DOI: 10.13382/j.jemi.B2205219

基于多层面压缩深度神经网络的轴承故障诊断*

刘 钊[1] 孙洁娣[1,2] 温江涛[3]

(1.燕山大学信息科学与工程学院 秦皇岛 066004;2.燕山大学河北省信息传输与信号处理重点实验室 秦皇岛 066004;3.燕山大学河北省测试计量技术及仪器重点实验室 秦皇岛 066004)

摘 要:针对在资源有限的工业环境中难以应用基于深度神经网络的故障诊断模型的问题,提出一种压缩深度神经网络的轴承故障诊断方法,将结构化剪枝、非结构化剪枝、参数量化及矩阵压缩多层面处理相结合,实现了网络多级压缩。首先用结构化剪枝剔除卷积层中输出低秩特征图对应的滤波器;再用非结构化剪枝去除全连接层中非重要性连接;最后通过对权重矩阵的参数量化减少参数表示所需比特数,并结合权值矩阵压缩存储方法进一步减小了网络的参数存储量。实验表明提出的压缩方法在保证较高诊断准确率的前提下,极大减少了网络的参数存储量和浮点运算量,缩短了网络训练时间,加快了网络响应速度,为深度神经网络方法的工业实际应用进行了有益探索。

关键词:轴承故障诊断;深度学习;模型压缩;网络剪枝;参数量化

中图分类号:TH165+.3 **文献标识码**:A **国家标准学科分类代码**:460.40

Bearing fault diagnosis method based on multi-dimension compressed deep neural network

Liu Zhao[1] Sun Jiedi[1,2] Wen Jiangtao[3]

(1. School of Information Science and Engineering, Yanshan University, Qinghuangdao 066004, China; 2. Hebei Key Laboratory of Information Transmission and Signal processing, Yanshan University, Qinhuangdao 066004, China; 3. Key Laboratory of Measurement Technology and Instrumentation of Hebei Province, Yanshan University, Qinghuangdao 066004, China)

Abstract: Aiming at the problem that it is difficult to apply the fault diagnosis model based on deep neural network in the industrial environment with limited resources, a bearing fault diagnosis method based on compressed deep neural network was proposed. Firstly, the filter corresponding to the output low-rank feature graph in the convolution layer is removed by structural pruning. Then unstructured pruning was used to remove non-important connections in the whole connection layer. Finally, the number of bits required for parameter representation is reduced by quantizing the parameters of the weight matrix, and the storage of parameters is further reduced by using the compression storage method of the weight matrix. Experimental results show that the proposed compression method can greatly reduce the parameter storage and floating point computation of the network, shorten the training time of the network and speed up the response of the network on the premise of high diagnostic accuracy, which provides a beneficial exploration for the industrial application of the deep neural network method.

Keywords: bearing fault diagnosis; deep learning; model compression; network pruning; parameter quantification

0 引 言

滚动轴承是旋转机械的重要组成部件,其运行状态往往直接影响机器能否正常工作[1]。随着设备工作时间的增加,轴承的磨损也会更加严重,从而会导致轴承产生故障,影响整个机械设备的正常工作,造成经济损失[2]。因此对滚动轴承的运行状态进行准确的检测和诊断在工业领域有着重大意义。

传统的轴承故障诊断方法主要是通过提取信号时域

收稿日期:2022-03-01 Received Date:2022-03-01

*基金项目:河北省自然科学基金(E2020203061)项目资助

特征、频域特征和时频特征[3-4]结合机器学习[5]的方法进行故障诊断,该类方法依赖于人工特征提取、特征降维、分类器选择,无法在少量特征中获取复杂故障的信息,造成了该方法适用的局限性。

近年来,深度学习已在诸多领域得到了广泛的应用,其自适应提取数据特征的能力,打破了传统人工提取特征的局限性,特别是深度卷积神经网络(deep convolutional neural network, DCNN)已在许多领域取得了显著的效果。DCNN 通常由多个卷积层和全连接层构成,这使其具备了强大的特征提取和分类能力,许多研究人员已将其应用于机械轴承故障诊断中。许多学者采用一维 DCNN 直接对采集的时域信号进行分析,实现了轴承故障诊断[6-8];有些学者则利用 CNN 在图像处理领域展现的强大学习及分类能力,研究多种变换方法先将一维时域故障信号转换为二维图像,再采用基于 CNN 的深度网络对图像展现的故障时频信息进行深度挖掘,从而实现复杂故障诊断。许多学者将一维时域信号转换为二维灰度图像,结合二维 CNN 和图像处理的方法对其进行诊断分类,能够充分发挥深度神经网络强大的学习能力,有利于数据的特征提取并消除了人工提取特征的影响[9-12]。Zhang 等[13]通过短时傅里叶变换将一维时域信号转换为二维图像的频域信号,并采用 CNN 提取数据特征。Lu 等[14]通过分段组合的方式将原始监测信号映射为二维矩阵,使用 DCNN 网络提取数据的高维特征,实现轴承故障诊断。

通过分析文献发现,虽然基于深度神经网络的方法识别准确率较高,但是主要通过构建复杂的网络模型、不断加深网络深度、层与层之间增加多种处理来实现诊断性能的提升,造成参数量庞大、对监测系统软硬件要求过高,训练和测试需要耗费大量的计算资源和更多的时间,极大的限制了此类方法在资源有限的工业环境及实时在线处理中的应用。研究表明,多层神经网络中通常存在大量冗余参数,这些参数对网络诊断性能影响较小,因此可考虑在兼顾深度网络诊断准确率及监测设备软硬件的情况下剔除模型的冗余参数,对深度网络结构进行压缩,实现在现场工业监测设备中部署深度神经网络诊断方法,改善在线监测及诊断性能。

目前深度神经网络的压缩方法主要有网络剪枝[15]、参数量化[16]、知识蒸馏[17-18]和轻量级网络设计[19]等。其中网络剪枝是常用的网络压缩方法之一,能同时对卷积层和全连接层进行裁剪,剪枝方法又分为结构化剪枝和非结构化剪枝[20],结构化剪枝通过压缩卷积层实现网络加速;非结构化剪枝通过修剪低于给定阈值的参数达到减小网络占用内存的目的。在非结构化剪枝方法研究方面,Han 等[15]提出删除网络中“不重要”的参数,在不降低准确率的情况下成功将网络参数量减少一个数量级,然而该方法修剪的参数无法恢复,错误的修剪可能会降低网络的准确率。随后 Guo 等[21]提出一种剪枝与拼接相结合的动态网络剪枝策略,通过剪枝修剪“不重要”的参数,利用拼接恢复错误修剪的参数,从而实现动态的网络剪枝,解决了因前置层参数的错误剪枝可能导致的网络准确率下降的问题。Han 等[22]结合参数剪枝、量化和霍夫曼编码的方法将 AlexNet 和 VGG16 模型的参数量压缩了 35~39 倍。Wang 等[23]结合稀疏学习和遗传算法,利用评估因子去除网络中的冗余分支,通过遗传算法动态的优化评估因子,从稀疏网络中选择最有效的剪枝方案,在不影响准确率的情况下降低了网络的计算复杂度。非结构化剪枝需要特定的硬件加速才能实现。在结构化剪枝方法研究方面,Lin 等[24]通过计算滤波器输出特征图秩的大小,删除输出低秩特征图对应的滤波器,减少了训练过程中的浮点运算。季繁繁等[25]提出一种二阶信息的结构化剪枝算法,加速了网络的收敛,且准确率有所提升。结构化剪枝方法无需硬件加速,但网络参数大量集中于全连接层,虽然能够达到网络加速但在减少网络内存占用方面效果不明显。

由于神经网络的训练速度主要取决于卷积层滤波器的数量和大小,参数主要存在于全连接层,因此仅使用结构化或非结构剪枝,难以同时实现网络训练加速及网络参数大幅度减少。而目前基于深度神经网络的轴承故障诊断方法多数是通过深层且复杂的网络模型来提高诊断准确率,庞大的参数量和浮点运算量使得模型部署需要大容量的存储资源和高性能的计算资源,导致难以在实际工业监测平台上应用。针对上述问题,本文提出一种结合结构化与非结构化剪枝、量化和矩阵压缩的混合模型压缩方法,同时压缩网络的卷积层和全连接层,并对其参数进行量化,最后再利用矩阵压缩的方法进一步减小量化参数所需的存储空间。本文提出的方法能够同时实现网络的压缩与加速,在保证诊断准确率的前提下,降低了对监测设备计算和存储能力的要求,为实际工业现场在线监测采用深度神经网络类方法进行了有益的探索。

1　本文提出的方法

目前深度神经网络方法在轴承故障诊断中存在参数量大,训练时间长,响应速度慢,对监测平台的计算和存储资源要求较高,导致实际应用困难。针对上述问题,本文提出了一种模型压缩方法处理深度神经网络模型,旨在解决应用此类模型所需的庞大计算开销和内存需求与监测设备资源受限之间的矛盾。

本文提出的网络压缩方法是一种结合结构化剪枝、非结构化剪枝、量化与矩阵压缩存储的多角度混合压缩方法。首先在结构化剪枝阶段剔除卷积层中不重要的滤

波器,减少网络的浮点运算次数,加速网络训练,并对剪枝后的网络进行微调使其恢复网络性能;其次利用非结构化剪枝方法裁剪全连接层中不重要的连接,从而减少参数量;最后对剩余的参数量化,通过聚类、权重共享降低表示权重参数的比特(bit)数。本文方法的处理过程如图 1 所示。

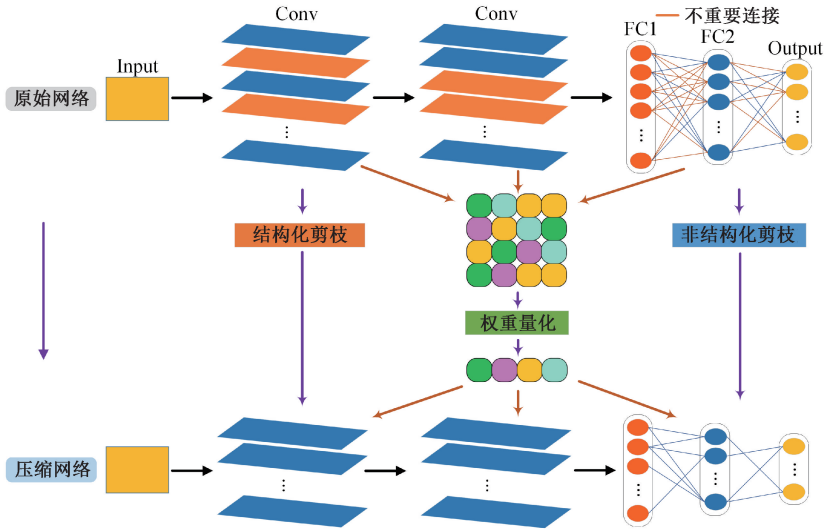


图 1　本文方法处理过程

Fig. 1　Schematic of proposed method

1.1　结构化剪枝

基于 CNN 的深度神经网络通常由多个卷积层和全连接层堆叠组成,每个卷积层中滤波器与该层输入数据的卷积运算往往会产生大量的参数和浮点运算,冗余滤波器产生的结果也在其中,剔除这些滤波器不会影响网络的性能。特征图作为滤波器与输入数据卷积运算输出的结果,本文通过计算网络中每个卷积核输出特征图的秩来衡量卷积核的重要性,剔除输出低秩特征图对应滤波器,从而实现网络压缩与加速。卷积层参数与浮点运算量(floating point operations, FLOPs)计算公式如下。

$$\begin{cases} Parameters = N_{out} \times N_{in} \times K_w \times K_h \\ FLOPs = M_w \times M_h \times N_{out} \times N_{in} \times (K_w \times K_h \times 2 - 1) \end{cases} \tag{1}$$

其中,N_{out},N_{in},K_w,K_h 分别表示各卷积层的输出通道数,输入通道数,卷积核的大小,M_w,M_h 为输入图像数据的大小。

1.2　非结构化剪枝

网络模型中参数主要集中于全连接层,利用非结构化剪枝可以裁剪网络全连接层中的冗余连接,从而减少网络参数量。对网络中每层的参数排序,将低于某一阈值的参数定义为冗余参数,剔除后不会对网络准确率造成太大的影响,通过剔除这些冗余参数而保留对网络准确率贡献大的参数,降低参数存储所需的设备存储容量。全连接层参数量与浮点运算量计算公式如下。

$$\begin{cases} parameters = N_{out} \times (N_{in} + 1) \\ FLOPs = N_{out} \times N_{in} \times 2 \end{cases} \tag{2}$$

其中,N_{out},N_{in} 分别为各全连接层输入与输出神经元个数,剪枝过程示意图如图 2 所示,图中虚线部分为移除的网络连接,H,W,C 分别表示最后一个卷积层输出特征图的大小和输出通道数。

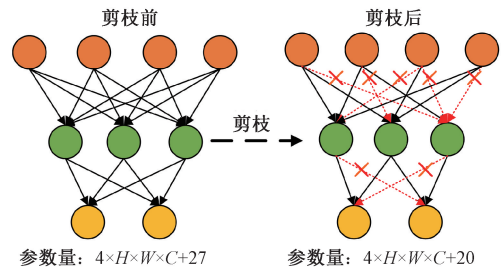


图 2　非结构化剪枝示意图

Fig. 2　Schematic diagram of pruning process

1.3　权重量化

网络量化通过减少表示每个权重所需的比特数来压缩原始网络[26-27],主要有权重共享和低比特表示两种方法,本文采用层内(层与层之间不共享)权重共享方法来实现权重量化。对模型各层的权重矩阵使用 K-means 聚类算法,得到聚类中心和对应的聚类索引,并将每个权重用其所在的聚类中心代替,最后只需存储其聚类中心和聚类索引,权重共享聚类过程如图 3 所示,其中相同颜色表示聚为一类。

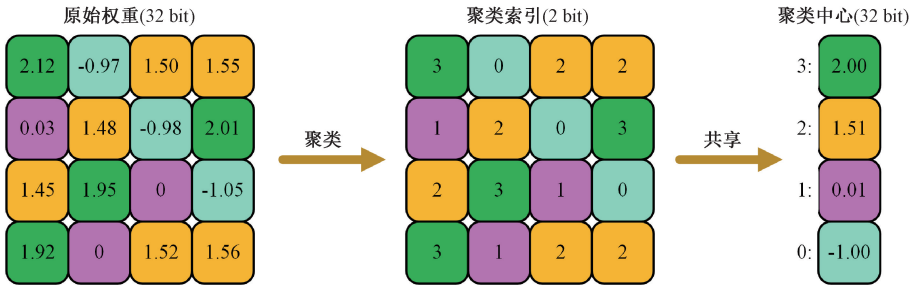


图 3　权重共享聚类过程

Fig. 3　Weight sharing clustering process

量化后的权重由原始的 32 bit 浮点数变为 2 bit 聚类索引和 32 bit 聚类中心表示，使得存储的数据量大大减少。若聚类类别为 k，则索引数量为 $\log_2(k)$ bit，在有 n 个权重的网络中，若每个权重用 b bit 表示，则压缩率 r 可表示如下：

$$r = \frac{nb}{n\log_2(k) + kb} \tag{3}$$

在图 3 中，原始权重矩阵大小为 4 × 4，即权重数量为 16，聚类类别为 4，每个权重由 32 位表示，得到压缩率为 16×32/(16×2+4×2)= 32。

1.4　权值矩阵压缩

在高阶的权值矩阵中往往存在许多相同值的元素，可对其稀疏化后利用矩阵压缩存储，进一步节省存储空间。目前稀疏矩阵的压缩存储方法主要包括按行压缩(compressed sparse row，CSR)和按列压缩(compressed sparse column，CSC)。本文采用按行压缩的方法，若稀疏矩阵中非零元素远少于零元素，则对每个非零元素可只存储其值、行索引及列索引 3 个元素。设矩阵 $\boldsymbol{A}$ 为 4×4 矩阵，a_{ij} 为 $\boldsymbol{A}$ 中第 i 行第 j 列元素，$i,j = 1,2,3,4$，其中 $a_{12},a_{21},a_{22},a_{33},a_{42}$ 不为 0，其余元素均为 0，则矩阵按行压缩存储的元素如表 1 所示。

表 1　矩阵按行压缩存储元素

Table 1　Matrices store elements in compressed rows

值	a_{12}	a_{21}	a_{22}	a_{33}	a_{42}
行索引	1	2	2	3	4
列索引	3	1	2	3	2

在量化后的聚类索引矩阵中并非所有索引矩阵都为稀疏矩阵，本文通过将索引矩阵中相同且最多元素归零处理，从而对其稀疏化，再利用矩阵压缩方法存储聚类索引矩阵。设索引矩阵 $\boldsymbol{A}$ 各元素为 a_{ij}，索引 k 的个数为 n_k，则稀疏化后的 a_{ij} 表示如下：

$$a_{ij}' = a_{ij} - \max\{n_i\} \tag{4}$$

针对基于 CNN 的深度神经网络在实际工业应用中存在的问题，本文采用多角度混合压缩方法，从网络的不同部分进行处理，实现了深度网络结构压缩，为深度网络类方法在工业现场的应用推广提供了新思路。

2　实验情况简介

2.1　测试网络简介

基于 CNN 的网络主要通过不同的卷积层和全连接层的组合改善特征提取性能，提高识别准确率，但随着层数的增加，模型参数量和浮点运算也急剧增加，为了验证本文提出的混合压缩方法，选取了文献[10]和[11]提出的两种不同深度的卷积神经网络模型来测试压缩效果，并进一步分析方法性能。由于文献[11]提出的模型深度比文献[10]提出的模型深度更深，因此本文分别将文献[10]和[11]的模型表示为 S-CNN、D-CNN，两个模型的结构如表 2 所示。

表 2　本文使用的模型结构

Table 2　Brief instruction for model structures tested

模型	输入	网络结构	输出
D-CNN	二维图像	(卷积+池化)×4+全连接×3	Softmax
S-CNN		(卷积+池化)×2+全连接×3	

2.2　数据集概述

1)数据集 I 情况简介

该数据集源自西安交通大学机械装备健康监测联合实验室[28]，实验设备为多级齿轮传动试验台，该试验台主要由电动机、定轴齿轮箱、行星齿轮箱等组成。数据样本采自定轴齿轮箱的 LDK UER204 型滚动轴承，轴承故障类型包括内圈、外圈、滚动体、保持架及混合故障 4 种类型。在转速和径向力不同的情况下数据样本又分为 3 种工况，分别为工况 1：转速 2 100r/min，径向力 12 kN；工况 2：转速 2 250 r/min，径向力：11 kN；工况 3：转速 2 400 r/min，径向力：10 kN，每种工况又包括 5 个轴承，本文分别在不同工况下尽可能选取轴承故障位置不同的 10 类故障类型，详细信息如表 3 所示。

表 3　数据集 I 详细信息

Table 3　The details of dataset I

工况	数据集	故障类型	故障标签
1	Bearing1_1	外圈	0
	Bearing1_4	保持架	1
	Bearing1_5	混合	2
2	Bearing2_1	内圈	3
	Bearing2_2	外圈	4
	Bearing2_3	保持架	5
	Bearing2_5	外圈	6
3	Bearing3_1	外圈	7
	Bearing3_2	混合	8
	Bearing3_3	内圈	9

2)数据集 II 情况简介

该数据集来自美国凯斯西储大学的轴承数据中心[29],数据集中的样本数据采自电动机驱动端的 SKF6205 型滚动轴承,在电机转速为 1 797 r/min,采样频率为 12 kHz,负载为 0 马力的情况下,本文选取的数据包括 4 种健康状况:(1)正常状态(N),(2)内圈故障(IF),(3)外圈故障(OF),(4)滚动体故障(BF),滚动体故障中心位置为 6 点钟方向,其中 IF、OF、BF 这 3 种故障的故障程度又分为故障直径为 0.18 mm、0.36 mm 和 0.54 mm 这 3 种情况。

3)数据集处理

为更好地提取数据特征和适应所使用的模型输入,本文将原始一维数据转换为二维灰度图像作为输入。按特定间隔在原始数据中截取长度为 M^2 的一维数据,设 $L(k),k=1,2,\cdots,M^2$ 表示原始信号第 k 点的值,$P(i,j),i,j=1,2,\cdots,M$ 表示在点 (i,j) 出的像素强度,两者的转换关系如下:

$$P(i,j)=round\left\{\frac{L((i-1)\times M+j)-\mathrm{Min}(L)}{\mathrm{Max}(L)-\mathrm{Min}(L)}\times 255\right\} \tag{5}$$

其中,$round(\cdot)$ 表示四舍五入函数,通过上述转换,将原始长度为 M^2 的一维信号转换为 $M\times M$ 大小,像素为 0~255 的灰度图像。

将处理后数据的 70%做为训练集,30%做为测试集,最终得到数据集 I 和数据集 II 的详细信息如表 4 所示。

表 4　数据集详细信息

Table 4　Details of dataset

数据集	训练样本数	测试样本数	样本大小
数据集 I	2 100	900	64×64
数据集 II	1 680	720	64×64

2.3　实验参数与性能指标

模型优化器均使用 SGD(stochastic gradient descent),epoch 为 50,训练集批处理大小 batch-size 为 64,测试集 batch-size 为 32,初始学习率为 0.01,分别在 30 和 40 个 epoch 后将学习率乘以 0.1。为了避免实验结果的偶然性,所有结果取 15 次重复实验结果的均值,实验均基于 Pytorch 框架实现,PC 配置为:Inter Core i5 3.30-GHz CPU,8-GB RAM。

模型压缩算法的评价指标主要有准确率、参数存储量和浮点运算量,其中准确率衡量了压缩后模型的诊断性能;参数存储量衡量了模型所占用监测设备的内存大小;浮点运算量衡量了模型的运算速度,浮点运算量减少量越大,表明模型加速效果越明显,诊断速度越快。

3　实验结果及分析

3.1　剪枝率选取

在网络的卷积层和全连接层中,裁剪不同数量的滤波器和参数对网络性能有不同的影响,通常使用剪枝率来表示对网络的裁剪程度。卷积层中剪枝率为移除的滤波器数量与原始滤波器数量的比值,全连接层中剪枝率为剔除的参数量与原始网络参数量的比值。为了得到最佳剪枝率,使得在剪枝后网络参数最少而不影响诊断准确率,在 S-CNN 模型上实验得到结构化和非结构化的剪枝率和网络诊断准确率的变化关系如图 4 所示。

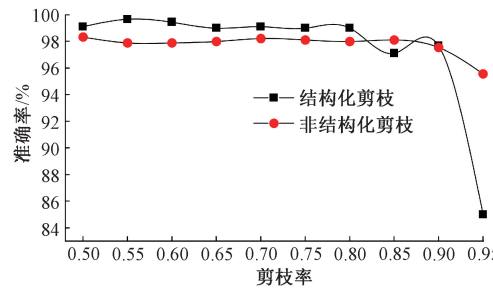


图 4　不同剪枝率的故障诊断准确率

Fig. 4　Diagnostic accuracies under different pruning rate

从图 4 可以看出,剪枝率为 0.9 时的结构化剪枝的识别准确率比剪枝率为 0.85 时高;而非结构化剪枝率在大于 0.9 后准确率急剧下降,这是由于裁剪大量的参数使得原始模型结构变化较大,影响了模型性能。综合考虑性能要求,本文选取结构化剪枝率为 0.9,非结构化剪枝率为 0.85,为了保证全连接层的输入数量,最后一个卷积层的剪枝率设为 0.85。

3.2　压缩性能分析

本节在数据集 I 和数据集 II 上对比了 D-CNN 原始网络(D-CNN-O)、D-CNN 压缩网络(D-CNN-C)、S-CNN 原始网络(S-CNN-O)、S-CNN 压缩网络(S-CNN-C)的分类准确率、参数存储量、浮点运算量及训练时间 4 个性能指标,各指标统计结果如表 5 所示。

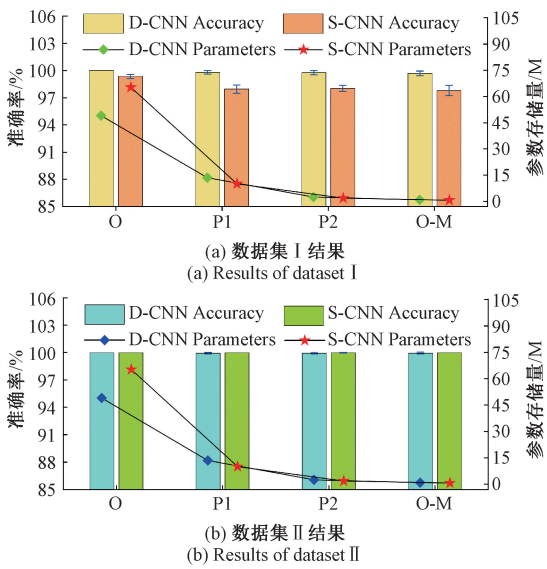
表 5 本文方法对不同网络的压缩结果

Table 5 Compression of the proposed method with different networks

数据集	模型	准确率/%	参数存储量/M	浮点运算量/M	训练时间/s
数据集 I	D-CNN-O	100±0.00	49.04	69.71	1 048.69
	D-CNN-C	99.70±0.25	0.89	4.30	786.82
	S-CNN-O	99.35±0.23	65.22	69.94	1 104.35
	S-CNN-C	97.79±0.56	0.66	3.52	682.18
数据集 II	D-CNN-O	100±0.00	49.04	69.71	841.24
	D-CNN-C	99.94±0.10	0.89	4.30	633.33
	S-CNN-O	100±0.00	65.22	69.94	830.88
	S-CNN-C	100±0.00	0.66	3.52	547.96

由表 5 可以看出，采用本文提出的压缩方法对 D-CNN 模型压缩后，在几乎不影响准确率的情况下将参数存储量压缩为原网络的 1.81%，浮点运算量仅为原网络的 6.17%，训练时间仅为原网络的 3/4；对 S-CNN 模型压缩后，参数存储量仅为原网络的 1.01%，浮点运算量仅为原网络的 5.03%，在数据集 I 上，训练时间为原网络的 8/13，在数据集 II 上，训练时间为原网络的 2/3，在数据集 I 上虽然准确率下降了 1.56%，但仍然有较高的识别准确率。采用本文提出的方法极大的减少了模型参数占用的监测设备存储空间，加快了网络训练，且在仅有少量训练数据的数据集 II 上，对压缩后的网络微调后依然可以获得较好的诊断性能。

图 5 为数据集 I 和数据集 II 上 D-CNN 和 S-CNN 不同压缩阶段参数存储量与准确率变化关系图，其中 O 表示原始无压缩模型，P1 表示结构化剪枝后，P2 表示非结构化剪枝后，Q-M 表示量化和矩阵压缩存储后。



(a) 数据集 I 结果
(a) Results of dataset I

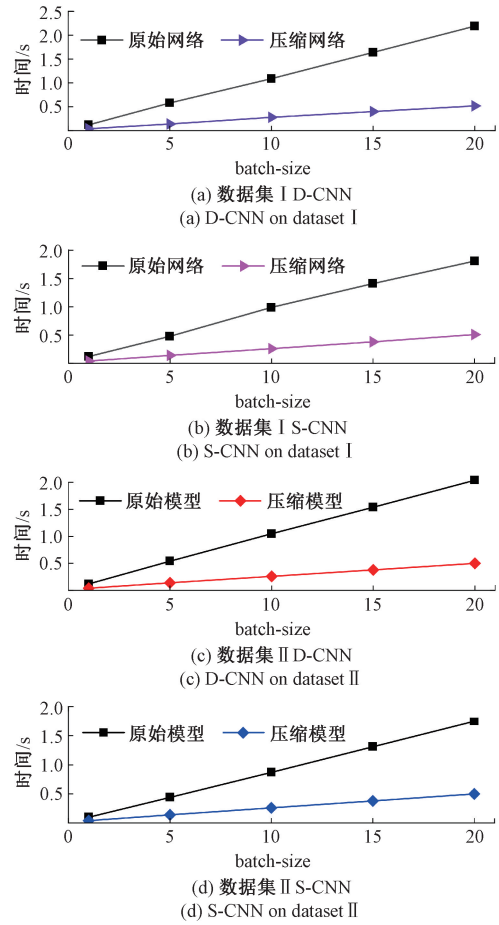
(b) 数据集 II 结果
(b) Results of dataset II

图 5 不同压缩阶段的参数存储量与准确率变化

Fig. 5 Changes of parameter volume and diagnosis accuracy of different compression stages

由图 5 可以看出，在数据集 I 上，模型在结构化剪枝后参数减少量最大，但准确率也受到较大影响；非结构化剪枝和量化与矩阵压缩阶段准确率变化不大。在数据集 II 上，本文提出的方法在不影响识别准确率的情况下，模型的参数存储量大幅度减少。本文方法在总体准确率较高的情况下压缩了大量参数，降低了模型参数的存储成本和训练时的计算成本。

为验证采用本文压缩方法压缩后的模型能更快的识别样本的故障类别，本文选取 1、5、10、15、20 个 batch-size 的测试数据进行实验，分别对比了压缩后的模型与原模型处理不同数量样本所需要的时间，在 D-CNN 和 S-CNN 上的实验结果如图 6 所示。



(a) 数据集 I D-CNN
(a) D-CNN on dataset I

(b) 数据集 I S-CNN
(b) S-CNN on dataset I

(c) 数据集 II D-CNN
(c) D-CNN on dataset II

(d) 数据集 II S-CNN
(d) S-CNN on dataset II

图 6 不同数量样本数据的响应时间

Fig. 6 Response time of different number of samples

由图 6(a)~(d)可以看出，压缩后的模型的诊断响应速度更快，所需时间更短。随着需要诊断样本数量的增加，压缩后的模型诊断响应速度受影响较小，用时增长较慢，而原始模型的诊断响应速度减缓明显，诊断耗时线性增加。由此可见，本文提出的压缩方法使得模型在实际的工业场景中能更快的诊断轴承故障，缩短了诊断时间，改善了诊断实时性，能快速识别轴承故障。

4　与其他方法对比分析

现有的压缩方法多数为单独的结构化剪枝或非结构化剪枝,以及剪枝与量化结合的方法。在仅有结构化剪枝和非结构化剪枝的情况下通常难以在加速模型训练的同时减少大量参数。而本文提出的多层面混合压缩方法在大幅度减少参数量的同时能加速模型的训练。下面将本文提出压缩方法与常见的压缩方法进行对比:

1) FPEI (filter pruning entropy importance)[30],该算法将滤波器输出特征图的熵的大小作为衡量滤波器重要性的指标,修剪熵较小的特征图对应的滤波器,最后对修剪后的网络再进行微调恢复网络的性能。

2) CHIP (Channel independence pruning)[31],该方法根据通道之间的相似性来判断冗余的滤波器,计算每个滤波器输出特征图与同一层中其他滤波器输出特征图的相似性,将相似的滤波器中可以被其他滤波器所表示的滤波器定义为冗余滤波器,将其移除后不会影响模型性能,最后保证了模型中滤波器的独立性,实现了模型压缩。

4.1　与其他压缩方法的压缩性能对比

在数据集 I 和数据集 II 上对比了本文方法与 FPEI 算法,CHIP 算法在 D-CNN 和 S-CNN 模型上的参数存储量、浮点运算量和训练时间 4 个指标,其统计结果如表 6 所示。

表 6　不同压缩方法压缩性能对比结果

Table 6　Compression performance comparisons of different methods

模型	压缩方法	数据集 I			数据集 II		
		参数存储量/M	浮点运算量/M	训练时间/s	参数存储量/M	浮点运算量/M	训练时间/s
D-CNN	FPEI	27.93	22.30	664.23	27.93	22.30	412.14
	CHIP	47.82	18.42	452.95	47.82	18.42	368.87
	本文方法	**0.89**	**4.30**	**786.82**	**0.89**	**4.30**	**547.96**
S-CNN	FPEI	33.07	22.60	494.67	33.07	22.60	401.17
	CHIP	65.07	28.14	670.53	65.07	28.15	542.72
	本文方法	**0.66**	**3.52**	**682.18**	**0.66**	**3.52**	**633.33**

从表 6 可以看出,在数据集 I 上,本文方法在参数存储量和浮点运算量上远小于 FPEI 算法和 CHIP 算法,由于本文方法需要多级压缩,因此在训练时间上比 FPEI 算法和 CHIP 算法的训练时间略长。在数据集 II 上,本文方法的模型参数存储量和浮点运算量最小,分别对不同压缩方法处理后两种网络的参数存储量和浮点运算量求和,得出本文方法模型参数存储量为 FPEI 算法模型参数存储量的 2.55%,CHIP 算法模型参数存储量的 1.37%。浮点运算量为 FPEI 算法浮点运算量的 17.41%,CHIP 算法浮点运算量的 16.79%。综上分析,本文方法在压缩模型的同时加速网络的训练,对比 FPEI 算法和 CHIP 算法,本文方法整体上最优。

表 7 为模型在数据集 I 和数据集 II 上采用本文方法、FPEI 算法和 CHIP 算法压缩后的模型诊断准确率对比。

表 7　不同压缩方法准确率对比

Table 7　Comparison of accuracy of different methods

模型	压缩方法	准确率/%	
		数据集 I	数据集 II
D-CNN	FPEI	100±0.00	100±0.00
	CHIP	99.86±0.18	99.89±0.19
	本文方法	**99.70±0.25**	**99.94±0.10**
S-CNN	FPEI	99.91±0.07	100±0.00
	CHIP	99.48±0.26	99.89±0.19
	本文方法	**97.79±0.56**	**100±0.00**

由表 7 可以看出,在数据集 I 上,对于网络较深的 D-CNN 模型,本文方法压缩后的模型准确率与 FPEI 算法、CHIP 算法压缩得到的准确率相差不大。在仅有两层卷积层的 S-CNN 模型上,采用本文提出的压缩方法处理模型后,虽然准确率较 FPEI 算法和 CHIP 算法处理后的准确率稍低,但本文方法处理后的模型参数远小于 FPEI 算法和 CHIP 算法处理后的模型参数。在数据集 II 上,采用本文方法压缩后的模型在参数存储量和浮点运算量上远小于 FPEI 算法和 CHIP 算法,而诊断准确率优于 CHIP 算法的模型诊断准确率,与 FPEI 算法的准确率基本一致,表明本文压缩方法具有较好的鲁棒性。

4.2　与其他压缩方法的响应时间对比

本文提出的混合压缩方法,从不同层面分析网络参数的特性,通过多级压缩实现了加快模型训练的同时减少模型参数量的目的;相比之下,FPEI 和 CHIP 算法从不同角度衡量滤波器的重要性,仅使用结构化剪枝的方法来实现网络的加速和减少模型参数,因此本文方法在模型训练时间上较两种对比算法略长,而在实际监测过程中,通常更关注的是诊断的响应时间,本节在数据集 I 和数据集 II、D-CNN 和 S-CNN 两种模型上分别对比了本文方法与 FPEI 算法、CHIP 算法压缩后的模型诊断时间,诊

断样本数据为测试集中抽取的 20 个 batch-size 的样本，实验结果如图 7 所示。

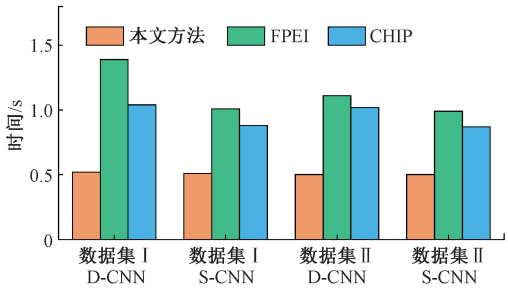


图 7 不同压缩方法的响应时间对比

Fig. 7 Comparison of response time with different methods

从图 7 可以看出，在不同数据集和不同模型上，本文方法压缩后的模型诊断时间明显少于 FPEI 算法和 CHIP 算法，本文方法能使得模型具有更好的实时诊断能力。

4.3 与其他压缩方法的参数压缩比例对比

为分析本文方法的压缩性能，在数据集 I 上，对比不同方法对 D-CNN 和 S-CNN 模型压缩后的参数量和浮点运算量，不同压缩方法处理模型后的参数剩余比例对比如图 8 所示。

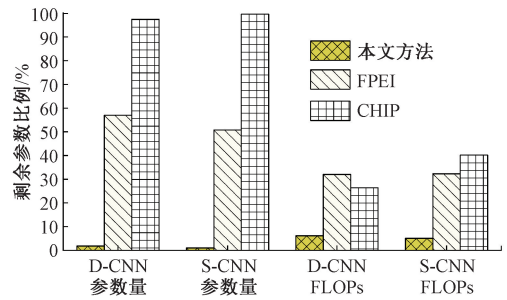


图 8 不同压缩方法参数剩余比例对比

Fig. 8 Comparison of remaining ratio of parameters in different compression methods

从图 8 可以看出，本文方法在 3 种方法中的参数量和浮点运算量剩余比例最小，表明本文提出的压缩方法在 3 种方法中的压缩性能最好。

4.4 与传统方法对比

传统机器学习方法在轴承故障诊断领域已取得了良好的成果，如支持向量机（SVM）[32]、K 最邻近法（KNN）[33]、自编码（SAE）[34] 等。为验证本文方法处理后的模型相比于传统机器学习方法的优异性，在数据集 II 上对比了本文方法压缩得到的模型与传统机器学习模型的诊断准确率和 640 个样本数据的诊断响应时间；同时本文还对比了有 1 个卷积层和 1 个全连接层的简单深度学习模型（simple deep learning model，SDLM），结构参数如表 8 所示。

表 8 SDLM 参数表

Table 8 Parameters of SDLM

Layer type	Parameters	Output size
Conv-BN-ReLU	$k=7$, $s=1$, $p=2$	64×31×31
MaxPool	$k=2$, $s=2$, $p=0$	64×31×31
FC	—	10

对比实验统计结果如表 9 所示。

表 9 与传统方法对比结果

Table 9 Comparisons with traditional models

模型	准确率/%	模型大小/M	响应时间/s
SVM	80.16±1.42	44.37	3.38
KNN	89.32±0.82	86.46	5.90
SAE	96.87±3.78	3.80	3.32
SDLM	99.48±0.54	2.36	1.11
S-CNN	**100±0.00**	**0.66**	**0.50**
D-CNN	**99.94±0.10**	**0.89**	**0.50**

由表 9 可以看出，本文方法处理后的深度神经网络模型在诊断准确率上高于传统机器学习模型和简单深度学习模型，且准确率标准差更小，表明模型更稳定。对于相同数量的数据样本，本文方法得到的模型可以更快的得到诊断结果，模型存储需要的存储资源最少，性能更好。

5 结 论

基于 CNN 的深度神经网络在轴承故障诊断研究中展示出良好的性能，但是堆叠多层网络虽然提高了故障识别准确率，但是造成参数量庞大、对监测系统软硬件要求过高，训练和测试需要耗费大量的计算资源和更多的时间，导致难以在实际工业监测设备中应用，为此本文提出了一种多层面混合网络压缩方法，在深度卷积神经网络的不同处理层进行网络模型的压缩，通过将结构化、非结构化剪枝并结合量化以及矩阵压缩方法，同时实现了压缩后参数量的急剧缩减以及浮点运算的大幅度减少，并且缩短了训练及响应时间。实验结果表明，本文方法在两种不同深度的诊断网络上均可取得较好的压缩与识别效果，与常用压缩方法及传统机器学习方法相比，本文方法在保持较高诊断准确率的前提下所需的监测设备存储容量更低，响应速度更快，为推进在实际工业监测设备上部署深度神经网络类方法、实现轴承快速故障诊断进行了有益的探索。

参考文献

[1] KHAN S, YAIRI T. A review on the application of deep learning in system health management[J]. Mechanical Systems and Signal Processing, 2018, 107: 241-265.

[2] 谢佳琪, 尤伟, 沈长青, 等. 基于改进卷积深度置信网络的轴承故障诊断研究[J]. 电子测量与仪器学报, 2020,34(2):36-43.
XIE J Q, YOU W, SHEN CH Q, et al. Research on bearing fault diagnosis based on improved convolution depth confidence network [J]. Journal of Electronic Measurement and Instrumentation, 2020,34(2):36-43.

[3] SUN J, YAN C, WEN J. Intelligent bearing fault diagnosis method combining compressed data acquisition and deep learning [J]. IEEE Transactions on Instrumentation and Measurement, 2018, 67(1): 185-195.

[4] 宫文峰, 陈辉, 张美玲, 等. 基于深度学习的电机轴承微小故障智能诊断方法[J]. 仪器仪表学报, 2020, 41(1): 195-205.
GONG W F, CHEN H, ZHANG M L, et al. Intelligent diagnosis method for minor faults of motor bearings based on deep learning [J]. Chinese Journal of Scientific Instrument, 2020,41(1): 195-205.

[5] 刘应东, 刘韬, 李华, 等. 变工况轴承的联合分布适应迁移故障诊断[J]. 电子测量与仪器学报, 2021, 35(5):69-75.
LIU Y D, LIU T, LI H, et al. Joint distribution adaptive migration fault diagnosis of bearings under variable working conditions [J]. Journal of Electronic Measurement and Instrumentation, 2021,35(5):69-75.

[6] WANG X, MAO D, LI X. Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network[J]. Measurement, 2021,173(6): 108518.

[7] INCE T, KIRANYAZ S, EREN L, et al. Real-time motor fault detection by 1-D convolutional neural networks [J]. IEEE Transactions on Industrial Electronics, 2016,63(11): 7067-7075.

[8] MO Z, ZHANG Z, TSUI K. The variational kernel-kased 1-D convolutional neural network for machinery fault diagnosis [J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-10.

[9] ZHANG J, SUN Y, GUO L, et al. A new bearing fault diagnosis method based on modified convolutional neural networks[J]. Chinese Journal of Aeronautics, 2020, 33(2): 439-447.

[10] ZHAO J, YANG S, LI Q, et al. A new bearing fault diagnosis method based on signal-to-image mapping and convolutional neural network[J]. Measurement, 2021, 176(1): 109088.

[11] WEN L, LI X, GAO L, et al. A new convolutional neural network-based data-driven fault diagnosis method[J]. IEEE Transactions on Industrial Electronics, 2018, 65(7): 5990-5998.

[12] WANG Z, ZHAO W, DU W, et al. Data-driven fault diagnosis method based on the conversion of erosion operation signals into images and convolutional neural network [J]. Process Safety and Environmental Protection, 2021,149(12):591-601.

[13] ZHANG Y, XING K, BAI R, et al. An enhanced convolutional neural network for bearing fault diagnosis based on time-frequency image[J]. Measurement, 2020, 157(99): 107667.

[14] LU C, WANG Z, ZHOU B. Intelligent fault diagnosis of rolling bearing using hierarchical convolutional network based health state classification [J]. Advanced Engineering Informatics, 2017,32: 139-151.

[15] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural network [J]. Advances in Neural Information Processing Systems, 2015,28: 1506-2626.

[16] CHOI Y, EL-KHAMY M, LEE J. Towards the limit of network quantization [J]. ArXiv Preprint, 2016, arXiv:1612.01543.

[17] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science, 2015,14(7): 38-39.

[18] LI X, LI S, OMAR B, et al. ResKD: Residual-guided knowledge distillation[J]. IEEE Transactions on Image Processing, 2021,30: 4735-4746.

[19] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6848-6856.

[20] 杨民杰, 梁亚玲, 杜明辉. 基于参数子空间和缩放因子的 YOLO 剪枝算法[J]. 计算机工程, 2021,47(2): 111-117.
YANG M J, LIANG Y L, DU M H. YOLO pruning algorithm based on parameter subspace and scaling factor [J]. Computer Engineering, 2021,47(2): 111-117.

[21] GUO Y, YAO A, CHEN Y. Dynamic network surgery for efficient dnns [J]. Advances in Neural Information Processing Systems, 2016: 29.

[22] HAN S, MAO H, DALLY W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding[J]. ArXiv Preprint, 2015, arXiv:1510.00149.

[23] WANG Z, LI F, SHI G, et al. Network pruning using sparse learning and genetic algorithm [J]. Neurocomputing, 2020,404: 247-256.

[24] LIN M, JI R, WANG Y, et al. HRank: Filter pruning using high-rank feature map [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1529-1538.

[25] 季繁繁, 杨鑫, 袁晓彤. 基于深度神经网络二阶信息的结构化剪枝算法[J]. 计算机工程, 2021, 47(2): 12-18.
JI F F, YANG X, YUAN X T. Structured pruning algorithm based on second-order information of deep neural network [J]. Computer Engineering, 2021, 47(2): 12-18.

[26] CHENG Y, WANG D, ZHOU P, et al. A survey of model compression and acceleration for deep neural networks[J]. ArXiv Preprint, 2017, arXiv: 1710.09282.

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

[28] 雷亚国, 韩天宇, 王彪, 等. XJTU-SY 滚动轴承加速寿命试验数据集解读[J]. 机械工程学报, 2019, 55(16): 1-6.
LEI Y G, HAN T Y, WANG B, et al. Interpretation of XJTU-SY rolling bearing accelerated life test data set[J]. Journal of Mechanical Engineering, 2019, 55(16): 1-6.

[29] SMITH W A, RANDALL R B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study[J]. Mechanical Systems and Signal Processing, 2015, 64-65: 100-131.

[30] WANG J, JIANG T, CUI Z, et al. Filter pruning with a feature map entropy importance criterion for convolution neural networks compressing [J]. Neurocomputing, 2021, 461: 41-54.

[31] SUI Y, YIN M, XIE Y, et al. CHIP: CHannel independence-based pruning for compact neural networks[J]. Advances in Neural Information Processing Systems, 2021, 34: 24604-24616.

[32] LI X, YANG Y, PAN H, et al. A novel deep stacking least squares support vector machine for rolling bearing fault diagnosis[J]. Computers in Industry, 2019, 110: 36-47.

[33] QIAN W, LI S, LU J. Adaptive nearest neighbor reconstruction with deep contractive sparse filtering for fault diagnosis of roller bearings [J]. Engineering Applications of Artificial Intelligence, 2022, 111: 104749.

[34] YU J, YAN X. A new deep model based on the stacked autoencoder with intensified iterative learning style for industrial fault detection [J]. Process Safety and Environmental Protection, 2021, 153: 47-59.

作者简介

刘钊, 2020 年于扬州大学获得学士学位, 现为燕山大学硕士研究生, 主要研究方向为智能故障诊断、深度学习理论及应用。
E-mail: liuzhaozz@foxmail.com

Liu Zhao received his B. Sc. degree in 2020 from Yangzhou University. Now he is a M. Sc. candidate at Yanshan University. His main research interests include intelligent diagnosis, deep learning and various applications.

孙洁娣(通信作者), 1998 年于河北师范大学获得学士学位, 2001 年于燕山大学获得硕士学位, 2008 年于天津大学获得博士学位, 现为燕山大学信息学院教授, 主要研究方向为智能故障诊断、深度学习理论及应用。
E-mail: wjtsjd@163.com

Sun Jiedi (Corresponding author) received her B. Sc. degree in 1998 from Hebei Normal University, M. Sc. degree in 2001 from Yanshan University and Ph. D. degree in 2008 from Tianjin University. She is now a professor with the School of Information Science and Engineering, Yanshan University. Her main research interests include intelligent diagnosis, deep learning and various applications.