快速识别,但依旧存在结构简单和学习非线性映射关系能力较弱等问题。

深度学习方法能够较好学习样本数据的内在规律和表示层次,其在变压器故障诊断领域得到了快速发展和广泛应用[10]。文献[11]提出融合分类模块提前筛选出可能被网络错误分类的样本进行单条数据分析,并在此基础上改进一维卷积神经网络(One-Dimensional Convolutional Neural Networks, 1D-CNN)实现变压器故障诊断。文献[12]提出基于软阈值改进的深度残差收缩网络和带可变权重的交叉熵函数,有效提升了变压器油色谱故障识别精度。由于深度学习模型的密集性和过参数化性,使其在计算机视觉开源大数据领域中应用广泛(如 CIFAR-100 和 ImageNet 等公用数据集),但针对少了 3~4 个数量级的变压器故障样本数据,直接应用深度学习模型会导致模型过拟合,影响模型性能[13]。因此,如果能对现有深度学习模型进行稀疏化,使原始模型的复杂度能够匹配样本数量和数据复杂度,就可以发挥深度学习模型的最佳性能。

目前对模型进行稀疏化的方法有紧凑网络、模型剪枝和参数量化等[14-16]。这些方法均为单一稀疏化方法,未针对特定的模型参数和结构进行较好结合,因此稀疏效果有限。在变压器故障诊断领域,基于深度学习模型的稀疏化的研究成果十分稀缺且应用较少。如文献[17]仅对视觉几何组 19(Visual Geometry Group 19, VGG19)和残差网络 50 (Residual Network 50, ResNet50)等传统卷积神经网络进行剪枝与量化并将其应用于变压器故障诊断。

本文提出了一种基于多重稀疏 MobileNetV2 模型的电力变压器故障诊断方法。首先,使用纺锤块和沙漏块对 MobileNetV2 中的倒残差块进行部分替换,初步优化模型;然后,提出基于依赖图

（Dependency Graph，DG）模型的组级剪枝方法，对模型进行剪枝与微调，删除模型中的冗余结构和参数；最后，提出 8 位宽对称均匀量化和量化感知训练（Quantization Aware Training，QAT）方法，将模型中每个 32 位宽网络浮点参数量化为 8 位宽整型参数，实现 MobileNetV2 模型的多重稀疏化。在 DGA 数据集上的试验结果表明，本文方法在提升变压器故障识别结果精度的前提下，实现了 MobileNetV2 模型的深度压缩加速和结构简化，有效解决了模型过参数化和过复杂化问题，提高了资源利用率并降低了能耗。

1 紧凑 MobileNetV2 模型

1.1 MobileNetV2 模型

Google 团队在 2018 年提出 MobileNetV2 深度学习模型[18]，MobilNetV2 模型由 1 个初始卷积层、17 个倒残差块、1 个平均池化层和 1 个线性层堆叠而成。由于 MobileNetV2 模型只能输入二维图像数据，故设输入数据的尺寸为 $D_{in}\times D_{in}$，卷积核尺寸为 $D_V\times D_V$，其输入、输出通道数分别为 C_{in} 和 C_{out}，输出数据尺寸为 $D_{out}\times D_{out}$。忽略偏置项时，MobileNetV2 模型的结构如图 1 所示。

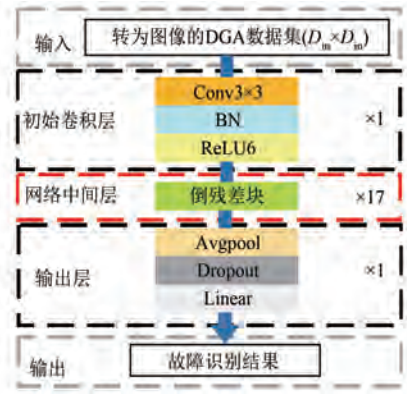

图 1 MobileNetV2 模型结构图

Fig. 1 Diagram of MobileNetV2 model structure

初始卷积层中卷积核的计算量 FP_{Conv} 和参数量 P_{Conv} 分别为

$$FP_{Conv}=(2\times C_{in}\times D_V^2-1)\times D_{out}^2\times C_{out} \quad (1)$$

$$P_{Conv}=C_{in}\times C_{out}\times D_V^2 \quad (2)$$

批量归一化（Batch Normalization，BN）层参数量 P_{BN} 为

$$P_{BN}=2\times C_{in} \quad (3)$$

倒残差块的计算量 FP_{Dwise} 和参数量 P_{Dwise} 分别为

$$FP_{Dwise}=(2\times D_V^2-1)\times C_{in}\times D_{out}^2+(2\times C_{in}-1)\times C_{out}\times D_{out}^2 \quad (4)$$

$$P_{Dwise}=D_V^2\times C_{in}+C_{in}\times C_{out} \quad (5)$$

假设线性层的输入、输出神经节点数分别为 N_{in}、N_{out}，则线性层的计算量 FP_{Linear} 和参数量 P_{Linear} 分别为

$$FP_{Linear}=(2\times N_{in}-1)\times N_{out} \quad (6)$$

$$P_{Linear}=N_{in}\times N_{out} \quad (7)$$

图 2 为不同模块结构图。MobileNetV2 模型采用了较为轻量的倒残差块，结构如图 2(a)所示。

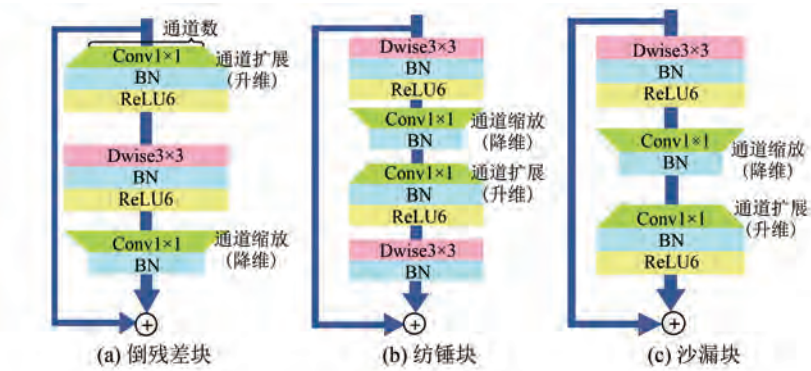

图 2 不同模块结构图

Fig. 2 Diagram of different module structures

倒残差块具有以下特点：①使用 1×1 标准卷积对输入的特征图进行升维，增加通道数；②使用 3×3 的深度卷积对每个通道独立进行卷积；③再次使用 1×1 标准卷积对特征图的独立通道特征进行融合并降维。

虽然 MobileNetV2 模型的结构相对简单，但

针对少量的变压器故障样本数据,该模型内部的结构和参数仍然显得冗余。因此,本文对其进行多重稀疏工作。

(1)MobileNetV2 模型的紧凑改进。针对模型结构进行优化改进,将倒残差块替换为更为简洁高效的纺锤块[19]和沙漏块[20],初步稀疏模型。

(2)MobileNetV2 模型的剪枝。针对模型中的冗余结构和参数进行改进,提出基于 DG 模型的组级剪枝方法[21]进行模型剪枝,进一步稀疏模型。

(3)MobileNetV2 模型的量化。针对模型中的参数分辨率进行改进,提出 8 位宽均匀对称量化和 QAT 方法[16]进行模型量化,再次稀疏模型。

1.2　紧凑 MobileNetV2 模型

针对倒残差块中存在的问题:①先升维后降维方式会增大参数量和计算量;②由于相邻倒残差块之间的表示维数较低,在瓶颈层之间构建短连接可能会阻止模型训练时的梯度传播[22]。

本文使用宽架构模块替换倒残差块,优势为:①在高维表示之间建立短连接,可以确保在特征提取时保留更多有效信息并促进梯度的跨层传播;②在高维表示中适当应用 3×3 的轻量级深度卷积,可以在不增加计算量和参数量的前提下,保留更多数据特征。

假设 $\boldsymbol{I}\in\mathbf{R}^{D_V\times D_V\times C_{in}}$ 为输入张量,$\boldsymbol{O}\in\mathbf{R}^{D_V\times D_V\times C_{out}}$ 为输出张量,在不考虑深度卷积和激活层的前提下,宽架构模块的数学表达式为

$$\boldsymbol{O}=\varphi_e[\varphi_r(\boldsymbol{I})]+\boldsymbol{I} \tag{8}$$

式中:φ_e 和 φ_r 分别为通道扩展和通道压缩的 2 个 1×1 标准卷积。

两种宽架构模块的基本结构分别如图 2(b)和图 2(c)所示。

纺锤块由 1 个 3×3 深度卷积,2 个 1×1 标准卷积、2 个 BN 层和 2 个 ReLU6 激活函数构成。纺锤块中的 2 个 1×1 标准卷积为先降维再升维。

在纺锤块的基础上加入 1 个 3×3 深度卷积置于模块尾部,形成沙漏块[20],该模块的数学表达式为

$$\boldsymbol{O}_1=\varphi_{1,p}[\varphi_{1,d}(\boldsymbol{I})] \tag{9}$$

$$\boldsymbol{O}_2=\varphi_{2,p}[\varphi_{2,d}(\boldsymbol{O}_1)]+\boldsymbol{I} \tag{10}$$

式中:$\varphi_{1,p}$、$\varphi_{2,p}$ 和 $\varphi_{1,d}$、$\varphi_{2,d}$ 分别为第 1、2 次 1×1 标准卷积和深度卷积;$\boldsymbol{O}_1$、$\boldsymbol{O}_2$ 分别为第 1、2 次输出张量。

本文基于纺锤块与沙漏块对 MobileNetV2 模型中的倒残差块进行部分替换,提出了一种简洁高效的紧凑 MobileNetV2 模型,该模型的结构如图 3 所示。

由图 3 可知,紧凑 MobileNetV2 模型由 1 个初始卷积层、5 个沙漏块、6 个倒残差块、7 个纺锤块、1 个平均池化层和 1 个全连接层堆叠而成。

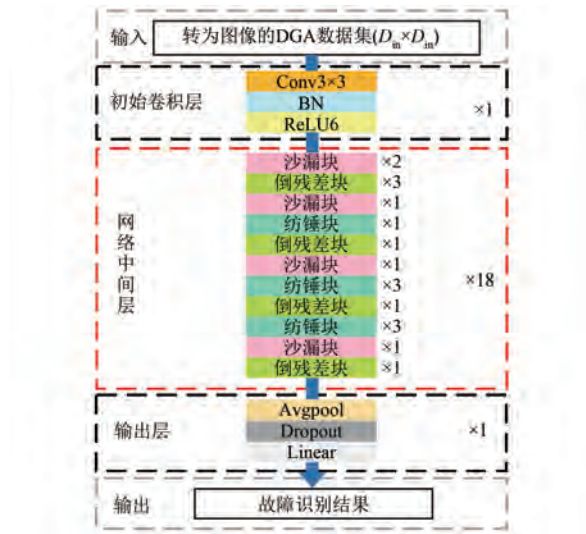

图 3　紧凑 MobileNetV2 模型结构图

Fig. 3　Diagram of compact MobileNetV2 model structure

1.3　不同模型的复杂性评价分析结果

根据式(1)~式(7),模型整体的计算量 FP_{All} 和参数量 P_{All} 由每一层累加而成,假设输入数据为一张 1×3×32×32 的图片,卷积核的输入和输出通道数分别为 32 和 64。则可计算出相关模型的复杂度大小,结果如表 1 所示。

表 1　不同模块和模型的参数量和计算量

Tab. 1　Number of parameters and computational complexity for different modules and models

模块、模型名称	计算量/万次	参数量/万个
倒残差块	130.46	1.06
沙漏块	56.93	0.28
纺锤块	52.43	0.20
MobileNetV2 模型	651.70	223.15
紧凑 MobileNetV2 模型	565.73	160.67

由表 1 可知,沙漏块和纺锤块的计算量和参数量较倒残差块明显减少;紧凑 MobileNetV2 模型的计算量和参数量分别为MobileNetV2模型的

86.8% 和 72%。

2 模型剪枝及 8 位宽量化方法

2.1 基于 DG 模型的组级剪枝方法

模型剪枝是指剔除网络模型中的冗余结构和参数，降低模型参数量和计算量。由于深度学习模型内部结构的耦合性和复杂性，因此需对模型中的耦合层进行同时剪枝，确保建立短连接的两个层级输出通道一致[23]。针对紧凑 MobileNetV2 模型中的深度卷积层进行剪枝，则需对与其耦合的标准卷积层和两个 BN 层进行同时剪枝。

本文提出一种基于 DG 模型的组级剪枝方法，如图 4 所示。该方法首先基于 DG 模型对模型中的耦合依赖参数进行自动分组，然后基于 L2 范数设计组级剪枝方法进行同时剪枝与微调。

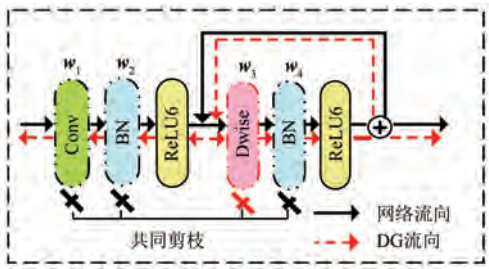

图 4 基于 DG 模型的组级剪枝方法

Fig. 4 Group-level pruning method based on DG model

2.1.1 DG 模型

本文提出了一个测量相邻层之间相互依赖性的模型—DG[21]。对于任意深度学习模型 $F(\boldsymbol{x},\boldsymbol{w})$，模型中第 i+1 层的输出 $\boldsymbol{y}_{i+1}$ 的表达式为

$$\boldsymbol{y}_{i+1}=F(\boldsymbol{w}_i\boldsymbol{x}_i+\boldsymbol{b}) \tag{11}$$

式中：$\boldsymbol{x}_i$ 为第 i 层的输入；$\boldsymbol{w}_i$ 为第 i 层的权值向量；$\boldsymbol{b}$ 为偏置向量。

假设 F 中存在参数组 $\boldsymbol{w}_g=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_i\}$，其相互具有依赖关系，即 $\boldsymbol{w}_1\Leftrightarrow\boldsymbol{w}_2\Leftrightarrow\cdots\Leftrightarrow\boldsymbol{w}_i$，则可对 $\boldsymbol{w}_g$ 进行分解，如式(12)所示：

$$\boldsymbol{w}_g=\{\boldsymbol{f}_1^-,\boldsymbol{f}_1^+,\boldsymbol{f}_2^-,\boldsymbol{f}_2^+,\cdots,\boldsymbol{f}_i^-,\boldsymbol{f}_i^+\} \tag{12}$$

式中：$\boldsymbol{f}_i^-$ 和 $\boldsymbol{f}_i^+$ 分别为 $\boldsymbol{w}_i$ 的输入和输出。

故可以将 $\boldsymbol{w}_g$ 分解为两种依赖关系：层间依赖和层内依赖。以图 3 为例，将 $\boldsymbol{w}_g$ 中的依赖关系展开，模型结构关系如图 4 和式(13)所示：

$$(\boldsymbol{f}_1^-,\boldsymbol{f}_1^+)\leftrightarrow(\boldsymbol{f}_2^-,\boldsymbol{f}_2^+)\cdots\leftrightarrow(\boldsymbol{f}_i^-,\boldsymbol{f}_i^+) \tag{13}$$

式中：↔为两个相邻层间的联通性，其中具有依赖关系的两个层 $\boldsymbol{f}_i^+\Leftrightarrow\boldsymbol{f}_j^-$ 相联通，$\boldsymbol{f}_i^+\leftrightarrow\boldsymbol{f}_j^-$ 称为层间依赖。

在第 i 层中具有依赖关系的输入和输出 $\boldsymbol{f}_i^+\Leftrightarrow\boldsymbol{f}_j^-$ 需共同剪枝，记为 $sch(\boldsymbol{f}_i^-)=sch(\boldsymbol{f}_i^+)$，称为层内依赖。基于 DG 模型的组级剪枝传播过程如图 5 所示。

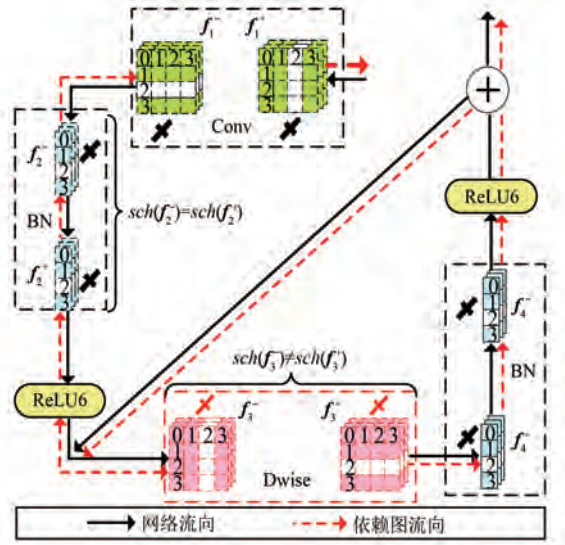

图 5 基于 DG 模型的组级剪枝传播过程

Fig. 5 Propagation process of group-level pruning based on DG model

相联通的两个层 $\boldsymbol{f}_i^+\leftrightarrow\boldsymbol{f}_j^-$，由于对应同一个中间特征(即待剪枝的权重张量的行或列相同)，因此层间依赖关系始终存在，如图 5 中的 $\boldsymbol{f}_1^+$ 和 $\boldsymbol{f}_2^-$ 及 $\boldsymbol{f}_3^+$ 和 $\boldsymbol{f}_4^-$ 均具有层间依赖关系。层内依赖关系分为两种剪枝方法：①对于 BN 层，如图 5 中的 $\boldsymbol{f}_2^-$ 和 $\boldsymbol{f}_2^+$ 及 $\boldsymbol{f}_4^-$ 和 $\boldsymbol{f}_4^+$，其输入和输出相同，故具有层内依赖关系，需同时进行剪枝，即 $sch(\boldsymbol{f}_i^-)=sch(\boldsymbol{f}_i^+)$；②对于卷积层，如图 5 中的 $\boldsymbol{f}_1^-$ 和 $\boldsymbol{f}_1^+$ 及 $\boldsymbol{f}_3^-$ 和 $\boldsymbol{f}_3^+$，其输入和输出不等则不具有层内依赖关系，需分别进行剪枝，即 $sch(\boldsymbol{f}_i^-)\neq sch(\boldsymbol{f}_i^+)$。综上，可以看出 $\boldsymbol{f}_2^-$ 和 $\boldsymbol{f}_2^+$、$\boldsymbol{f}_4^-$ 和 $\boldsymbol{f}_4^+$、$\boldsymbol{f}_1^+$ 和 $\boldsymbol{f}_2^-$ 及 $\boldsymbol{f}_3^+$ 和 $\boldsymbol{f}_4^-$ 均具有依赖关系，故可以同时对这些层进行剪枝。

因此，DG 模型的数学表达为

$$\begin{aligned}DG(\boldsymbol{f}_i^-,\boldsymbol{f}_i^+)=I_d[\boldsymbol{f}_i^+\leftrightarrow\boldsymbol{f}_j^-]\ \vee\\ I_d[i=j\wedge sch(\boldsymbol{f}_i^-)=sch(\boldsymbol{f}_i^+)]\end{aligned} \tag{14}$$

式中：∨和∧分别为逻辑“或”和“与”；I_d 为一个指示函数，表示返回“真”条件成立；$I_d[\boldsymbol{f}_i^+\leftrightarrow\boldsymbol{f}_j^-]$ 为判断第 i 层和第 j 层是否具有层间依赖；$I_d[i=j\wedge sch(\boldsymbol{f}_i^-)=sch(\boldsymbol{f}_i^+)]$ 为判断具有层内依赖关系的第 i 层输入和输出是否共享相同剪枝方法。

2.1.2　组级剪枝优化策略

对网络参数进行分组后,需评估涉及耦合层的分组参数的重要性以进行组级剪枝。本文根据 L2 范数设计出一种组级剪枝优化策略。利用 L2 范数 $I(\boldsymbol{w})=\|\boldsymbol{w}\|_2$ 为每个权值 $\boldsymbol{w}(\boldsymbol{w}\in\boldsymbol{w}_g)$ 产生独立的分数,则可通过计算每个 $\boldsymbol{w}_g$ 的总分数 $I(\boldsymbol{w}_g)=\sum_{\boldsymbol{w}\in\boldsymbol{w}_g}I(\boldsymbol{w})$ 来估计其重要性。但若直接对模型进行剪枝可能会导致模型中某些层的梯度消失或爆炸,从而影响模型的训练收敛性和最终性能[15]。

因此,本文提出一种稀疏训练方法对组级参数进行稀疏化。具体是对于每个 k 维的 $\boldsymbol{w}$[索引为 $\boldsymbol{w}(k)$],在稀疏训练时,引入一个简单的正则化项,如式(15)所示:

$$R(\boldsymbol{w}_g,k)=\sum_{k=1}^{K}\gamma_k\cdot I_{\boldsymbol{w}_g,k}=\sum_{k=1}^{K}\sum_{\boldsymbol{w}\in\boldsymbol{w}_g}\gamma_k\|\boldsymbol{w}(k)\|_2^2 \tag{15}$$

式中:$I_{\boldsymbol{w}_g,k}=\sum_{\boldsymbol{w}\in\boldsymbol{w}_g}\|\boldsymbol{w}(k)\|_2^2$ 为第 k 维的重要性分数;γ_k 为收缩强度,其表达式为

$$\gamma_k=2^{\theta_k}=2^{\theta(I_{\boldsymbol{w}_g}^{\max}-I_{\boldsymbol{w}_g,k})/(I_{\boldsymbol{w}_g}^{\max}-I_{\boldsymbol{w}_g}^{\min})} \tag{16}$$

式中:θ_k 为控制 γ_k 的参数,$\theta_k\in[2^0,2^\theta]$,在本文中所有试验使用恒定超参数 $\theta=4$;$I_{\boldsymbol{w}_g}^{\max}$ 和 $I_{\boldsymbol{w}_g}^{\min}$ 分别为参数组的最大和最小重要性分数。

稀疏训练后,本文使用相对分数 $\hat{I}_{\boldsymbol{w}_g,k}$ 来识别和剪枝网络中不重要的参数,其表达式为

$$\hat{I}_{\boldsymbol{w}_g,k}=n\cdot I_{\boldsymbol{w}_g,k}/\sum[Topn(I_{\boldsymbol{w}_g})] \tag{17}$$

式中:n 为组数;$Topn(I_{\boldsymbol{w}_g})$ 为前 n 个参数组的重要性分数。

在剪枝过程中,根据各组的重要性分数对其排序,同时引入稀疏率,根据其比率大小,删除比率相对靠后的参数。

模型剪枝流程如图 6 所示,关键步骤如下。

(1)稀疏训练:首先基于 DGA 训练集对模型进行稀疏训练,其目的是提前模拟并适应模型剪枝引入的截断误差,以提升模型泛化能力,保持模型性能;

(2)模型剪枝:在模型稀疏训练的基础上,对模型进行剪枝;

(3)微调模型:基于稀疏训练中最优的模型权重,对剪枝后的模型在 DGA 训练集上进行微调,以恢复模型性能。

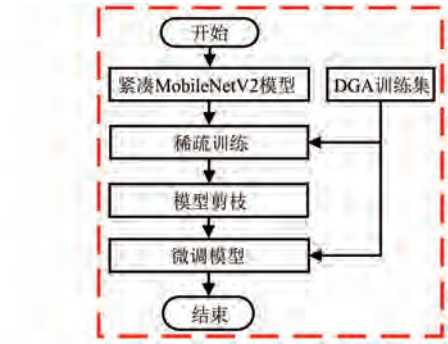

图 6　模型剪枝流程

Fig. 6　Model pruning workflow

2.2　8 位宽对称均匀量化方法

参数量化是指将卷积神经网络中 32 位宽高精度浮点参数转换成低精度整型参数(如 8 位宽、4 位宽和 2 位宽)参与网络计算。目前,通用的网络训练方法和硬件平台不支持 2 位宽和 4 位宽参数量化场景,必须单独设计专用的软硬件系统框架[14]。因此,本文采用 8 位宽量化方法和 QAT 方法降低紧凑 MobileNetV2 模型的计算量。

2.2.1　8 位宽对称均匀量化

8 位宽量化是指通过某种量化算子 Q 将深度学习模型中连续分布的 32 位宽浮点参数(权值向量 $\boldsymbol{w}$ 和激活值向量 $\boldsymbol{v}$)映射为离散分布的 8 位宽整型参数[16]。以网络中 $\boldsymbol{w}_i$ 的 8 位宽量化为例,其计算方法如式(18)和式(19)所示:

$$S=\frac{\beta-\alpha}{2^b-1} \tag{18}$$

$$\widetilde{\boldsymbol{w}}_i=Q(\boldsymbol{w}_i)=Int\left(\frac{\boldsymbol{w}_i}{S}\right)-\boldsymbol{Z} \tag{19}$$

式中:S 为量化步长;β 和 α 分别为 $\boldsymbol{w}_i$ 中最大值和最小值;b 为量化位宽,本文取 $b=8$;Int 算子为舍入操作;$\boldsymbol{Z}$ 为零点偏移向量,对称均匀量化时其值为 0;$\widetilde{\boldsymbol{w}}_i$ 为量化后的权值向量,其值域为 $[\tilde{\alpha},\tilde{\beta}]=[-127,127]$。

本文提出的 8 位宽对称量化原理如图 7 所示。

$\widetilde{\boldsymbol{w}}_i$ 的值组成的集合$\{(-2^7+1)S,\cdots,-S,0,S,\cdots,(2^7-1)S\}$为量化电平,其与 8 位有符号整型参数的范围对应。若量化电平之间间隔 d 相等则称为均匀量化,如图 7(b)所示。

在卷积神经网络训练的过程中,式(16)中的

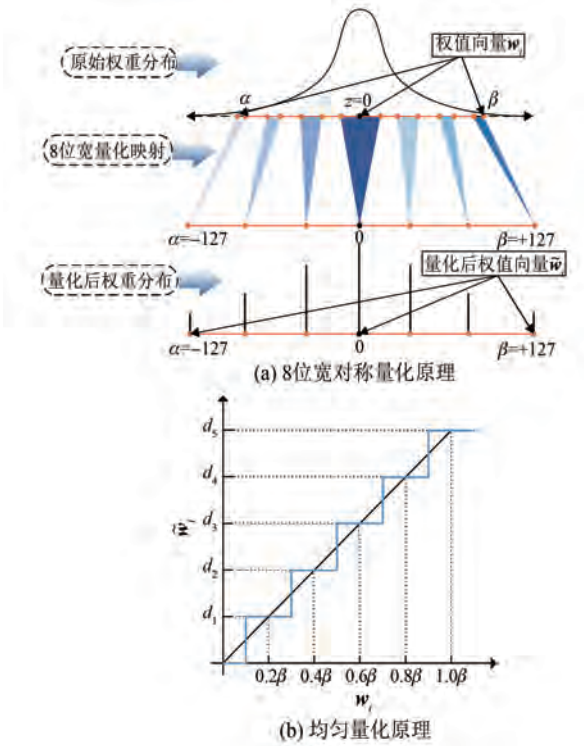

图 7　量化原理

Fig. 7　Quantization principles

Int 算子会导致反向传播的参数梯度为 0，导致模型无法训练，因此需将 $\tilde{\boldsymbol{w}}_i$ 反量化回 32 位宽浮点参数以更新权重，数学表达式为

$$\hat{\boldsymbol{w}}_i = DQ(\tilde{\boldsymbol{w}}_i) = S(\tilde{\boldsymbol{w}}_i + \boldsymbol{Z}) \tag{20}$$

式中：$\hat{\boldsymbol{w}}_i$ 为反量化后的权值向量；*DQ* 为反量化算子。

$\hat{\boldsymbol{w}}_i$ 与 $\boldsymbol{w}_i$ 之间存在一定的误差，该误差可以看作对原始值的噪声干扰，会使网络模型从原来训练收敛的局部最优点向更差的方向偏移，降低模型性能。

2.2.2　QAT

QAT 过程如图 8 所示。插入 *Q* 和 *DQ* 参与模型训练，将输入的 32 位浮点参数量化为 8 位整型参数参与模型的前向传播，再对其反量化回 32 位浮点参数参与模型的后向传播。在训练优化过程中 *Q* 和 *DQ* 可以学习量化和反量化的尺度信息，更新最优量化步长 *S* 和对应的 $\tilde{\boldsymbol{w}}_i$ 的值域。

2.3　故障识别流程

基于多重稀疏 MobileNetV2 模型的轻量化变压器故障识别流程如图 9 所示，关键步骤如下。

（1）数据转化：采用格拉姆角场与数据增强方法将一维 DGA 样本数据集扩充为 DGA 图像数据集，并划分测试集与训练集。

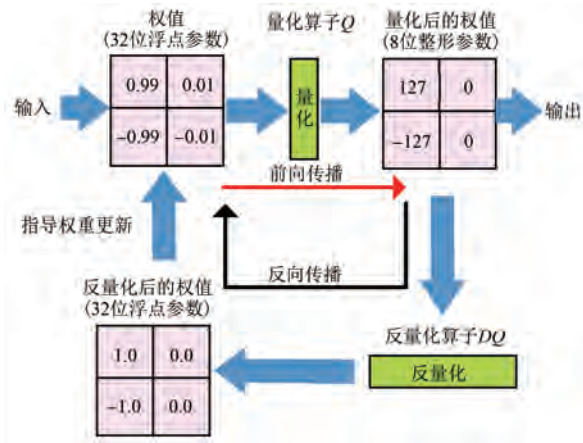

图 8　QAT 过程

Fig. 8　QAT process

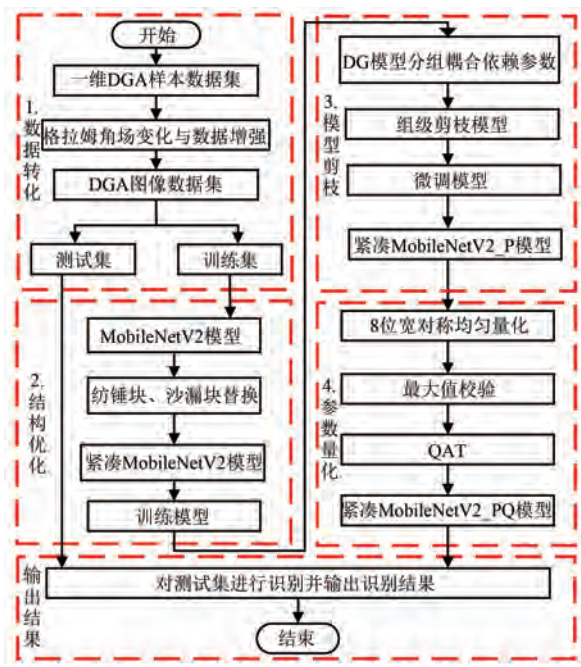

图 9　多重稀疏化变压器故障识别流程

Fig. 9　Fault identification workflow of multi-level sparse transformer

（2）结构优化：将 MobileNetV2 模型中的倒残差块替换为纺锤块与沙漏块，得到紧凑 MobileNetV2 模型，并使用 DGA 训练集对其进行训练，在测试集中验证模型性能。

（3）模型剪枝：基于 DG 模型对紧凑 MobileNetV2 模型中的耦合依赖参数进行自动分组后开展稀疏训练工作，进行组级剪枝并微调，得到紧凑 MobileNetV2_P 模型。

（4）参数量化：插入 8 位宽对称均匀量化器

与最大值校验器并进行 QAT,得到紧凑 MobileNetV2_PQ 模型,完成模型的多重稀疏化,最后在测试集中验证模型性能。

3　试验及结果分析

3.1　数据构成、算例介绍及试验设置

本试验所用的 DGA 数据集(包含 1 624 条一维样本数据)来源于某变电站采集的现场数据和部分文献中使用的数据[24]。根据 IEEE C57.104-2019 标准,将其分为 6 类,包括正常类别和 5 类故障类别(高能放电、低能放电、高温过热、中低温过热和放电兼过热)。

由于紧凑 MobileNetV2 只能输入二维数据,因此本文参考文献[25]的数据转换和增强方法,将 1 624 条一维 DGA 数据集转化成 9 750 张分辨率为 32×32 的增强图像数据集。相较于含有 1 400 万张图片和 2 万个类别的 ImageNet 数据集,本文数据集属于少样本数据集。6 种类别样本图如图 10 所示,图中每张图片的分辨率为 370×370。

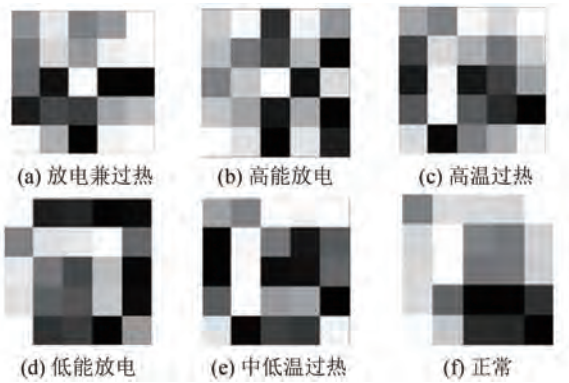

图 10　6 种类别样本图

Fig. 10　Sample diagrams of six categories

按 8:2将原始数据划分为训练集和测试集,其详细信息如表 2 所示。

表 2　DGA 数据集划分表

Tab. 2　DGA dataset division

样本类别	训练集/张	验证集/张
正常	1 109	277
高能放电	1 681	419
低能放电	1 172	292
高温过热	2 319	597
中低温过热	1 273	317
放电兼过热	178	44

本文试验采用 PyTorch 构建,试验运行环境为搭载了 Intel 酷睿 i5 12400 的中央处理器和 NVIDIA Geforce RTX4070 图形处理器的计算机。Python 版本为 3.9,PyTorch 版本为 1.13,Cuda 版本为 10,TensorRT 版本为 8.6。

采用随机梯度下降法与交叉熵损失函数进行模型训练,其中预训练、稀疏训练、剪枝微调和 QAT 轮数均为 500,初始学习率为 0.001,所有深度学习模型均为初始设置,即默认动量为 0.9,无 Dropout,权重衰减为 5e-4。为了更加稳定地训练模型并加快模型收敛速度,使用余弦退火方法[26]来更新学习率,其数学表达式为

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left[1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right] \tag{21}$$

式中:η_t 为当前学习率;$\eta_{\min}$ 和 $\eta_{\max}$ 分别为最小学习率和最大学习率;T_{cur} 为当前迭代次数;$T_{\max}$ 为最大迭代次数。当模型训练迭代次数达 $T_{\max}$ 时,η_t 会重新回到最大学习率。

采用模型大小、参数量、计算量、识别 1 000 张图片的推理时间和模型精度进行性能评估。模型精度 A 的表达式为

$$A = \frac{T_p}{T_p + F_p} \times 100\% \tag{22}$$

式中:T_p、F_p 分别为正确分类、错误分类的样本总数。

3.2　紧凑 MobileNetV2 模型性能评估

本节开展 6 项消融试验以验证纺锤块和沙漏块对 MobileNetV2 模型性能的影响。在 DGA 数据集上进行相同的预训练,训练过程如图 11 所示。

由图 11(a)和图 11(b)可以看出,相比其他 MobileNetV2 模型,本文所提紧凑 MobileNetV2 模型在预训练阶段正确率较高且损失较小,其性能更接近经典深度学习模型:VGG19 和 ResNet50。

6 种模型在 DGA 数据集上的试验结果如表 3 所示。由表 3 可知,MobileNetV2 模型的参数量、计算量和模型大小远低于 VGG19 和 ResNet50,精度为 95.1%;仅加入纺锤块的纺锤 MobileNetV2 模型计算量最小仅为 5.34 M,但参数量较大为 2.17 M,且相较于 MobileNetV2 模型其精度略微降低;仅加入沙漏块的沙漏 MobileNetV2 模型较 MobileNetV2 模型,其参数量、计算量和模型大小分别下降至 1.93 M、6.45 M 和 7.69 MB,同时精度

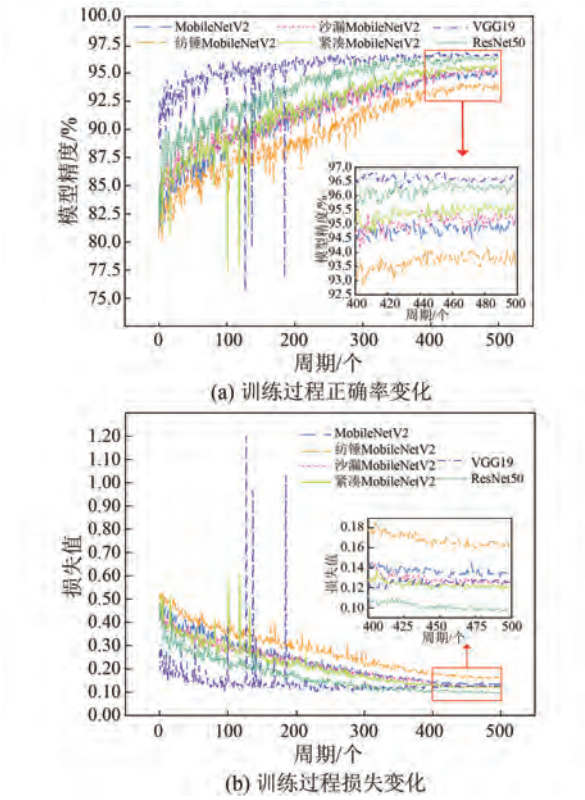

图 11　6 种模型训练过程

Fig. 11　Training processes of 6 models

有小幅提升；而本文所提紧凑 MobileNetV2 模型较 MobileNetV2 模型，其参数量、计算量和模型大小分别下降至 1.61 M、5.66 M 和 6.35 MB，同时精度提升最高，达 95.7%。

表 3　6 种模型在 DGA 数据集上的试验结果

Tab. 3　Experimental results of 6 models on DGA dataset

故障识别模型	模型精度/%	模型大小/MB	参数量/M	计算量/M
VGG19	96.9	148.00	38.93	418.08
ResNet50	96.6	145.00	38.10	1495.61
MobileNetV2	95.1	8.75	2.23	6.52
纺锤 MobileNetV2	94.1	8.46	2.17	5.34
沙漏 MobileNetV2	95.5	7.69	1.93	6.45
紧凑 MobileNetV2	95.7	6.35	1.61	5.66

基于图 10 的 6 种类别样本图，分别对 MobileNetV2 模型和紧凑 MobileNetV2 模型的中间层特征图进行可视化，其结果如图 12 和 13 所示，图中图像分辨率大小为 369×462。

由图 12 和图 13 可知，MobileNetV2 模型中间层输出的各个类别特征图之间差异性较小；而紧凑 MobileNetV2 模型中间层输出的各个类别特征图之间差异性较大。例如，图 13(b)中间部分像素点呈现出大面积黑点与其他类别相比更具差异性。

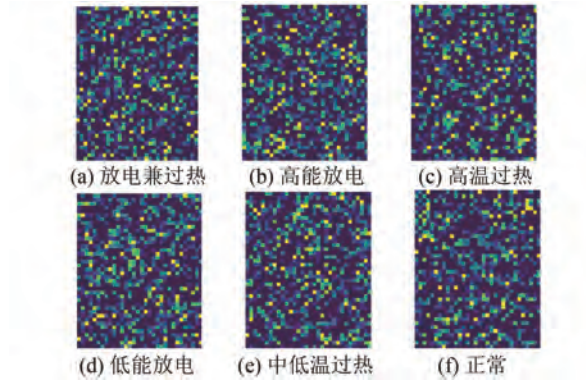

图 12　MobileNetV2 模型中间层特征图

Fig. 12　Feature maps of intermediate layers in MobileNetV2 model

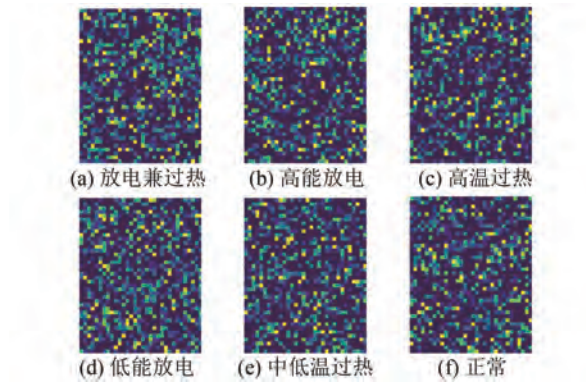

图 13　紧凑 MobileNetV2 模型中间层特征图

Fig. 13　Feature maps of intermediate layers in compact MobileNetV2 model

综上可见，本文所提紧凑 MobileNetV2 模型优化了 MobileNetV2 模型的原始结构，提升了模型性能并降低了模型复杂度。

3.3　DG 剪枝方法及其性能评估

本试验对紧凑 MobileNetV2 模型先进行稀疏训练，再进行剪枝与微调得到紧凑 MobileNetV2_P 模型。在 DGA 数据集上，稀疏率 R 从 0~0.9 的变化时，模型稀疏训练与剪枝微调的过程如图 14 所示。

由图 14 可知，随 R 增大，微调精度呈现先降后升再持续下降的趋势，约在 $R=0.4$ 后，模型微调精度急剧下降且微调损失值急剧上升。

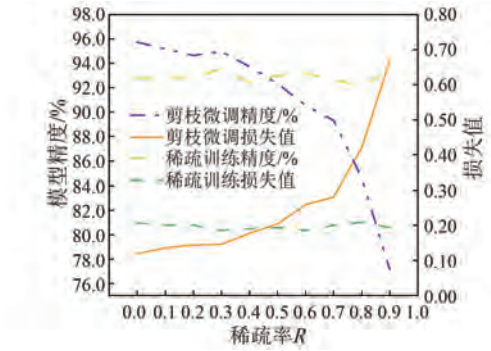

图 14　不同 R 下稀疏训练与剪枝微调变化过程

Fig. 14　Variation process of sparse training and pruning fine-tuning under different R values

紧凑 MobileNetV2_P 模型在不同 R 下的试验结果如表 4 所示。

表 4　紧凑 MobileNetV2_P 模型在不同 R 下的试验结果

Tab. 4　Experimental results of compact MobileNetV2_P model under different R values

稀疏率 R	模型精度/%	模型大小/MB	参数量/M	计算量/M
0.0	95.7	6.35	1.61	5.66
0.1	95.2	5.18	1.30	4.59
0.2	94.7	4.14	1.03	3.70
0.3	95.0	3.23	0.80	2.92
0.4	93.7	2.43	0.59	2.22
0.5	92.3	1.77	0.42	1.65
0.6	90.5	1.20	0.27	1.09
0.7	89.3	0.75	0.16	0.67
0.8	84.6	0.43	0.07	0.36
0.9	77.1	0.21	0.02	0.13

由表 4 可知,随 R 增大,模型整体指标均呈现下降趋势。其中,$R\in[0,\ 0.4]$ 时,模型大小、参数量和计算量下降较快,且剪枝对模型精度的影响较小;$R=0.3$ 时相较于 $R=0.2$ 时模型精度反而有所提升。$R\in[0.5,0.9]$ 时,模型精度下降迅速且模型大小、参数量和计算量下降减缓,其中 $R=0.9$ 时模型精度最低,为 77.1%。

综上,$R=0.4$ 时能保证紧凑 MobileNetV2_P 模型在模型精度与复杂度之间做到较好平衡,精度达 93.7%,模型大小、参数量和计算量较 MobileNetV2 模型分别减少约 72.2%、73.5% 和 66%。

为验证本文提出的组级剪枝策略,选取 $R=0.4$ 时的几何中值的过滤器剪枝(Filter Pruning via Geometric Median, FPGM)策略、基于层自适应幅度的剪枝(Layer-Adaptive Magnitude-based Pruning, LAMP)策略、网络瘦身(Network Slimming,NS)策略以及本文所提剪枝策略进行对比,结果如表 5 所示。

表 5　$R=0.4$ 时不同剪枝策略对比

Tab. 5　Comparison of different pruning strategies at $R=0.4$

剪枝策略	模型精度/%	模型大小/MB	参数量/M	计算量/M
FPGM[27]	92.1	2.38	0.58	2.21
LAMP[28]	92.8	2.43	0.59	2.22
NS[29]	91.4	2.43	0.59	2.22
本文策略	93.7	2.43	0.59	2.22

由表 5 可知,本文所提组级剪枝策略与其他剪枝策略相比,在模型大小、参数量和计算量减少基本一致的情况下,模型精度最高,达 93.7%。

文献[30]及表 4 的试验结果表明,对模型进行稀疏化操作后,减小了模型复杂度且提升了精度,这是因为模型稀疏化操作属于截断操作,通过删除模型中的冗余参数和结构,减少了模型过拟合问题,从而让稀疏后模型的复杂度匹配样本数量和数据复杂度,提升模型性能。

3.4　8 位宽对称均匀量化方法及其性能评估

对紧凑 MobileNetV2_P 模型进行 8 位宽对称均匀量化得到紧凑 MobileNetV2_PQ 模型。试验在 TensorRT 框架下,使用 DGA 数据集对各模型进行 QAT,采用最大值校验器,最后进行性能评估。紧凑 MobileNetV2_PQ 模型在不同 R 下的试验结果如表 6 所示。

由表 6 可知,$R\in[0,\ 0.4]$ 时,随 R 增大紧凑 MobileNetV2 _ PQ 模型精度下降较慢;$R\in[0.5,0.9]$ 时,随 R 增大模型精度下降迅速,最低为 79.7%。相较于紧凑 MobileNetV2_P 模型,MobileNetV2_PQ 模型的精度有所提高,原因是:①QAT 本身就是微调优化的过程;②插入最大值校验器后会动态优化量化步长 S,以此优化模型。MobileNetV2_PQ 模型大小较 MobileNetV2_PQ 模型更大的原因是:进行量化时,TensorRT 框架自带

一定的参数和结构(量化器、最大值校验器等)，导致模型大小增加。

表 6 紧凑 MobileNetV2_PQ 模型在不同 R 下的试验结果

Tab. 6 Experimental results of compact MobileNetV2_PQ model under different R values

R	模型精度/%	模型大小/MB	参数量/M	计算量/M
0.0	96.1	3.71	1.61	0.33
0.1	96.2	3.44	1.30	0.30
0.2	95.6	3.24	1.03	0.26
0.3	95.7	3.06	0.80	0.23
0.4	95.2	2.73	0.59	0.20
0.5	94.4	2.69	0.42	0.17
0.6	93.0	2.57	0.27	0.13
0.7	91.1	2.36	0.16	0.10
0.8	86.3	2.17	0.07	0.06
0.9	79.7	1.81	0.02	0.03

$R=0.4$ 时，MobileNetV2_PQ 模型 QAT 过程如图 15 所示。由图 15 可以看出，模型收敛较快，约在第 400 个周期后模型精度趋于平稳。

综上，在 $R=0.4$ 时，紧凑 MobileNetV2_PQ 模型相较于紧凑 MobileNetV2_P 模型，精度提升了 1.5%，计算量仅为 MobileNetV2_P 模型的 9.0%；相较于 MobileNetV2 模型，精度提升 0.1%，模型大小、参数量和计算量约分别降至 MobileNetV2 模型的 31.2%、26.5% 和 3.1%。

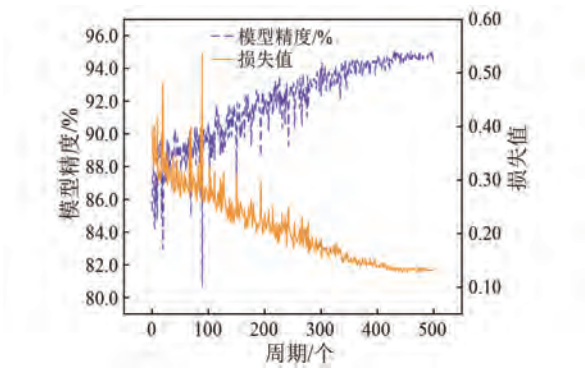

图 15 $R=0.4$ 时 QAT 过程

Fig. 15 QAT process at $R=0.4$

在 $R=0.4$ 时，将本文所提 QAT 方法和训练后量化(Post-Training Quantization，PTQ)[14] 方法进行对比，结果如表 7 所示。

由表 7 可知，相较于 PTQ 方法，本文所提 QAT 方法在模型计算量相同的情况下，模型精度更高。

表 7 在 $R=0.4$ 时的不同量化方法对比

Tab. 7 Comparison of different quantization methods at $R=0.4$

量化方法	模型精度/%	模型大小/MB	参数量/M	计算量/M
PTQ[14]	93.0	2.73	0.59	0.20
QAT	95.2	2.73	0.59	0.20

3.5 本文方法与其他故障诊断方法的对比试验

基于一维变压器油色谱故障样本数据集，选取与本文模型大小相似的传统三比值法[4]、RF[5] 和 1D-CNN[11] 进行对比试验。同时为验证本文方法的通用性，选取三个深度学习模型(VGG19、ResNet50、MobileNetV2)与本文模型进行剪枝和量化的对比试验，取 $R=0.4$，用 DGA 数据集对其分别多重稀疏化后，得到四种稀疏化模型：VGG19_PQ[17]、ResNet50 _ PQ[17]、MobileNetV2 _ PQ[18] 和本文所提紧凑 MobileNetV2_PQ。试验结果如表 8 所示。

表 8 本文方法与传统方法、机器学习方法和其他深度学习方法对比

Tab. 8 Comparison of the proposed method with traditional methods, machine learning methods, and other deep learning methods

模型名称	模型精度/%	模型大小/MB	参数量/M	计算量/M	推理时间/s
三比值法[4]	24.0	/	/	/	/
RF[5]	84.0	1.29	/	/	/
1D-CNN[11]	80.6	2.73	/	/	/
VGG19_PQ[17]	97.4	15.00	14.00	0.62	1.75
ResNet50_PQ[17]	94.0	16.20	13.68	1.69	4.45
MobileNetV2_PQ[18]	94.4	2.87	0.83	0.24	0.76
紧凑 MobileNetV2_PQ	95.2	2.73	0.59	0.20	0.66

由表 8 可知，相较于传统方法和机器学习方法，本文所提紧凑 MobileNetV2_PQ 模型精度较高，可达 95.2%；模型大小也较小，为 2.73 MB。相较于其他深度学习模型，本文所提模型的模型精度仅次于 VGG19_PQ 模型；但参数量和计算量明显最小，分别为 0.59 M 和 0.20 M；且识别 1 000 张图片所用的推理时间最少，仅为 0.66 s，因此本文

方法为最佳稀疏化方法。

4　结语

本文提出了一种基于多重稀疏 MobileNetV2 模型的变压器故障识别方法。并在 DGA 数据集上开展数值试验和性能评估,得到如下结论。

(1)基于纺锤块和沙漏块构建紧凑 MobileNetV2 模型,初步稀疏模型,较 MobileNetV2 模型,紧凑 MobileNetV2 模型的参数量和计算量分别减少 27.8% 和 13.2%,且精度提高 0.6%。

(2)对紧凑 MobileNetV2_P 模型进行 8 位宽对称均匀量化得到紧凑 MobileNetV2_PQ 模型。紧凑 MobileNetV2_PQ 模型在将故障识别精度提升至 95.2% 的前提下,最大程度上减少了网络复杂度,较 MobileNetV2 模型其参数量、计算量和模型大小分别降低约 73.5%、96.9% 和 68.8%,且识别 1 000 张图片的推理时间仅为 0.66 秒。本文方法将紧凑改进、模型剪枝和参数量化三种单一稀疏化方法进行了有效结合,在保证模型精度的前提下,实现了深度学习模型的多重稀疏化。

利益冲突声明

所有作者声明不存在利益冲突。

All authors disclose no relevant conflict of interests.

作者贡献

刘航和史智予进行了方案设计、内容总结和论文撰写与修改,罗琳灵和李明进行了数据收集,牛犇参与了相关文献综述查询,刘志坚参与了论文审核。所有作者均阅读并同意了最终稿件的提交。

The scheme design, content summary, paper writing and revision were conducted by Liu Hang and Shi Zhiyu. The data collection was conducted by Luo Linglin and Li Ming. The research of literature review was conducted by Niu Ben. The manuscript was reviewed by Liu Zhijian. All authors have read and approved the final version of the paper for submission.

参 考 文 献

[1] YANG D, QIN J, PANG Y, et al. A novel double-stacked autoencoder for power transformers DGA signals with an imbalanced data structure [J]. IEEE Transactions on Industrial Electronics, 2021, 69(2): 1977-1987.

[2] ZHANG Y, DING X, LIU Y, et al. An artificial neural network approach to transformer fault diagnosis [J]. IEEE Transactions on Power Delivery, 1996, 11(4): 1836-1841.

[3] HOSSEINI M, STEWART B G, KEARNS M, et al. Construction of a transformer DGA health index based on DGA screening processes [C]//2020 IEEE Conference on Electrical Insulation and Dielectric Phenomena, East Rutherford, 2020.

[4] 刘航, 王有元, 梁玄鸿, 等. 基于多因素的变压器油中溶解气体体积分数预测方法[J]. 高电压技术, 2018, 44(4): 1114-1121.
LIU H, WANG Y Y, LIANG X H, et al. Prediction method of the dissolved gas volume fraction in transformer oil based on multi factors [J]. High Voltage Engineering, 2018, 44(4): 1114-1121.

[5] HAQUE N, JAMSHED A, CHATTERJEE K, et al. Accurate sensing of power transformer faults from dissolved gas data using random forest classifier aided by data clustering method [J]. IEEE Sensors Journal, 2022, 22(6): 5902-5910.

[6] 崔星, 陈静, 孙婧琪, 等. 基于 ICEEMDAN 多尺度模糊熵和 MVO-KELM 的变压器绕组铁心机械故障诊断[J]. 电机与控制应用, 2023, 50(10): 81-90.
CUI X, CHEN J, SUN J Q, et al. Mechanical fault diagnosis for transformer winding core based on ICEEMDAN multi-scale fuzzy entropy and MVO-KELM [J]. Electric Machines & Control Application, 2023, 50(10): 81-90.

[7] 臧旭, 张甜瑾, 邵心悦, 等. 基于时变滤波经验模态分解和 SSA-LSSVM 的变压器内部机械故障诊断方法[J]. 电机与控制应用, 2023, 50(9): 49-56.
ZANG X, ZHANG T J, SHAO X Y, et al. A transformer internal mechanical fault diagnosis method based on TVFEMD and SSA-LSSVM [J]. Electric Machines & Control Application, 2023, 50(9): 49-56.

[8] 薛健侗，马宏忠. 基于 VMD 和 WOA-SVM 的变压器绕组松动故障诊断[J]. 电机与控制应用，2023，50(8)：84-90.
XUE J T，MA H Z. Fault diagnosis for winding looseness of transformer based on VMD and WOA-SVM [J]. Electric Machines & Control Application，2023，50(8)：84-90.

[9] 刘可真，姚岳，赵现平，等. 基于样本集成学习和 SO-SVM 的变压器故障诊断[J]. 电机与控制应用，2023，50(12)：21-31.
LIU K Z，YAO Y，ZHAO X P，et al. Transformer fault diagnosis based on sample integration learning and SO-SVM [J]. Electric Machines & Control Application，2023，50(12)：21-31.

[10] ARCHANA R，JEEVARAJ P S E. Deep learning models for digital image processing：A review [J]. Artificial Intelligence Review，2024，57(1)：11.

[11] 李平，胡根铭. 基于改进神经网络与比值法融合的变压器故障诊断方法[J]. 高电压技术，2023，49(9)：3898-3906.
LI P，HU G M. Transformer fault diagnosis method based on the fusion of improved neural network and ratio method [J]. High Voltage Engineering，2023，49(9)：3898-3906.

[12] 马鑫，尚毅梓，胡昊，等. 基于数据特征增强和残差收缩网络的变压器故障识别方法[J]. 电力系统自动化，2022，46(3)：175-183.
MA X，SHANG Y Z，HU H，et al. Identification method of transformer fault based on data feature enhancement and residual shrinkage network [J]. Automation of Electric Power Systems，2022，46(3)：175-183.

[13] QIN J，YANG D S，WANG N，et al. Convolutional sparse filter with data and mechanism fusion：A few-shot fault diagnosis method for power transformer [J]. Engineering Applications of Artificial Intelligence，2023，124：106606.

[14] 高晗，田育龙，许封元，等. 深度学习模型压缩与加速综述[J]. 软件学报，2021，32(1)：68-92.
GAO H，TIAN Y L，XU F Y，et al. Survey of deep learning model compression and acceleration [J]. Journal of Software，2021，32(1)：68-92.

[15] LIANG T L，GLOSSNER J，WANG L，et al. Pruning and quantization for deep neural network acceleration：A survey [J]. Neurocomputing，2021，461：370-403.

[16] 杨春，张睿尧，黄泷，等. 深度神经网络模型量化方法综述[J]. 工程科学学报，2023，45(10)：1613-1629.
YANG C，ZHANG R Y，HUANG L，et al. A survey of quantization methods for deep neural networks[J]. Chinese Journal of Engineering，2023，45(10)：1613-1629.

[17] LIU Z J，HE W，LIU H，et al. Fault identification for power transformer based on dissolved gas in oil data using sparse convolutional neural networks [J]. IET Generation，Transmission & Distribution，2024，18(3)：517-529.

[18] SANDLER M，HOWARD A，ZHU M L，et al. Mobilenetv2：Inverted residuals and linear bottlenecks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition，Salt Lake City，2018.

[19] LIU Z，MAO H Z，WU C Y，et al. A convnet for the 2020s [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition，New Orleans，2022.

[20] ZHOU D，HOU Q，CHEN Y，et al. Rethinking bottleneck structure for efficient mobile network design [C]//Computer Vision-ECCV 2020：16th European Conference，Glasgow，2020.

[21] FANG G，MA X，SONG M，et al. DepGraph：Towards any structural pruning [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition，Vancouver，2023.

[22] SANKARARAMAN K A，DE S，XU Z，et al. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent [C]//International Conference on Machine Learning，Virtual，2020.

[23] YOU Z H，YAN K，YE J M，et al. Gate decorator：Global filter pruning method for accelerating deep convolutional neural networks [C]// 33rd Conference on Neural Information Processing Systems，Vancouver，2019.

[24] SHRIVASTAVA K，CHOUBEY A. Data mining approach with IEC based dissolved gas analysis for fault diagnosis of power transformer [J]. International Journal of Engineering Research and Technology，2013，2(3)：1-11.

[25] 刘志坚，何蔚，刘航，等. 基于格拉姆角场变换和深度压缩模型的变压器故障识别方法[J]. 电网

技术, 2023, 47(4): 1478-1490.
LIU Z J, HE W, LIU H, et al. Fault identification method for power transformer based on Gramian angular field transformation and deep compression model [J]. Power System Technology, 2023, 47(4): 1478-1490.

[26] 刘志坚, 孟欣雨, 刘航, 等. 基于改进 ResNet34 网络的变电站设备巡检图像分类识别的方法[J]. 电机与控制应用, 2024, 51(5): 50-60.
LIU Z J, MENG X Y, LIU H, et al. Method for substation equipment inspection image classification and recognition based on improved ResNet34 network [J]. Electric Machines & Control Application, 2024, 51(5): 50-60.

[27] HE Y, LIU P, WANG Z W, et al. Filter pruning via geometric median for deep convolutional neural networks acceleration [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019.

[28] LEE J, PARK S, MO S, et al. Layer-adaptive sparsity for the magnitude-based pruning [C]// 2021 The Ninth International Conference on Learning Representations, Vienna, 2021.

[29] LIU Z, LI J G, SHEN Z Q, et al. Learning efficient convolutional networks through network slimming [C]// 2017 IEEE International Conference on Computer Vision, Venice, 2017.

[30] HOEFLER T, ALISTARH D, BEN-NUN T, et al. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks [J]. Journal of Machine Learning Research, 2021, 22(241): 1-124.

收稿日期:2024-12-18
收到修改稿日期:2025-02-26

作者简介:

刘　航(1992-),男,博士,讲师,研究方向为电力设备状态评估与故障诊断,liuhangsheep@163.com;

*通信作者:刘志坚(1975-),男,博士,教授,研究方向为电力系统运行与控制等,248400248@qq.com。