基于注意力机制 – Inception – CNN 模型的滚动轴承故障分类

朱 浩[1]，宁 芊[1,2]，雷印杰[1]，陈炳才[3]，严 华[1]
(1. 四川大学 电子信息学院，成都 610065；2. 新疆师范大学 物理与电子工程学院，乌鲁木齐 830054；
3. 大连理工大学 计算机科学与技术学院，辽宁 大连 116024)

摘 要： 针对传统机器学习方法需要大量专家知识和高昂经济成本，研究了一种基于注意力机制和 Inception 网络结构的卷积神经网络。其注意力机制是对数据的不同特征维度赋予不同的权重，抽取出更加关键和重要的信息，使模型做出更加准确的判断。其 Inception 网络结构则是拓宽网络的宽度并增加网络对卷积核尺度的适应性，以提取到更加丰富的特征。为了提高模型的泛化能力，在每个卷积层和全连接层后又添加了一个 DropBlock 层。最后结果显示该模型不仅在同负载的情况下获得很高的滚动轴承故障分类准确率和稳定性，并且在不同负载情况、不同规模的滚动轴承数据集上依旧能保持高的准确率与稳定性。

关键词： 卷积神经网络；注意力机制；Inception 网络结构；DropBlock；故障诊断
中图分类号： TH165 + .3 **文献标志码：** A DOI: 10.13465/j.cnki.jvs.2020.19.013

Fault classification of rolling bearing based on attention mechanism – inception – CNN Model

ZHU Hao[1], *NING Qian*[1,2], *LEI Yinjie*[1], *CHEN Bingcai*[3], *YAN Hua*[1]
(1. School of Electronic Information Engineering, Sichuan University, Chengdu 610065, China;
2. School of Physics and Electronic Engineering, Xinjiang Normal University, Urumqi 830054, China;
3. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: Aiming at traditional machine learning method requiring a lot of expert knowledge and high economic costs, a convolutional neural network (CNN) based on attention mechanism and Inception network structure was proposed. Its attention mechanism was used to assign different weights to different feature dimensions of data, extract more critical and important information, and make the model do more accurate judgments. Its Inception network structure was used to broaden the network width, increase the network's adaptability to the convolution core scale, and extract more abundant features. In order to improve the generalization ability of the model, a DropBlock layer was added to follow each convolution layer and full connection layer. Finally, the results showed that the proposed model can not only achieve higher fault classification accuracy and stability under the same load, but also maintain higher accuracy and stability under different load conditions and rolling bearing data sets with different scales.

Key words: convolutional neural network (CNN); attention mechanism; Inception network structure; DropBlock; fault diagnosis

滚动轴承是旋转机械装置的常用零部件，它的正常运行对整个装置起着至关重要的作用，一旦滚动轴承出了故障，则极有可能造成巨大的经济损失甚至引起安全险情[1]。因此，及时诊断滚动轴承的工作状态便显得格外重要。

传统的故障诊断方法主要分为人工提取特征和故障分类两步。用于从原始的一维数据中提取特征的常用方法有时域统计分析[2]，小波变换[3]等，接着一般会经过特征选择用以去除冗余特征和不敏感特征，这时采用的方法主要是 PCA[4] 及其变形方法，最后用通过筛选后的特征对诸如 KNN[5]，SVM[6] 等算法进行训练得到模型，然后用测试集进行测试来验证模型是否有效。

近年来随着深度学习[7]的崛起，该方法被越来越多的应用到故障诊断领域。孙文珺等[8]利用稀疏自编码器有效的实现了感应电动机的故障诊断。Kong 等[9]采用深度卷积神经网络对滚动轴承的能量进行监控，并以此进行电机故障的分类。雷亚国等[10]提出了

基金项目：国家自然科学基金(61771089)
收稿日期：2019 – 01 – 17 修改稿收到日期：2019 – 07 – 19
第一作者 朱浩 男，硕士，1993 年生
通信作者 宁芊 女，博士，副教授，1969 年生

一种基于深度学习的机械装备大数据健康监测方法，该方法通过建立深层模型，摆脱了对大量信号处理技术与诊断经验的依赖，直接从频域信号中自适应地提取故障特征，实现大量数据下故障特征的自适应提取与健康状况的智能诊断。

Inception 网络结构和注意力机制作为深度学习发展过程中的重大突破，被应用在诸多领域，展示了其巨大的优越性并表明其在故障诊断领域中进行应用的可能。Szegedy 等[11] 通过对 Inception 网络结构进行堆叠组成 GoogLeNet 使之运用在图像分类领域并取得了当时最好成绩。Chen 等[12] 将注意力机制运用在 CNN 上，提出一个新颖的卷积神经网络 SCA-CNN，在 CNN 中加入了 Spatial Attention 和 Channel-wise Attention 两种注意力机制进而完成图像标注任务。

正如上述所言，深度学习中的方法越来越多的被运用到故障诊断领域，虽然这些方法大多数结果的准确率都会比传统机器学习方法高，但也会有自身的限制和缺点。以往多数提出的深度学习模型的泛化能力不强，往往只能适用于当前负载情况下的零部件，一旦负载情况有所变化（如噪声变化，工作环境变化等），准确率就会降低。另外有的网络模型设计的层数过多，这样一方面网络越深则越难以训练甚至发生退化（Degradation），另一方面网络越深则越容易过拟合出现训练集准确率高而测试集准确率低的情况。一维故障数据相较于图片或是其它高维数据来说，是相对简明的，清晰的，所以我们可以通过相对简单的模型来充分表达数据背后的信息。同时有的网络模型的输出结果稳定性不理想，准确率难以一直保持在一个高水平线上。

针对以上问题，本文研究提出了一种卷积神经网络，它能够有效解决上述问题。其主要贡献有：① 提出了一种简单高效的端对端卷积神经网络框架，它直接提取一维原始数据作为输入，并能够对数据进行诊断且直接给出分类结果；② 该框架不仅能够在同负载情况下获取高准确率，它还能够在不同负载情况下保持高准确率；③ 该框架结构相对简单，易于训练，能够很好的保证测试集准确率的稳定性。

1　相关模型简介

卷积神经网络（Convolutional Neural Network，CNN）是一种常见的深度学习网络架构，受生物自然视觉认知机制启发而来。20 世纪 90 年代，Le Cun 等确立了 CNN 的现代结构。本文模型是一个融合了 Inception 网络结构和注意力机制的卷积神经网络，同时为了增强泛化能力，也添加了 DropBlock 层以适应不同的负载情况。下面将简要介绍模型所用到的几种网络结构。

1.1　卷积层（Convolution Layer）

卷积层的作用主要是学习输入数据的特征表示，卷积层由很多的卷积核（Convolutional Kernel）组成，卷积核用来计算不同的特征图（Feature Map）。激活函数（Activation Function）给卷积神经网络引入了非线性，常用的有 sigmoid、tanh、Relu 函数。具体的卷积层公式如下所示

$$y^{l(i,j)} = \sum_{j'=0}^{n-1} W_i^{l(j')} x^{l(j+j')} \tag{1}$$

式中：$W_i^{l(j')}$ 为第 l 层的第 i 个卷积核的第 j' 个权值；$x^{l(j+j')}$ 为第 l 层中第 j 个被卷积的局部区域；n 为卷积核的宽度；$y^{l(i,j)}$ 为对应的输出。$y^{l(i,j)}$ 经过激活函数（这里我们采用 Relu 函数）的公式为

$$a^{l(i,j)} = f(y^{l(i,j)}) = \max\{0, y^{l(i,j)}\} \tag{2}$$

1.2　池化层（Pooling Layer）

池化层的作用主要是降低卷积层输出的特征向量，同时改善结果，使结构不容易出现过拟合。典型的操作包括平均池化和最大化池化。通过卷积层与池化层，可以获得更多的抽象特征。具体的池化层公式如下所示（这里我们选择常用的最大池化层）

$$p^{l(i,j)} = \max_{(j-1)n+1 \leqslant t \leqslant jn} \{a^{l(i,t)}\} \tag{3}$$

式中：$a^{l(i,t)}$ 为第 l 层第 i 帧第 t 个神经元的激活值；n 为池化区域的宽度；$p^{l(i,j)}$ 为相应的池化层输出。

1.3　全连接层（Fully Connected Layer）

将卷积层和池化层堆叠起来以后，就能够形成一层或多层全连接层，在整个卷积神经网络中起到“分类器”的作用。全连接层运用的激活函数多是 sigmoid 或 softmax 函数。具体的全连接层正向传播公式如下

$$y^{l+1(j)} = \sum_{i=1}^{n} W_{ij}^{l} x^{l(i)} + b_j^{l} \tag{4}$$

式中：W_{ij}^{l} 第 l 层第 i 个神经元与第 $l+1$ 层第 j 个神经元之间的权值；$x^{l(i)}$ 代表第 l 层第 i 个神经元的值；b_j^{l} 第 l 层所有神经元对第 $l+1$ 层第 j 个神经元的偏置值；$y^{l+1(j)}$ 代表第 $l+1$ 层第 j 个输出神经元的 logits 值。$y^{l+1(j)}$ 经过 softmax 激活函数的公式为

$$a^{l+1(j)} = \frac{e^{y^{l+1(j)}}}{\sum_j e^{y^{l+1(j)}}} \tag{5}$$

1.4　Inception 网络结构

Inception 网络结构也是一个卷积网络，它的提出主要是考虑了多个不同大小的卷积核能够增强网络的适应力并提取到更加丰富的不同尺度的特征。同时通过采用 NIN[13] 模型，Inception 网络结构能够大大减小模型的参数使网络在不损失模型特征表示能力的前提下，尽量减少卷积核的数量，达到降低模型复杂度的目的。本文模型的 Inception 网络结构层被设计成由 NIN 和卷积核大小分别为 3，5，7 的卷积层组成，这样既能

保证获得不同尺度的特征，也能保证模型结构保持相对简单，不会过于复杂。

1.5 注意力机制(Attention Mechanism)

注意力机制通过借鉴人类的注意力方式被广泛使用在自然语言处理、图像识别及语音识别等各种不同类型的深度学习任务中，是深度学习技术中最值得关注与深入了解的核心技术之一。

1.6 DropBlock 层

DropBlock 是由 Ghiasi[14] 等提出的一种正则化方法。作者认为原来的 Dropout 方法多是作用在全连接层上，在卷积层应用 Dropout 方法作用不大，原因是每个特征的位置都有一个感受野范围，仅仅对单个元素位置进行 Dropout 并不能降低特征学习的范围，也就是说网络仍可以通过该位置的相邻位置元素去学习对应的语义信息，也就不会促使网络去学习更加鲁棒的特征。所以 DropBlock 通过设置大小来对一块一块的元素去 Dropout 就可以学习到更加鲁棒的特征，这样就可以在全连接层和卷积层同时使用 DropBlock 方法。当 DropBlock 大小设置为 1 时，它就相当于是普通的 Dropout。

2 注意力机制-Inception-CNN 模型结构

2.1 注意力机制的引入

在本文模型中，区别于 SCA-CNN 我们设计的注意力机制模型主要集中在对特征维度这一层面进行处理。在原始故障数据进入一个卷积层后便会得到不同的特征，这时为了衡量这些特征的重要程度，便引入了注意力机制。在本文中，我们将经过卷积层后得到的特征送入 softmax 函数中，它能够通过计算得到不同特征维度的各自的权重系数，然后得到的系数在与输入的特征进行对应元素相乘组成新的特征，而这些新特征便是经过权重分配的特征，能够帮助模型获得更关键和重要的信息，使模型做出更加准确的判断。也即注意力机制能够将更多的注意力集中在更加重要的特征上。

2.2 模型结构

模型的整体结构见图 1。模型以一维故障数据作为输入，然后进入一个拥有大卷积核的卷积层，该层的作用是为了能够在原始震动信号的中低频段获得更多更有效的信息[15]。Srivastava 等[16] 提出，Dropout 可以被用作一种添加噪声的方法并克服过拟合，所以我们在每个卷积层和全连接层的后面接入了一层 DropBlock 层，一方面可以模拟噪声，另一方面可以通过 DropBlock 来提高模型的泛化能力。在经过 DropBlock 层后，选择接入了最大池化层，它可以非常有效地缩小参数矩阵的尺寸，从而减少最后全连层中的参数数量帮助简化模型，同时使用最大池化层也有助于加快计算速度。可以把卷积层，DropBlock 层，最大池化层合在一起称之为一个卷积块。在输入数据经过第一个卷积块处理之后，已经获得了不同特征维度信息，但为了获得更多以及更高维度的特征信息和给更重要的特征赋予更多的权重，我们将特征信息引入注意力层，然后得到的新张量做为下一个卷积块(Inception 块)的输入。第二个卷积块由 Inception 层，DropBlock 层和最大池化层组成，其中 Inception 层由大小为 3，5，7 的卷积层，大小为 3 的最大池化层和用于简化网络减少参数的 NIN 层构成，通过对这些卷积层核池化层的拼接联合，拓宽了网络的宽度并提取到了多尺度和更加丰富的特征。接着进入的 DropBlock 层和最大池化层的作用与第一个卷积块的作用相同。最后通过一个全连接层把前两个卷积块得到的局部特征重新通过权值矩阵组装成完整的图，在经过 DropBlock 层后传入 softmax 层得到结果分类输出。

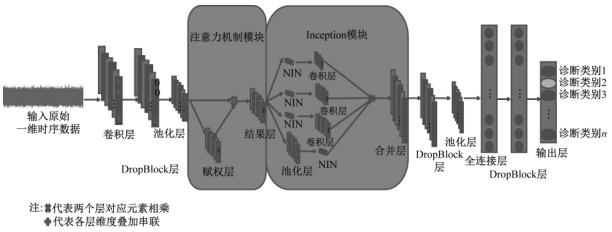


图 1 注意力机制-Inception-CNN 模型结构

Fig. 1 Attention mechanism-Inception-CNN model structure

该模型横向只有两个卷积层，结构简单，因此不用担心训练时发生退化现象。

2.3 模型训练

信号在神经网络的输出应该与目标值一致，评价这种一致性的函数叫做损失函数(Loss Function)，本模型采用的损失函数是交叉熵损失函数(Cross-entropy Loss Function)。假设信号在神经网络的 softmax 输出为 $q(x)$，其目标值为 $p(x)$。则交叉熵损失函数为

$$\text{Loss} = -\sum_{x}\left(p(x)\lg q(x) + (1-p(x))\lg(1-q(x))\right) \tag{6}$$

训练的目标就是尽可能地降低 Loss，Loss 越低表明模型的输出值与真实值就越接近。为了更好地训练模型，采用了 Adam[17] 优化器来对模型进行优化，它是一种可以替代传统随机梯度下降过程的一阶优化算法，基于训练数据迭代地更新神经网络权重，是一种高效简洁的优化器。

3 试验与性能分析

3.1 多故障类别多负载数据集试验

本节的试验数据来自于美国凯斯西储大学(CWRU)滚动轴承数据中心[18]，它是世界公认的轴承

故障诊断标准数据集。CWRU 轴承中心数据采集系统如图 2 所示。本试验的试验对象为图中的驱动端轴承，型号为 SKF6205，系统的采样频率为 12 kHz，故障由电火花加工制作而成。被诊断的轴承一共有 3 种缺陷位置，分别是滚动体损伤，外圈损伤与内圈损伤，损伤直径的大小分别为包括 0.007 inch，0.014 inch 和 0.021 inch，共计 9 种损伤状态。同时加上正常状态，所以一共有十种故障分类状态。

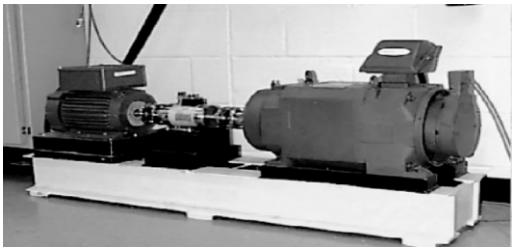


图 2　CWRU 轴承中心数据采集系统

Fig. 2　Data acquisition system of bearing center of CWRU

试验中为了加快卷积网络的训练，对数据进行归一化处理，公式如下

$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{7}$$

试验一共准备了 3 个数据集。数据集 A，B 和 C 分别是在负载为 1 hp，2 hp 和 3 hp 下的数据集，每次采集的样本长度为 2 048。每个数据集分别包含训练样本(用以训练估计模型)，验证样本(用以调整模型的相关参数)和测试样本(用以测试模型性能的好坏)，具体数据集描述见表 1。

表 1　CWRU 试验数据集描述

Tab. 1　Description of CWRU data set

损伤位置		外圈损伤			内圈损伤			滚动体损伤			正常
损伤直径/inch		0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021	0
数据集 A	训练集	700	700	700	700	700	700	700	700	700	700
	验证集	200	200	200	200	200	200	200	200	200	200
	测试集	100	100	100	100	100	100	100	100	100	100
数据集 B	训练集	700	700	700	700	700	700	700	700	700	700
	验证集	200	200	200	200	200	200	200	200	200	200
	测试集	100	100	100	100	100	100	100	100	100	100
数据集 C	训练集	700	700	700	700	700	700	700	700	700	700
	验证集	200	200	200	200	200	200	200	200	200	200
	测试集	100	100	100	100	100	100	100	100	100	100

3.1.1　参数设置

当我们配置参数时，需要注意到虽然模型越复杂可能会提高模型的表达能力，但同时也可能会导致模型有过拟合和训练难度加大的缺点，因此我们应该在参数设置上保持一种平衡，具体的参数设置由试验和故障诊断领域的相关经验决定。如由于故障数据是一维时序数据，所以第一个卷积层不仅需要大的卷积核同时也需要大的步幅，如此做是为了让网络的感受野至少大于一个周期，以便于获得与相位无关的特征。这样做才能提取到更加关键有效的特征信息以帮助模型做出更精准的判断。DropBlock 层的丢失块的大小采用原论文默认的参数设置为 7。同时为了能够降低参数量，要求 Inception 网络的每个 NIN 模块的数量少于上层的输出维度。池化层的数量都是和上层的卷积层数量保持一致。全连接层的大小原则上不能少于分类数。

有了上述的选择参数的思路，经过重复的试验比较，最终得到了以下的参数设置，见表 2。

Keskar 等[19] 证明，使用更小的 batch size 训练神经网络有助于增强模型的泛化能力。所以在模型训练时我们选择的 batch size 为 32 以增强模型的泛化能力。Xavier 等[20] 提出了一种名叫 Xavier 均匀初始化的方法，文章主要的目标就是使得每一层输出的方差应该尽量相等，近年来许多深度学习模型都使用它作为神经网络的初始化方法并取得了不错的效果，因此本文也采用 Xavier 均匀初始化的方法对神经网络进行初始化且一共进行 20 个 epoch 以充分训练数据。在学习率选择上面，我们设置一开始的学习率为 0.001，当验证集的损失在两个 epoch 不在下降时，我们将学习率折半以进一步训练。学习率设置的最小值为 0.000 01。

3.1.2　同负载故障诊断结果及与其它方法的比较

试验中，由于神经网络的初始权值是随机生成，且每次的输入数据也是随机获取，所以为了验证本方法的稳定性，每个试验重复进行 10 次。为了验证本方法是否能够获得较高的准确率和较好的稳定性，与普通的多层感知机和当前几种以一维数据作为输入的卷积网络方法作为对比。参与比较的方法的各网络结构如表 3 所示。

表 2 本文模型参数

Tab. 2 The parameters of this model

层数	层的类型	(卷积核)大小	卷积核步幅	卷积核数量	保留率	丢失块大小
1	卷积层	32 × 1	16 × 1	32		
2	DropBlock 层				0.5	7
3	池化层	4 × 1	4 × 1	32		
4	(注意力模块)					
	全连接层	32				
5	(Inception 模块)					
	NIN/卷积层	1 × 1/3 × 1	1 × 1	24/48		
	NIN/卷积层	1 × 1/5 × 1	1 × 1	24/64		
	NIN/卷积层	1 × 1/7 × 1	1 × 1	16/64		
	池化层/NIN	3 × 1/1 × 1	1 × 1	32/24		
6	DropBlock 层				0.5	7
7	池化层	2 × 1	2 × 1	200		
8	全连接层	500				
9	DropBlock 层				0.5	1
10	(Softmax 输出层)					
	全连接层	10				

表 3 各方法网络结构

Tab. 3 Network structure of each method

方法	网络结构
普通 MLP	Input [2048] →FC [64] →FC [32] →FC [10]
Ince 等[21] 的工作	Input [240] →60C [9] →60P [4] →40C [5] →40P [4] →40C [5] →40P [4] →FC [20] →FC [10]
Abdeljaber 等[23] 的工作	Input [128] →64C [41] →64P [2] →32C [41] →32P [2] →FC [10] →FC [10]
Zhang 等[15] 的工作	Input [2048] →16C [64] →16P [2] →32C [3] →32P [2] →3 [64C [3] →64P [2]] →FC [100] →FC [10]

普通 MLP 是为了对比卷积网络而自行设计的一个模型，它通过简单三个全连接层串联而成，属于最简单的神经网络。其它方法都是近年来取得不错效果的卷积神经网络，Ince 等主要设计了一个三层卷积层的卷积神经网络，Abdeljaber 等主要设计了一个两层卷积层的卷积神经网络，Zhang 等则设计了一个较为复杂的网络，共有 5 个卷积层。我们的方法共有两层卷积层，其中第二层为 Inception 层，对比上述三个方法，我们的模型比 Zhang 等设计的简单，复杂度大致和 Ince 等的模型相当，但又因为层数只有两层更加容易训练，不会发生梯度爆炸或梯度消失的情况，也即我们的模型框架相对简单，容易训练。各试验对比的详细结果如表 4 和表 5。

表 4 各方法平均准确率

Tab. 4 Average accuracy of each method

方法	A	B	C
普通 MLP	91.85%	91.76%	96.32%
Ince 等的工作	99.01%	99.34%	99.54%
Abdeljaber 等的工作	98.08%	98.34%	98.74%
Zhang 等的工作	98.64%	99.21%	99.79%
本文方法	99.23%	100.00%	99.98%

表 5 各方法稳定性

Tab. 5 Stability of each method

方法	A	B	C
普通 MLP	91.65% ±1.25%	92.15% ±2.35%	95.60% ±2.00%
Ince 等的工作	98.95% ±0.75%	99.00% ±1.00%	99.55% ±0.45%
Abdeljaber 等的工作	97.75% ±1.15%	97.50% ±2.30%	98.45% ±1.15%
Zhang 等的工作	98.55% ±1.25%	96.05% ±13.95%	99.05% ±0.95%
本文方法	99.20% ±0.20%	100.00% ±0.00%	99.95% ±0.05%

通过表 4 我们发现，当训练表和测试表处于同负载的情况下，即便是简单的 MLP 模型也能取得不错的成绩。而上面的几种卷积神经网络都相较于 MLP 而言又有了进一步提升，Abdeljaber 等的模型在这几个卷积模型中最为简单且准确率也最低，所以可能是因为模型对数据的表达能力有些许不足，但同时 Zhang 等的模型最为复杂却不能获得最好的效果，说明有可能发生了过拟合。相比而言，本文方法的准确率达到了最

高且同时模型保持相对简单。

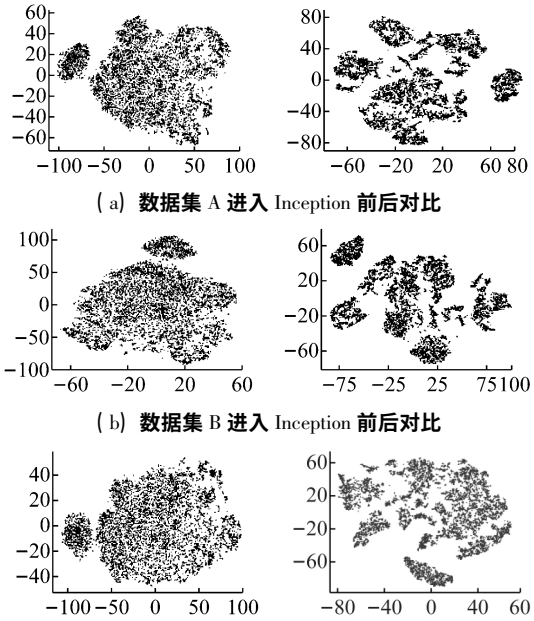
通过表 5 可以清楚地观测到，在同负载环境下各模型的稳定性情况。本文方法每次测出准确率的偏差不超过 1%，表现优于其它几种方法。

同时为了直观上了解 Inception 的作用，我们将训练数据在进入 Inception 层之前和之后做了可视化表达，结果如图 3。

我们可以发现，在经过第一个卷积层后，此时模型可以将数据分为两类，分别为正常数据和故障数据，表明模型已有初步的诊断能力，但并没有能够将故障大类进行细分，在经过 Inception 层后，模型有效的将故障大类进行细分，划分为不同的小块，结合试验对比结果来看，体现出了 Inception 层的有效性和实用性。

3.1.3　不同负载环境下故障诊断结果及与其它方法的比较

由上述结果发现，在同负载情况，上述各方法都能获得不错的准确率。为此，我们试验了在不同负载环境下各方法的准确率情况。具体来说，我们分别用数据集 A，B，C 来训练样本，然后在其余的数据集上测试样本，如 A→B 代表在数据集 A 上训练样本，在数据集 B 上进行测试，同样也是每个方法试验十次。详细结果如表 6 和表 7。



(a) 数据集 A 进入 Inception 前后对比

(b) 数据集 B 进入 Inception 前后对比

(c) 数据集 C 进入 Inception 前后对比

图 3　CWRU 数据集可视化

Fig. 3　Visualization of CWRU data set

表 6　各方法不同负载情况下的平均准确率

Tab. 6　Average accuracy under different loads of each method

方法	A→B	A→C	B→A	B→C	C→A	C→B
普通 MLP	33.67%	32.38%	38.10%	35.33%	39.83%	34.32%
Ince 等的工作	98.33%	91.98%	93.21%	88.34%	81.09%	86.61%
Abdeljaber 等的工作	89.82%	66.22%	90.87%	84.69%	80.52%	88.87%
Zhang 等的工作	99.37%	89.82%	92.70%	93.25%	77.93%	81.51%
本文方法	99.96%	98.55%	97.19%	99.89%	94.98%	97.79%

表 7　不同负载情况下的稳定性

Tab. 7　Stability under different load of each method

方法	A→B	A→C	B→A	B→C	C→A	C→B
普通 MLP	34.00% ±1.90%	33.05% ±3.85%	38.05% ±1.05%	34.60% ±3.20%	39.75% ±3.65%	34.10% ±4.10%
Ince 等的工作	96.45% ±3.05%	89.45% ±8.55%	91.35% ±4.85%	87.40% ±12.00%	83.00% ±3.70%	86.40% ±3.50%
Abdeljaber 等的工作	86.25% ±10.75%	64.60% ±13.80%	89.30% ±5.00%	80.95% ±11.95%	84.35% ±6.05%	89.10% ±4.00%
Zhang 等的工作	98.00% ±2.00%	87.85% ±10.65%	92.40% ±6.20%	91.05% ±8.95%	75.55% ±5.45%	86.65% ±9.15%
本文方法	99.90% ±0.10%	97.60% ±2.20%	96.95% ±0.85%	99.80% ±0.20%	94.35% ±2.85%	96.45% ±3.35%

通过表 6 我们可以看出，当训练集和测试集分处在不同负载环境中时，普通 MLP 模型的准确率出现了大幅度下降，此时我们可以认为这个模型不具有通用性。其它方法也都有了不同程度的下降，说明模型不能很好表达不同分布的数据，而本文方法的准确率只是出现了些许下降，说明我们的模型对不同分布的数据也能有很好的表达。

通过表 7 我们发现，其它卷积方法对于不同负载情况下的故障诊断波动性都比较大，表明它们都不具备良好的泛化能力，更多的依靠偶然性来进行诊断，即当随机打乱而获得的训练集和测试集的分布比较接近时，可能会获得较好的准确率，但是当训练集和测试集的分布差异较大时，那准确率可能就会大幅度降低。但本文的方法可以看到，是波动最小的一种方法，对于数据分布的依赖性最小，表明我们是这几种方法中最为稳定的一种方法。

3.2　全生命周期故障数据集试验

本节的试验数据来自美国辛辛那提大学智能维护中心(Center for Intelligent Maintenance Systems，IMS) 提供的滚动轴承运行的全生命周期实验数据[23]。实验台装置如图 4 所示，装置中的四个型号为 ZA-2115 的相同滚动轴承安装在主轴上，交流电机通过皮带摩擦驱

动主轴转动。数据采样频率为 20.48 kHz，每次采样的数据长度为 20 480。并将轴承转速恒定在 2 000 r/min。

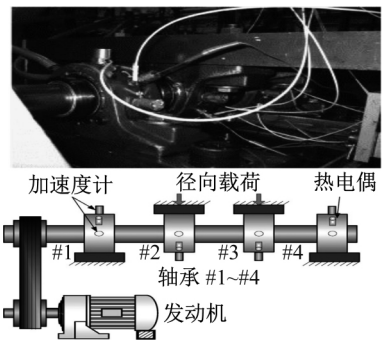


图 4 辛辛那提数据集实验台装置

Fig. 4 Cincinnati data set experimental platform

该试验轴承连续运行，采集的数据是滚动轴承从健康状态到发生故障再到直到失效全过程，时间跨度达到七天，从 2004 年 2 月 12 日 10 点 32 分开始，到 2004 年 02 月 19 日 06 时 22 分停止，每十分钟采集一次。我们取 2004 年 2 月 12 日 10 点 32 分到 2004 年 2 月 12 日 13 时 42 分轴承一的数据作为正常状态的数据，取 2004 年 2 月 19 日 03 时 42 分到 2004 年 2 月 19 日 06 时 02 分轴承一的数据作为外圈损伤状态的数据，取 2004 年 2 月 17 日 20 时 42 分到 2004 年 02 月 17 日 23 时 52 分轴承三的数据作为内圈损伤状态的数据，轴承四的数据作为滚动体损伤状态的数据。

该试验设置了一个数据集，总共包含滚动体损伤，外圈损伤，内圈损伤与正常状态共四种状态的数据集，每次采集的样本长度为 2 048。数据集具体描述如表 8 所示。

表 8 辛辛那提试验数据集描述

Tab. 8 Description of Cincinnati data set

损伤位置		外圈损伤	内圈损伤	滚动体损伤	正常状态
数据集	训练集	700	700	700	700
	验证集	200	200	200	200
	测试集	100	100	100	100

模型结构与凯斯西储大学类似，只是把最后的输出由 10 类改成了 4 类。同样也是每个方法测试 10 次，与其他方法的比较的试验结果如表 9 所示。

表 9 不同方法的结果比较

Tab. 9 Comparison of results of different methods

方法	平均准确率	稳定性
普通 MLP	79.55%	80.50% ±4.25%
Ince 等的工作	99.63%	99.63% ±0.38%
Abdeljaber 等的工作	99.21%	99.06% ±0.94%
Zhang 等的工作	99.73%	99.50% ±0.50%
本文方法	99.95%	99.75% ±0.25%

通过表 9 可知，MLP 模型准确率相对于西储大学数据集的准确率退步明显，猜想有可能是由于只有四类样本所以导致总的训练数据较少没有办法充分训练模型。为了验证该猜想我们将每类样本数量从 1 000 提高到 2 000 进行训练，训练样本，验证样本和测试样本的各自比例不变，结果如表 10 所示。

表 10 扩大数据集后的各方法结果比较

Tab. 10 Comparisons of results of various methods after enlarging data sets

方法	平均准确率	稳定性
普通 MLP	92.33%	92.19% ±2.06%
Ince 等的工作	99.71%	99.75% ±0.25%
Abdeljaber 等的工作	99.18%	98.94% ±0.69%
Zhang 等的工作	99.46%	97.38% ±2.63%
本文方法	99.80%	99.75% ±0.25%

通过表 10 可知，数据越多则越能充分训练模型，当加大数据集后 MLP 再次达到了 90% 以上的水平，同时也说明其他的卷积神经网络在样本相对较少的情况下也能达到高准确率。可以发现这几种卷积网络对辛辛那提数据集的识别率都很高且都保持了高的稳定性，基本都处在同样的高水平线上，但也注意到 Zhang 等的模型稳定性有波动，发现是 10 次试验中有 1 次试验的准确率出现了明显下降，准确率为 94.75%，其它 9 次准确率都接近或就是 100%。同时发现表 9 中 Zhang 等的模型并没有发生准确率突降的情况，表明过于复杂的模型可能有准确率突然下降的风险，同时再一次证明我们的模型相对于 Zhang 等的模型，设计的更加简单，也更易于训练。综上，本文方法无论是准确率还是稳定性表现都稍微好于其它几种卷积神经网络。

我们同时也对数据进入 Inception 前后做了可视化表达，如图 5。

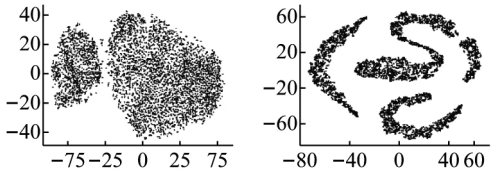


图 5 辛辛那提数据集可视化

Fig. 5 Visualization of Cincinnati data set

可以观察到，我们的模型同样在第一层卷积层区分出了故障数据和正常数据，但是不具备更加细分的能力，此时将数据送入 Inception 层后，我们可以发现数据得到了更加细致的划分，再一次证明 Inception 层的有效性和实用性。

3.3 IEEE PHM 2012 数据集试验

为了进一步验证本文模型的应用性，我们采用 IEEE PHM 2012 挑战赛[24] 的数据作为本节试验的数据集。该数据集由法国的 FEMTO-ST 研究所提供，提供了同种类型的轴承由开始到失效的全生命周期的数

据，轴承都是在一个名为 PRONOSTIA 的实验平台上进行试验，试验装置如图 6 所示，它能够模拟各种不同的工作环境以进行实验研究。FEMTO-ST 研究所分别在三种负载状态下对 17 个轴承进行试验，使得我们可以测试本文模型在不同负载情况下的表现，三种负载状态分别是：转速 1 800 r/min/负载力 4 000 N（7 个轴承是在此负载下运行），转速 1 650 r/min/负载力 4 200 N（7 个轴承是在此负载下运行）和转速 1 500 r/min/负载力 5 000 N（3 个轴承是在此负载下运行）。由于该数据集并未提供不同位置故障的数据，因此我们只能将数据集分为正常和故障两种类型，我们随机取每类负载条件下的第一个轴承的前面一段时间的振动信号作为正常数据并随机取该轴承最后一段时间的振动信号作为故障数据。

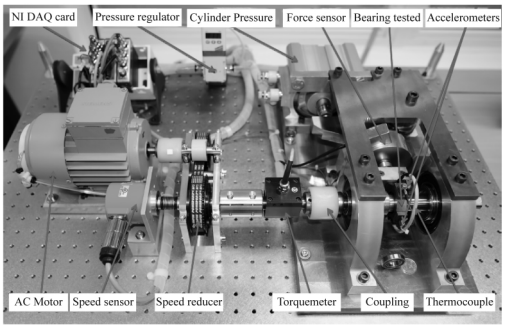


图 6　PHM2012 数据集实验台装置

Fig. 6　PHM2012 data set experimental platform

本试验设置了三个数据集，分别表示三种不同的负载状态，数据集 A 表示转速 1 800 r/min/负载力 4 000 N 的数据，数据集 B 表示转速 1 650 r/min/负载力 4 200 N 的数据，数据集 C 表示转速 1 500 r/min/负载力 5 000 N 的数据。每个数据集只包含正常和故障两种类型，每次采集的样本长度为 2 048。每个数据集同样包括训练集，验证集和测试集，数据集的具体描述如表 11 所示。

表 11　PHM 2012 数据集

Tab. 11　PHM 2012 data set

		故障	正常
数据集 A	训练集	1 400	1 400
	验证集	400	400
	测试集	200	200
数据集 B	训练集	1 400	1 400
	验证集	400	1 400
	测试集	200	200
数据集 C	训练集	1 400	1 400
	验证集	400	400
	测试集	200	200

3.3.1　同负载故障诊断结果及与其它方法的比较

模型结构与上面示例类似，由于最后的输出只有正常和故障两种类别，所以把最后的输出改成了 2 类。同样也是每个方法测试 10 次，与其他方法的比较的试验结果如表 12 和表 13 所示。

表 12　各方法平均准确率

Tab. 12　Average accuracy of each method

方法	A	B	C
普通 MLP	93.45%	95.85%	92.08%
Ince 等的工作	100.00%	100.00%	98.70%
Abdeljaber 等的工作	100.00%	99.65%	93.48%
Zhang 等的工作	100.00%	100.00%	97.83%
本文方法	100.00%	100.00%	100.00%

表 13　各方法稳定性

Tab. 13　Stability of each method

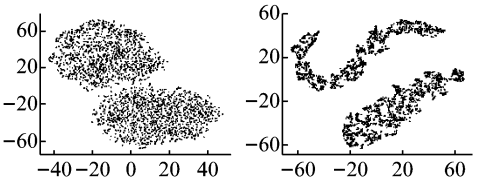
方法	A	B	C
普通 MLP	93.25% ±2.50%	95.88% ±1.87%	91.50% ±2.50%
Ince 等的工作	100.00% ±0.00%	100.00% ±0.00%	97.63% ±2.37%
Abdeljaber 等的工作	100.00% ±0.00%	98.25% ±1.75%	93.25% ±5.50%
Zhang 等的工作	100.00% ±0.00%	100.00% ±0.00%	94.38% ±5.62%
本文方法	100.00% ±0.00%	100.00% ±0.00%	100.00% ±0.00%

通过表 12 我们发现，在同负载的情况下，上述模型都能有着不错的成绩，但我们的模型在数据 A，B，C 都保持 100% 的准确率，表现稍微好于其它几种卷积神经网络。普通 MLP 与其它的方法有着明显的差异，表明仅使用普通 MLP 并不能够很好的对数据进行分类，同时发现除了我们的方法以外，其它方法对数据集 C 的诊断效果均不如数据集 A 和数据集 B，可能的原因是数据集 C 的正常数据和故障数据的分布相较于数据集 A 和数据集 B 最为接近，所以也就最难进行分类。

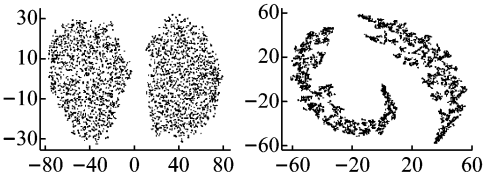
通过表 13 可以清楚的观测到在同负载环境下各模型的稳定性情况。上述模型在数据集 A，B 上都能保持不错的稳定性，但其它模型在数据集 C 上的稳定性波动较大，进一步说明了数据集 C 的正常数据和故障数据的分布相较于数据集 A 和数据集 B 最为接近，不利于模型区分。当某次试验随机选取的正常数据和故障数据的分布接近时，其它模型的分类结果就会较为准确，但是当某次试验随机选取的正常数据和故障数据分布差异较大时，可能其它模型的分类结果就会不如人意，从而导致了稳定性的波动比较大。

我们同时也对数据进入 Inception 前后做了可视化

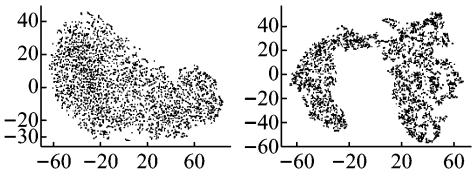
表达,如图 7。



(a) 数据集 A 进入 Inception 前后对比



(b) 数据集 B 进入 Inception 前后对比



(c) 数据集 C 进入 Inception 前后对比

图 7 PHM 数据集可视化

Fig. 7 Visualization of PHM data set

正如上述所言,模型在第一层卷积层过后就能够很好的区分出数据集 A,B 的故障数据和正常数据,但对于数据集 C 却不能很好的区分,表明数据集 C 的正常数据和故障数据分布差异相对较小。但同时,在经过 Inception 之后,我们可以发现三个数据集便都能够很好的区分出故障数据和正常数据。这也再一次体现了 Inception 的有效性和实用性,也表明我们的模型能够很好的对数据集 C 的数据进行分类。

3.3.2 不同负载环境下故障诊断结果及与其它方法的比较

由上述结果发现,在同负载情况,上述各方法都能获得不错的准确率。为此,我们试验了在不同负载环境下各方法的准确率情况。具体来说,我们分别用数据集 A,B,C 来训练样本,然后在其余的数据集上测试样本,如 A→B 代表在数据集 A 上训练样本,在数据集 B 上进行测试,同样也是每个方法试验十次。具体情况如表 14 和表 15 所示。

表 14 各方法不同负载情况下的平均准确率

Tab. 14 Average accuracy under different loads of each method

方法	A→B	A→C	B→A	B→C	C→A	C→B
普通 MLP	89.30%	63.70%	67.33%	89.65%	71.40%	87.98%
Ince 等的工作	91.15%	88.70%	57.70%	99.83%	86.60%	73.45%
Abdeljaber 等的工作	86.15%	80.40%	73.10%	99.50%	86.13%	97.70%
Zhang 等的工作	98.98%	100.00%	93.90%	100.00%	81.73%	85.55%
本文方法	100.00%	100.00%	94.63%	100.00%	98.05%	100.00%

表 15 不同负载情况下的稳定性

Tab. 15 Stability under different load of each method

方法	A→B	A→C	B→A	B→C	C→A	C→B
普通 MLP	88.88% ±5.13%	63.75% ±3.25%	67.50% ±3.25%	89.38% ±2.13%	71.88% ±2.13%	88.00% ±4.50%
Ince 等的工作	89.25% ±9.50%	80.00% ±20.00%	58.38% ±3.88%	99.50% ±0.50%	86.88% ±9.13%	75.38% ±21.88%
Abdeljaber 等的工作	85.25% ±7.50%	81.38% ±18.13%	78.00% ±18.50%	99.00% ±1.00%	86.00% ±5.75%	97.13% ±2.63%
Zhang 等的工作	97.13% ±2.88%	100.00% ±0.00%	93.50% ±6.50%	100.00% ±0.00%	79.00% ±14.25%	78.75% ±21.00%
本文方法	100.00% ±0.00%	100.00% ±0.00%	92.38% ±4.38%	100.00% ±0.00%	94.88% ±5.13%	100.00% ±0.00%

通过表 14 可知,我们的模型在所有的六种情况中都是表现最好的,表明我们的模型能够在不同负载情况下保持高准确率。同时发现普通 MLP 在所有的情况下的表现都一般,所以我们认为在不同负载环境下,我们可以放弃普通 MLP 这种模型,并且可以发现有的模型在某种情况下的准确率很低,例如 Ince 等的模型在 B→A,C→B 的情况下表现很差,表明它不能够在这种情况下对数据进行很好的表达,即其适用性受到了限制。

通过表 15 我们发现,其它方法对于不同负载下的故障诊断波动性在有的情况下比较大,表明该模型在随机打乱而获得的训练集和测试集的分布比较接近时,可能会获得较好的准确率,但是当训练集和测试集的分布差异较大时,那准确率可能就会大幅度降低。但本文的方法可以看到,是波动最小的一种方法,对于原始数据分布的依赖性最小,表明我们是这几种方法中最为稳定的一种方法。

4 结 论

本文提出了一种新的卷积神经网络用于处理滚动轴承故障诊断问题。它是基于注意力机制 Inception 网络结构搭建而成,是一种高效的滚动轴承故障诊断框架。主要结论如下:

(1) 本网络直接对原始一维故障信号进行处理,

免去了前期以往繁杂的人工提取特征的过程，极大提高了故障诊断工作的效率。

(2) 由试验验证可知，本网络不仅对同负载的数据有着极高的准确率，对不同负载的数据也有较高的准确率，且准确率表现优于最近提出的几种卷积神经网络。

(3) 同时由试验验证可知，本网络在测试集测试过程中，其准确率一直保持相对稳定的高水平状态，其稳定性表现也优于相比较的其它几种卷积神经网络。

参 考 文 献

[1] HE Q, WANG J, HU F, et al. Wayside acoustic diagnosis of defective train bearings based on signal resampling and information enhancement [J]. Journal of Sound and Vibration, 2013, 332(21): 5635-5649.

[2] WANG X, ZHENG Y, ZHAO Z, et al. Bearing fault diagnosis based on statistical locally linear embedding [J]. Sensors, 2015, 15(7): 16225-16247.

[3] LEE W, PARK C G. Double fault detection of cone-shaped redundant IMUs using wavelet transformationand EPSA [J]. Sensors, 2014, 14(2): 3428-3444.

[4] TIPPING M E, BISHOP C M. Probabilistic principal component analysis [J]. Journal of the Royal Statistical Society, 2010, 61(3): 611-622.

[5] PANDYA D H, UPADHYAY S H, HARSHA S P. Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN [J]. Expert Systems with Applications, 2013, 40(10): 4137-4145.

[6] SANTOS P, VILLA L F, REÑONES A, et al. An SVM-based solution for fault detection in wind turbines [J]. Sensors, 2015, 15(3): 5627-5648.

[7] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[8] 孙文珺，邵思羽，严如强. 基于稀疏自动编码深度神经网络的感应电动机故障诊断 [J]. 机械工程学报，2016，52(9)：65-71.
SUN Wenjun, SHAO Siyu, YAN Ruqiang. Induction motor fault diagnosis based on sparse auto-encoder deep neural network [J]. Journal of Mechanical Engineering, 2016, 52(9): 65-71.

[9] KONG Q, CUI G, YEO S S, et al. DBN wavelet transform denoising method in soybean straw composition based on near-infrared rapid detection [J]. Journal of Real-Time Image Processing, 2016, 13(3): 1-14.

[10] 雷亚国，贾峰，周昕，等. 基于深度学习理论的机械装备大数据健康监测方法 [J]. 机械工程学报，2015，51(21)：49-56.
LEI Yaguo, JIA Feng, ZHOU Xin, et al. Large data health monitoring method for mechanical equipment based on deep learning theory [J]. Journal of Mechanical Engineering, 2015, 51(21): 49-56.

[11] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C] // IEEE Conference on Computer Vision and Pattern Recognition. Boston: CVPR, 2015.

[12] CHEN L, ZHANG H, XIAO J, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning [C] // IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: CVPR, 2017.

[13] LIN M, CHEN Q, YAN S. Network in network [C] // International Conference on Learning Representations. Banff: ICLR, 2014.

[14] GHIASI G, LIN T Y, QUOC V L. DropBlock: a regularization method for convolutional networks [C] // Neural Information Processing Systems Conference. Montreal: NIPS, 2018.

[15] ZHANG W, PENG G, LI C, et al. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals [J]. Sensors, 2017, 17(2): 425.

[16] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[17] KINGMA D, BA J. ADAM: a method for stochastic optimization [C] // International Conference on Learning Representations. San Diego: ICLR, 2015.

[18] LOU X, LOPARO K A. Bearing fault diagnosis based on wavelet transform and fuzzy inference [J]. Mechanical Systems and Signal Processing, 2004, 18(5): 1077-1095.

[19] KESKAR N S, MUDIGERE D, NOCEDAL J, et al. On large-batch training for deep learning: generalization gap and sharp minima [C] // International Conference on Learning Representations. Toulon: ICLR, 2017.

[20] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks [J]. Journal of Machine Learning Research, 2010(9): 249-256.

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

[22] ABDELJABER O, AVCI O, KIRANYAZ S, et al. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks [J]. Journal of Sound & Vibration, 2017, 388: 154-170.

[23] QIU H, LEE J, LIN J, et al. Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics [J]. Journal of Sound & Vibration, 2006, 289(4): 1066-1090.

[24] NECTOUX P, GOURIVEAU R, MEDJAHER K, et al. PRONOSTIA: an experimental platform for bearings accelerated degradation tests [C] //IEEE International Conference on Prognostics and Health Management. Denver, 2012.