GhostConv轻量级网络设计及故障诊断研究

赵志宏[1]，李春秀[2]，杨绍普[1]

（1.石家庄铁道大学省部共建交通工程结构力学行为与系统安全国家重点实验室，河北 石家庄 050043；
2.石家庄铁道大学信息科学与技术学院，河北 石家庄 050043）

摘要：提出一种GhostConv轻量级网络模型并将其用于故障诊断。GhostConv利用常规卷积生成一小部分特征图，然后在生成的特征图上进行多次特征提取来生成其余特征图，最大程度地节约了常规卷积中生成冗余特征图的成本，减少了模型参数，保证了模型的性能。采用连续小波变换对振动信号进行时频变换生成二维时频图，之后利用设计的GhostConv搭建轻量级网络模型进行故障诊断。采用凯斯西储大学轴承数据集进行验证，并与其他卷积结构网络模型进行参数量、计算量以及识别准确率的对比。实验结果表明，与其他模型相比，所使用的网络模型在参数量和计算量较少的条件下依旧有较高的识别精度，且具有较好的鲁棒性和泛化能力，具有一定的工程应用价值。

关键词：故障诊断；滚动轴承；轻量级网络；GhostConv；时频图
中图分类号：TH165⁺.3；TH133.33 **文献标志码：**A **文章编号：**1004-4523(2024)01-0182-09
DOI：10.16385/j.cnki.issn.1004-4523.2024.01.018

引 言

轴承作为旋转机械最重要的组成部分，在运行过程中出现故障会导致安全事故的发生，造成巨大的经济损失。因此，对滚动轴承的故障诊断越来越受到研究人员的重视[1]。目前，关于轴承故障诊断的研究已有多种方法，例如，Lu等[2]使用遗传算法和经验模式分解提取特征，然后使用支持向量机对故障进行分类和识别。Mao等[3]提出了一种结合多孔排列熵和支持向量机的诊断方法，对轴承故障类型进行分类。

随着计算机技术的发展，基于深度学习的智能故障诊断方法受到越来越多的关注[4]。这些方法将故障特征提取和特征分类相结合，从原始信号数据中自动提取出代表性特征，然后进行分类。在深度学习中，卷积神经网络、长短期记忆网络以及自编码器等神经网络在机械故障诊断领域的应用都取得了进展。侯文擎等[5]提出了一种改进堆叠降噪自编码器的方法，将其应用于轴承故障诊断中。Pan等[6]建立了基于长短期记忆网络和卷积神经网络的模型以进行轴承的故障诊断，取得了较好的诊断结果。

随着深度学习的快速发展，为了得到更高的故障诊断识别精度，模型变得越来越复杂，如VGG[7]，ResNet[8]等模型的参数量高达上百兆，在移动端执行深度模型推理任务时往往受限于智能移动终端的计算资源及存储资源而存在高延迟、高能耗等问题[9]，为了实现卷积神经网络在现实场景中的低延迟运行，轻量级卷积神经网络受到了研究者的关注[10]，一些典型的方法[11-13]被提出。目前基于网络轻量化的方法有模型压缩和轻量化卷积结构两种，模型压缩技术包括知识蒸馏、网络剪枝和权值量化[14]等。相比于模型压缩方法，基于轻量化卷积结构的网络得到了较多的关注和应用。目前，常用的轻量化卷积结构有分组卷积、深度可分离卷积等，与之相关的轻量级网络有MobileNet，Xception[15]，ShuffleNet等。这些卷积结构及网络也已被成功地应用在故障诊断中，如刘恒畅等[16]提出了一种基于多分支的深度可分离卷积网络模型，将其应用在滚动轴承故障诊断中，降低了模型的参数量，取得了较好的诊断结果。邓飞跃等[17]通过对ShuffleNet V2模型添加注意力机制，提高了模型的性能，得到了99%以上的故障诊断结果。

不同于轻量化卷积结构网络模型，SqueezeNet[18]主要采用小卷积核来减少参数量，在减小模型大小的基础上保持了相当的精度。MobileNet V1[19]网络模型则主要采用深度可分离卷积操作来减少模型的参数量，但是深度可分离卷积中

收稿日期：2022-04-16；**修订日期：**2022-06-30
基金项目：国家重点研发计划资助项目(2020YFB2007700)；国家自然科学基金资助项目(11972236，11790282)；石家庄铁道大学研究生创新资助项目(YC2022059)。

存在通道之间信息混合不充分的问题，会降低模型的性能。MobileNet V2[20]在延用深度可分离卷积的同时，对模型的结构进行了改进，提升了模型的性能，但是深度可分离卷积特征提取能力不足的问题依旧存在。基于此，ShuffleNet V1[21]采用分组卷积对深度可分离卷积进行改进，将逐点卷积与分组卷积进行结合，并引入通道混洗操作来弥补通道之间信息混合不充分的问题。ShuffleNet V2[22]网络依旧采用深度可分离卷积，通过对模型的结构进行通道划分等操作，在ShuffleNet V1模型的基础上降低了参数量，进一步提升了模型的性能。尽管如此，分组卷积和深度可分离卷积仍然存在着特征提取能力不足的问题。

GhostNet[23]模型提出了一种新的卷积结构——Ghost模块，为了降低模型的参数量和计算量，Ghost模块首先利用逐点卷积生成部分特征图，然后利用深度卷积获取其余特征信息，降低了常规卷积生成冗余特征的成本，同时获得了较好的模型性能，但其所用的逐点卷积依旧存在参数量大和特征提取能力不足的问题。为解决上述问题，本文在Ghost模块的基础上，提出一种GhostConv结构，将Ghost模块的逐点卷积改进为常规卷积，通过扩大卷积核以提取更多的特征信息；然后通过在同一特征图上进行多次特征提取操作来获得其余特征图，减少模型的参数量。

1　卷积结构

卷积作为模型的重要组成部分，可以有效地提取输入数据的特征信息。随着深度学习的发展，卷积神经网络模型的种类也越来越多，已发展出很多类型的轻量级卷积结构。

1.1　深度可分离卷积

分组卷积将输入特征图平均划分为G组，在每一组内进行卷积操作，之后将G组输出特征图进行拼接。深度可分离卷积是一种特殊的分组卷积，包括深度卷积和逐点卷积两部分，是轻量级网络中最常用的卷积方式之一，其分组数为输入特征数，具体结构如图1所示。从图1中可以看出，在深度卷积操作中，对输入样本的每个通道都进行了单通道卷积核的卷积操作，从而得到与输入通道数相同的特征图数，这样可以显著降低网络中的参数规模和运算量。在逐点卷积中利用卷积核大小为1×1的常规卷积操作，将不同通道在同一像素位置上的信息进行组合利用，并生成最终所需的特征图。因为逐点卷积卷积核比较小，所以深度可分离卷积存在模型特征提取能力不足的问题。

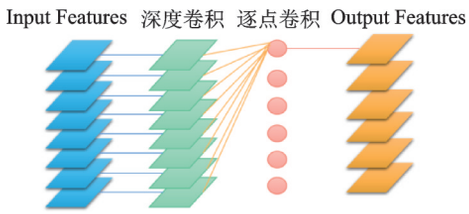

图1　深度可分离卷积结构

Fig. 1　Architecture of depthwise separable convolution

1.2　Ghost模块

常规卷积过程中，除了模型的参数量和运算量过大之外，其生成特征图的冗余度也很高。Ghost模块卷积结构为了减少常规卷积中特征冗余部分的参数，将普通的卷积层分解为两个部分，第一部分为逐点卷积操作，生成部分特征图，然后利用生成的特征图进行线性运算生成其余特征图，最后将两部分特征图进行拼接。Ghost模块结构如图2所示，其中，设输入特征数为8，输出特征数为8。Ghost模块中，第一步采用逐点卷积生成一半特征图，即图中的①部分，第二步根据第一步的特征图进行深度卷积生成另一半特征图，为②部分，最后进行特征图拼接生成完整的输出特征图。

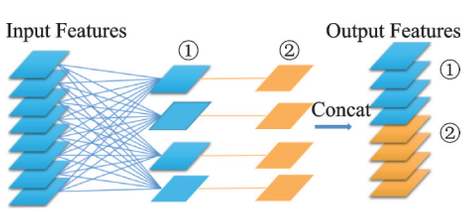

图2　Ghost模块结构

Fig. 2　Architecture of Ghost module

1.3　不同卷积复杂度分析

模型的参数量和计算量是评价模型复杂度的两个常用指标。模型的参数量是指网络中变量的总数，一般指可训练的参数。计算量则是指整个神经网络中所有浮点计算的总运算量。模型的参数量与计算量大小决定着模型的运行速度以及能否被应用在移动设备上。这里对以上不同卷积结构的参数量及计算量进行分析，为了分析的简便性，所有的分析都忽略了偏置项。

设输入的特征图维度大小为$H\times W\times C_1$，其中H，W和C_1分别代表特征图的高、宽和通道数，卷积核个数为C_0，尺寸为K，这样每一个卷积核大小为$K\times K\times C_1$，不同卷积结构的参数量与计算量如表1所示。对于常规卷积来说，模型的参数量是卷积核

大小与卷积核个数的乘积。分组卷积结构是将特征图分为G组，每一个维度大小是常规卷积的$1/G$，参数量也是常规卷积的$1/G$。深度可分离卷积分为两部分，第一部分为深度卷积，此时分组数为输入特征通道数，所以第一部分参数量为$K\times K\times C_1$，第二部分逐点卷积的参数量为$K\times K\times C_1\times C_0$，此时$K$为1。Ghost模块第一部分为常规卷积中的逐点卷积，所以参数量为$K\times K\times C_0\times C_0/2$，此时K为1。第二部分深度卷积参数量为$C_0/2\times K\times K$。模型的计算量为参数量大小与特征图大小的乘积。

表1 不同卷积复杂度分析

Tab. 1 Complexity analysis of different convolutions

卷积结构	卷积核大小	参数量	计算量
常规卷积	(K,K,C_1)	$K\times K\times C_1\times C_0$	$H\times W\times K\times K\times C_1\times C_0$
分组卷积	$(K,K,C_1/G)$	$K\times K\times C_1/G\times C_0$	$H\times W\times K\times K\times C_1/G\times C_0$
深度可分离卷积	$(K,K,1)+(1,1,C_1)$	$K\times K\times C_1+C_1\times C_0$	$H\times W\times C_1\times(K\times K+C_0)$
Ghost模块	$(1,1,C_1)+(K,K,1)$	$C_1\times C_0/2+(C_0-C_0/2)\times K\times K$	$H\times W\times C_1\times C_0/2+H\times W\times(C_0-C_0/2)\times K\times K$

2 GhostConv轻量级网络的故障诊断方法

为了更好地利用特征冗余部分来降低模型的参数量，本文借鉴Ghost模块卷积思想设计了GhostConv结构进行轻量级网络模型搭建。首先将Ghost模块的逐点卷积改为常规卷积，扩大卷积核以提取更多的空间信息特征；然后将第二部分设置为在同一特征图上进行多次特征提取的分组卷积，最后进行特征拼接。GhostConv结构在降低模型参数的同时又减少了分组卷积通道信息不交互带来的影响。

设置特定的比例来对第一部分生成的特征图个数与第二部分生成的特征图个数进行划分，采用字母s代表比例设定，当输出特征数为N时，则常规卷积生成N/s个特征图，分组卷积生成$N/s\times(s-1)$个特征图。这里以输入通道数和输出通道数为8、s为4为例，具体的GhostConv结构如图3所示。第一步进行常规卷积生成2个特征图，第二步在第一步特征图的基础上进行操作生成其余6个特征图，最后将两部分特征图进行拼接，这种在同一特征图上多次特征提取的操作可以更高效地降低常规卷积中的特征冗余部分参数量。

为了在降低模型参数量的同时保证模型的性能，对于s数值的选取既不能太大，也不能太小，s太大时，模型的性能会有所下降，s太小时，则模型的参数量过大。本文设计$s=8$的GhostConv卷积，第一步常规卷积的卷积核尺寸设置为3，第二步分组卷积卷积核的尺寸为3，步长为1。对于输入的特征图，先进行常规卷积操作生成输出通道数的1/8，之后采用生成的1/8部分进行分组卷积生成其余7/8部分特征图。为了更好地对比不同卷积结构的性能，本文参考MobileNet V1模型来设计简单的直线型模型结构，模型结构如图4所示，图中灰色部分为GhostConv结构，×8为采用8层GhostConv进行堆叠。模型第一层采用常规卷积结构进行特征提取，最后采用全连接层进行分类。具体的网络参数如表2所示。

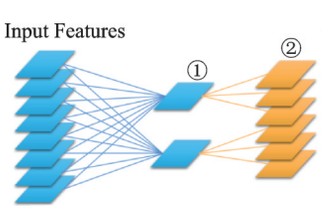

图3 GhostConv结构

Fig. 3 Architecture of GhostConv

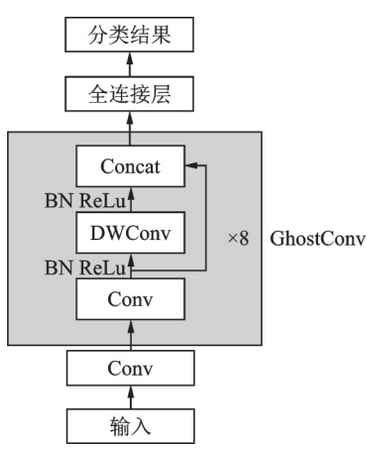

图4 模型结构

Fig. 4 Structure of model

本文设计的轻量级网络模型为二维模型，为了与实验数据兼容，参考文献[24]中的时频图转换方法将轴承一维数据进行连续小波变换生成二维时频图，将得到的时频图输入轻量级网络进行轴承故障诊断，具体流程如图5所示。

表 2　网络参数

Tab. 2　Network parameters

层数	网络层	重要参数
1	Conv1	通道数32、卷积核尺寸3、步长2
2	GhostConv1	通道数64、卷积核尺寸3、步长1
3	GhostConv2	通道数128、卷积核尺寸3、步长2
4	GhostConv3	通道数128、卷积核尺寸3、步长1
5	GhostConv4	通道数256、卷积核尺寸3、步长2
6	GhostConv5	通道数256、卷积核尺寸3、步长1
7	GhostConv6	通道数512、卷积核尺寸3、步长2
8	GhostConv7	通道数1024、卷积核尺寸3、步长2
9	GhostConv8	通道数1024、卷积核尺寸1、步长1
10	AvgPool	卷积核尺寸7、步长1
11	FC	神经元个数为故障类别数

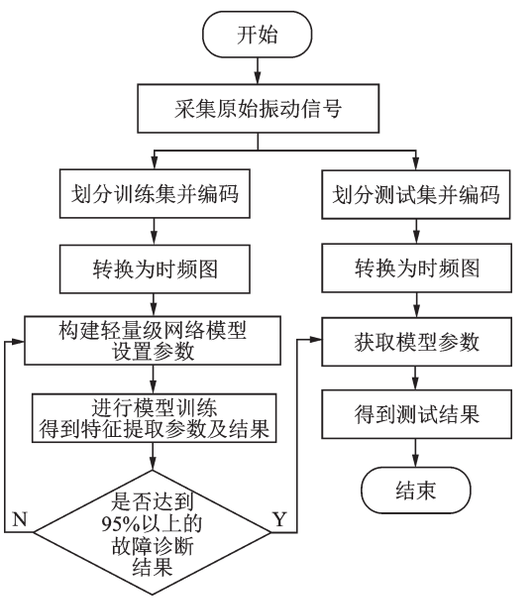

图 5　故障诊断流程

Fig. 5　Fault diagnosis process

3　实验数据

实验采用美国凯斯西储大学公开的轴承数据集[25]，故障类型包括轴承内圈损伤、外圈损伤、滚动体损伤3类，每种故障类型又包括0.1778 mm，0.3556 mm，0.5334 mm三种损伤程度，加上轴承健康状态共10类轴承工况数据，采样频率为12 kHz，轴承转速为1797 r/min。

基于轴承振动信号具有一定的周期性特征，需要选择合适的数据长度进行样本划分。实验选取约两个周期的数据长度864个采样点，针对每一工况，随机划分样本并对标签进行one-hot编码[26]。利用复值小波基函数对划分好的数据进行小波变换，生成时频图。以故障直径为0.1778 mm的内圈、外圈以及滚动体三处故障信号和正常信号为例，生成的时频图如图6所示，可以看出正常状态时频图和不同故障时频图明显不同。根据上述处理方法，每类数据生成300张时频图，其中240张为训练集，60张为测试集；10类数据训练集共2400张，测试集600张。

4　实验结果与分析

4.1　实验设置及结果

本实验在参数的选择上采用了深度神经网络研究中使用较多的网格搜索法。对于不同参数的设置进行了多次实验，最终设置批量大小为16，固定学习率值为0.0001，训练集的迭代次数为70。

基于网络模型轻量化的目的，需要对GhostConv结构中的s参数进行选择。s越大，GhostConv中分组卷积生成特征图的数量就越多，此时模型的参数量和计算量会越小，因此参数s的选择依据是在保证模型准确率的条件下使参数量和计算量最小。这里对参数s的变化进行实验分析，s从2开始，模型的参数量、计算量以及故障诊断实验结果如表3所示。从表3中可以看出，随着s的增大，模型的参数量和计算量不断减少；当s=8时，模型依旧有较好的故障诊断结果；当s继续增大时，模型参数量和计算量进一步减少，但是模型的故障诊断识别准确率也随之降低，因此，综合考虑之下，将GhostConv结构中的s参数设置为8。

将表2中层数为2～9的网络层分别替换为常规卷积和深度可分离卷积来进行CNN模型和DW-CNN模型的搭建。将三种不同卷积结构的模型进行实验对比，模型参数量、计算量及故障诊断结果如表4所示，表4中故障诊断识别准确率为5次实验结果的平均值。

从表4中可以看出，三种不同的卷积结构模型都得到了较好的故障诊断识别准确率，但是模型的参数量和计算量相差较大。DWCNN模型参数量是CNN模型参数量的1/8.8，本文模型参数量是CNN模型参数量的1/15.7；与DWCNN模型相比，本文模型的参数量更少，在模型计算量上，二者相差不多；但是与CNN模型相比，本文模型的计算量只有其1/8.5。实验结果表明，在故障诊断模型准确率差距不大的条件下，本文模型在参数量和计算量上更有优势。

本文模型的故障诊断模型准确率和损失率曲线如图7所示。从图7(a)中分析可知，模型在迭代30

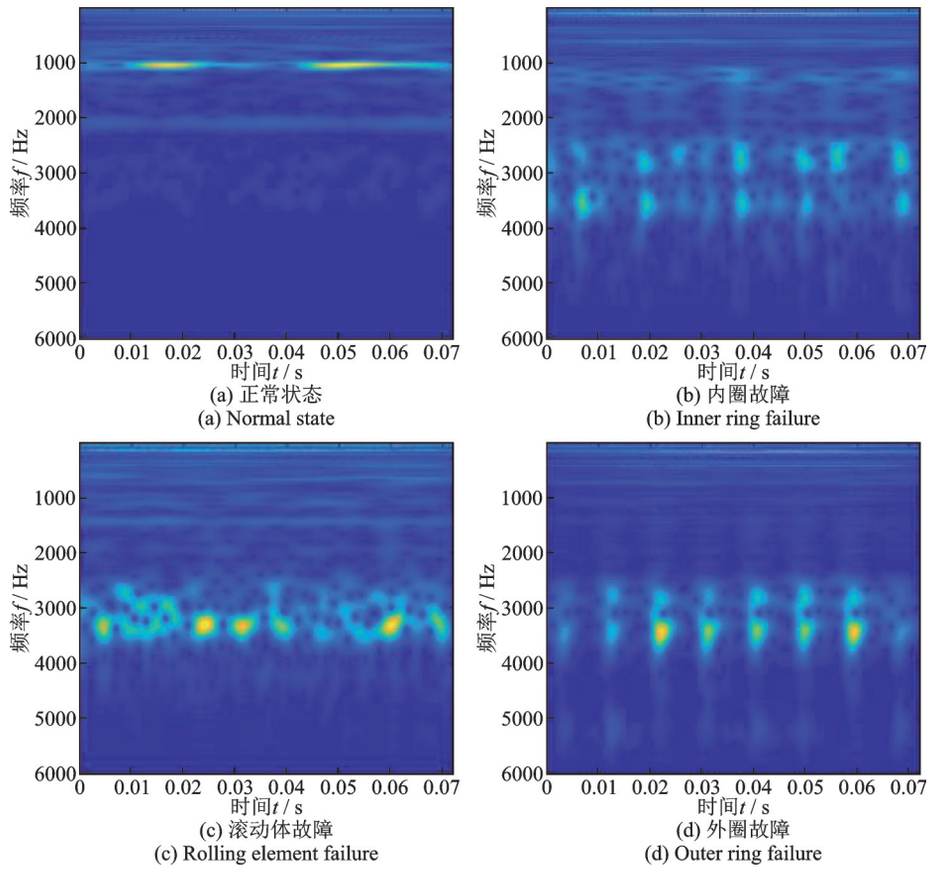

图 6 轴承不同工况下 Complex Morlet 小波时频图

Fig. 6 Complex Morlet wavelet time-frequency diagram of bearing under different working conditions

表 3 不同 s 时模型参数对比

Tab. 3 Comparison of model parameters at different s

比例 s	参数量/M	计算量/M	故障诊断准确率/%
2	4.07	1090	100
3	2.74	749.40	100
4	2.06	566.70	100
5	1.66	466.79	100
6	1.40	397.35	100
7	1.21	350.29	100
8	1.05	302.65	100
9	0.95	284.03	98.67

表 4 不同卷积结构模型参数对比

Tab. 4 Comparison of model parameters of different convolution structures

模型	参数量/M	计算量/M	故障诊断准确率/%
CNN	16.48	2560	100
DWCNN	1.87	318.48	99.60
本文模型	1.05	302.65	99.88

轮之前，训练集和测试集的准确率均呈上升趋势，但是存在较小的波动，在 30 轮之后，准确率趋于稳定并趋近于 1；从图 7(b)中可以看出，模型的损失率曲线一直呈下降趋势，趋近于 0。实验结果混淆矩阵如图 8 所示。从图 8 中可以看出，该模型在识别中仅

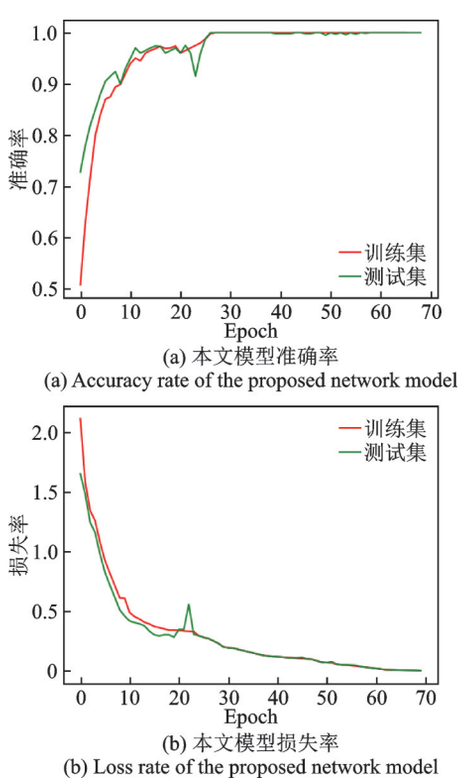

图 7 本文模型准确率及损失率

Fig. 7 The accuracy rate and loss rate of the proposed network model

出现 1 个错误，将 0.3556 mm 的滚动体故障误判为 0.5334 mm 的内圈故障，其余故障都被准确识别。

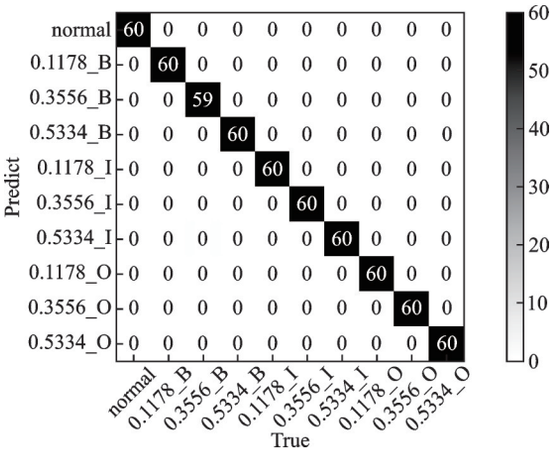

图 8　混淆矩阵

Fig. 8　Confusion matrix

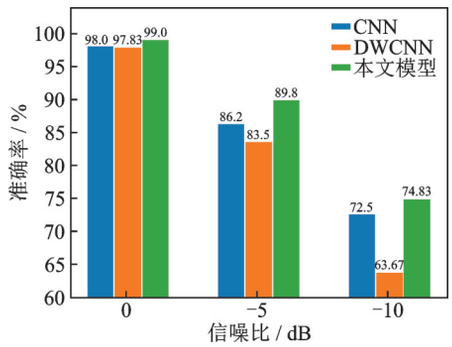

图 9　不同噪声程度下测试精度比较

Fig. 9　Comparison of measurement accuracy under different noise levels

4.2　加噪实验

在数据获取过程中，通过传感器所获得的振动信号会受到不同程度的噪声干扰，因此需要对故障诊断中模型对噪声的适应性进行分析和验证。为了探究本文模型在噪声情况下的特征提取能力，对原始数据混合不同强度的高斯白噪声，形成不同信噪比（Signal Noise Ratio，SNR）的信号来进行实验。SNR代表信号功率与噪声功率的比值，SNR越小，表明噪声功率越大。实验设置含噪信号的SNR分别为－10 dB、－5 dB以及0 dB，之后对含噪信号进行处理生成二维时频图。

4.2.1　不同卷积结构模型对比实验

本文设计的不同卷积结构模型在不同噪声程度下的故障诊断结果如图9所示。由图9可知，当原始振动信号混合信噪比为0 dB时，基于3种卷积结构的模型都能得到95%以上的故障识别准确率，表明模型具有一定抗噪能力。然而随着信噪比的下降，不同模型的准确率都有一定程度的下降，混合信噪比为－5 dB时，本文模型的故障识别准确率为89.8%，比DWCNN模型的准确率高6%左右，比CNN模型的准确率高3%。当混合信噪比下降为－10 dB时，所有方法的准确率均明显下降，此时DWCNN模型故障识别准确率最低，为63.67%，比CNN低8%左右，本文模型的故障识别准确率最高，为74.83%，表明本文模型的抗噪能力较好。

进一步对图9进行分析，在不同信噪比的条件下，本文模型和CNN模型都取得了相对较好的识别准确率，本文模型准确率略高，这表明本文模型的抗噪能力更好；而DWCNN模型在信噪比下降时，故障识别准确率最低，这是由于深度可分离卷积结构中的逐点卷积因为卷积核太小而很难提取时间维度上的相关信息特征。

4.2.2　与其他轻量级网络对比实验

以SNR＝－10 dB的含噪信号为分析对象，选取经典的轻量级网络MobileNet V2，ShuffleNet V2以及GhostNet与本文模型进行参数量、内存占用、计算量以及故障诊断识别准确率等方面的对比，进一步验证本文模型的有效性，具体的对比结果如表5所示。

表 5　与其他轻量级网络对比

Tab. 5　Comparison with other lightweight networks

模型	参数量/M	内存占用/M	计算量/M	准确率/%
本文模型	1.05	29.86	302.65	74.83
MobileNet V2	2.23	74.25	318.97	68.50
ShuffleNet V2	1.26	20.84	149.58	58.67
GhostNet	3.91	40.05	149.41	63.83

分析表5可知，ShuffleNet V2模型的内存占用最少，计算量也较低，但是其故障诊断识别准确率较差；GhostNet模型在计算量上较有优势，但是其参数量相对较多且故障诊断识别准确率与本文模型相差11%；MobileNet V2有相对较好的故障诊断识别准确率，但是在参数量、内存占用以及计算量上都没有明显的优势；本文模型在内存占用和计算量上并不是最优，但是本文模型参数量最少且故障诊断识别准确率最高。

4.3　泛化实验

为了进一步验证模型的性能，进行模型泛化性能实验。实验采用不同转速下的数据分别构建训练集和测试集。按照数据预处理方法对转速为1772 r/min和17950 r/min的10类数据集进行处理，生成二维时频图。不同卷积结构模型的实验结果如图10所示。在图10中，1797→1772代表采用转速为1797 r/min的数据集作为训练集，采用转速为1772 r/min的数据集作为测试集，AVG表示6组实验结果的平均值。

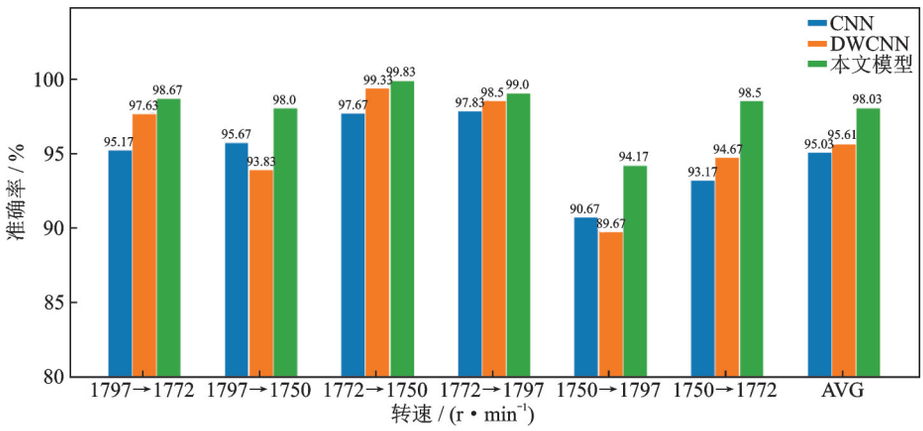

图 10 转速改变时模型的准确率

Fig. 10 Accuracy rate of the model when the rotating speed changes

由图 10 可知，在 6 组实验结果中，三种模型均在以转速为 1750 r/min 的数据集作为训练集的两组实验中得到了较差的结果，表明在低转速数据集进行训练时，模型的泛化性能较差。从图中还可以看出，DWCNN 模型在 1797→1750 上的准确率比 1797→1772 低 4% 左右，在 1750→1797 上的准确率比 1750→1772 低 5%，由两组实验结果可知，DWCNN 模型在转速相差较大时泛化性能较差。CNN 模型和 DWCNN 模型在几组实验中都取得了 95% 以上的平均准确率，而本文模型取得了 98.03% 的平均准确率，这是因为本文模型拥有较好的特征提取能力且模型参数量更少，在训练的过程中不容易发生过拟合。综上实验结果表明本文模型的泛化性能较好。

4.4 不同数据处理方法对比

为了进一步验证本文模型的有效性，选取不同的图像转换技术进行数据处理。对比方法包括短时傅里叶变换、灰度图以及格拉姆角差场(GADF)图像转换技术。得到的实验结果如表 6 所示。从表 6 中可以看出，灰度图以及短时傅里叶变换都得到了较好的实验结果，格拉姆角差场识别准确率相对较低，但也达到了 98% 以上的准确率，实验结果表明，在不同的数据处理方法下，本文模型都具有较好的特征提取能力。

表 6 不同数据处理方法对比

Tab. 6 Comparison of different data processing methods

数据处理方法	准确率/%
短时傅里叶变换	99.83
灰度图	100
GADF	98.75
本文模型	100

5 结 论

本文提出一种 GhostConv 轻量级网络的轴承故障诊断方法，进行了实验研究并与其他不同卷积结构构造的模型进行了对比分析，从中可以得到以下结论：

(1)通过与常规卷积、深度可分离卷积结构等搭建的模型进行实验对比，在模型结构相同的条件下，GhostConv 网络模型参数最少，仅为常规卷积模型参数量的 6%，为深度可分离卷积模型参数量的 56%。

(2)进行了不同噪声强度的故障诊断实验，实验结果表明，GhostConv 网络模型的抗噪能力最好，在高强度的噪声干扰下，得到了高于常规卷积模型 2% 以上的故障诊断识别准确率。

(3)选用不同转速的数据对模型进行泛化实验。实验结果表明，本文模型在测试数据集的转速发生改变时，依旧取得了较高的故障诊断识别准确率，具有较好的泛化性能。

(4)进行了不同数据处理方法的实验对比，结果表明，采用短时傅里叶变换、灰度图以及格拉姆角差场进行数据处理时，本文模型都具有较好的特征提取能力。

参考文献：

[1] Jiao J Y, Zhao M, Lin J, et al. Deep coupled dense convolutional network with complementary data for intelligent fault diagnosis[J]. IEEE Transactions on Industrial Electronics, 2019, 66(12): 9858-9867.

[2] Lu L, Yan J H, de Silva C W. Dominant feature selection for the fault diagnosis of rotary machines using mod-

ified genetic algorithm and empirical mode decomposition[J]. Journal of Sound and Vibration, 2015, 344: 464-483.

[3] Mao W T, He L, Yan Y J, et al. Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine[J]. Mechanical Systems and Signal Processing, 2017, 83(1): 450-473.

[4] Zhao R, Yan R Q, Chen Z H, et al. Deep learning and its applications to machine health monitoring[J]. Mechanical Systems and Signal Processing, 2019, 115: 213-237.

[5] 侯文擎，叶鸣，李巍华. 基于改进堆叠降噪自编码的滚动轴承故障分类[J]. 机械工程学报, 2018, 54(7): 87-96.
HOU Wenqing, YE Ming, LI Weihua. Rolling element bearing fault classification using improved stacked de-noising auto-encoders[J]. Journal of Mechanical Engineering, 2018, 54(7): 87-96.

[6] Pan H H, He X X, Tang S, et al. An improved bearing fault diagnosis method using one-dimensional CNN and LSTM[J]. Strojniški vestnik-Journal of Mechanical Engineering, 2018, 64: 443-452.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014.

[8] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 770-778.

[9] Qin Qing, Ren Jie, Yu Jialong, et al. To compress, or not to compress : characterizing deep learning model compression for embedded inference[C]//Proceedings of the 2018 IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom). Melbourne, Australia, 2018: 729-736.

[10] 葛道辉，李洪升，张亮，等. 轻量级神经网络架构综述[J]. 软件学报, 2020, 31(9): 2627-2653.
GE Daohui, LI Hongsheng, ZHANG Liang, et al. Survey of lightweight neural network[J]. Journal of Software, 2020, 31(9): 2627-2653.

[11] He Y H, Lin J, Liu Z J, et al. AMC: AutoML for model compression and acceleration on mobile devices[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 784-800.

[12] Zoph B, Vasudevan V, Shlens J V, et al. Learning transferable architectures for scalable image recognition[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE, 2018: 8697-8710

[13] Tan M X, Chen B, Pang R M, et al. MnasNet: platform-aware neural architecture search for mobile[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 2820-2828.

[14] HAN S, MAO H Z, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding[J]. Computer Science, 2015.

[15] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258.

[16] 刘恒畅，姚德臣，杨建伟，等. 基于多分支深度可分离卷积神经网络的滚动轴承故障诊断研究[J]. 振动与冲击, 2021, 40(10): 95-102.
LIU Hengchang, YAO Dechen, YANG Jianwei, et al. Fault diagnosis of rolling bearings based on a multi branch depth separable convolutional neural network [J]. Journal of Vibration and Shock, 2021, 40(10): 95-102.

[17] 邓飞跃，吕浩洋，顾晓辉，等. 基于轻量化神经网络Shuffle-SENet的高速动车组轴箱轴承故障诊断方法[J]. 吉林大学学报(工学版), 2022, 52(2): 474-482.
DENG Feiyue, LÜ Haoyang, GU Xiaohui, et al. Fault diagnosis of high-speed train axle bearing based on a lightweight neural network Shuffle-SENet[J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(2): 474-482.

[18] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: Alex-Net-level accuracy with 50× fewer parameters and <0.5 MB model size[J]. arXiv preprint arXiv: 1602.07360, 2016.

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

[20] Sandler M, Howard A, Zhu M L, et al. MobileNet V2: inverted residuals and linear bottlenecks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE, 2018: 4510-4520.

[21] Zhang X Y, Zhou X Y, Lin M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[22] Ma N N, Zhang X Y, Zheng H T, et al. ShuffleNet

V2: practical guidelines for efficient CNN architecture de sign[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 116-131.

[23] Han K, Wang Y H, Tian Q, et al. Ghostnet: more features from cheap operations[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 1580-1589.

[24] 张龙,甄灿壮,熊国良,等.基于深度时频特征的机车轴承故障诊断[J].交通运输工程学报,2021,21(6):247-258.
ZHANG Long, ZHEN Canzhuang, XIONG Guoliang, et al. Locomotive bearing fault diagnosis based on deep time-frequency features[J]. Journal of Traffic and Transportation Engineering, 2021, 21(6): 247-258.

[25] Case Western Reserve University Bearing Data Center [EB/OL].2018. https://csegroups.case.edu/bearing-datacenter/pages/download-data-file.

[26] HO T H, AHN K K. Modeling and simulation of hydrostatic transmission system with energy regeneration using hydraulic accumulator[J]. Journal of Mechanical Science and Technology, 2010, 24(5): 1163-1175.

GhostConv lightweight network design and research on fault diagnosis

ZHAO Zhi-hong[1], *LI Chun-xiu*[2], *YANG Shao-pu*[1]

(1.State Key Laboratory of Mechanical Behavior and System Safety of Traffic Engineering Structures, Shijiazhuang Tiedao University, Shijiazhuang 050043, China; 2.School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China)

Abstract: With the advent of the era of big data, the mechanical equipment fault diagnosis method based on deep learning has attracted more attention. However, the traditional deep network model seriously limits its application in practical engineering due to the excessive amount of parameters and calculations. Based on this, a GhostConv lightweight network model is proposed and used for fault diagnosis. GhostConv generates a small part of the feature maps through conventional convolution, and performs multiple feature extraction on the generated feature maps to generate the remaining feature maps. Contact the feature maps of the two parts to obtain a complete feature map. GhostConv structure saves the cost of generating redundant feature maps in conventional convolution to the maximum extent, and reduces the model parameters to ensure the performance of the model. In the experiment, the continuous wavelet transform is used to transform the vibration signal to generate a two-dimensional time-frequency diagram, and then the designed GhostConv is used to establish a lightweight fault diagnosis network model. The original dataset and noisy dataset of Case Western Reserve University are used for experimental verification, and compared with the conventional convolution structure network model and depth separable convolution structure model in terms of parameters, calculation and recognition rate. The experimental results show that the GhostConv lightweight network model still has high recognition accuracy and strong anti-noise ability under the condition of fewer parameters and calculations with good robustness and generalization ability. The parameters of the model are only 6% of the conventional convolution model and 56% of the deep separable convolution model. Under the condition of strong noise interference, the fault diagnosis and recognition rate is still higher than that of the conventional convolution model, which confirms its engineering application value.

Key words: fault diagnosis; rolling bearing; lightweight network; GhostConv; time-frequency diagram

作者简介: 赵志宏(1972—),男,博士,教授。E-mail: hb_zhaozhihong@126.com。