数据增强的深度卷积网络不变性机理的可解释性研究

刘妮[1] 雷娇娇[1] 邢振蓉[1] 韩文静[1] 何立火[2]
（1.长安大学信息工程学院，710064，西安；2.西安电子科技大学电子工程学院，710071，西安）

摘要：在图像分类任务中，数据增强是提升深度卷积网络（ConvNets）分类不变性的通用手段。针对当前深度卷积网络不变性研究中主要聚焦网络层的整体不变性度量，而忽视了位置、方向等信息的表征机制，且缺乏对特征综合机理的系统阐释的问题，本文受认知神经科学领域相关研究的启发，采用圆形/方形二分类任务构建简化实验场景，有效规避复杂特征的干扰。采用最大激活方法，对全连接层与全局平均池化（AvgPool）层进行特征可视化分析，研究了经数据增强训练后的全连接层与AvgPool层的不变性学习特性，并采用t分布随机邻域嵌入（t-SNE）降维可视化揭示了层级间高维权重向量的分类决策机理。结果表明：（1）全连接层首层通过逐步扩大位置感知域的方式实现了平移不变性；（2）全连接层通过线性加权机制有效整合分布式局部特征，遵循同类别特性正向激发、异类别特征负向抑制的连接逻辑；（3）ResNet架构中AvgPool层展现更优的平移不变性，而全连接结构则表现出更强的旋转和缩放不变性。

关键词：深度卷积网络；数据增强；平移不变性；特征可视化
中图分类号：TP18 **文献标志码：**AA
文章编号：0253－987X(XXXX)XX－0001－19

Interpreting the Invariance Mechanism of Data Augmentation on Deep Convolutional Networks

LIU Ni[1],LEI Jiao Jiao[1],XING ZHEN RONG[1],HAN WEN JING[1],HE LI HUO[2]
(1. School of Information Engineering, Chang'an University, Xi'an 710064, China;
2. School of Electronic Engineering, Xidian University, Xi'an 710071, China)

Abstract: In image classification tasks, data augmentation is a common means to improve the classification invariance of deep convolutional networks (ConvNets). To address the problems that current research on Convolutional Neural Network invariance primarily focuses on global metrics while neglecting the representation mechanisms of position and direction information, and lacks a systematic explanation of feature integration mechanisms, this study, inspired by relevant research in cognitive science and neuroscience, constructed a simplified experimental scenario using a circular/square binary classification task to effectively eliminate interference from complex features. By employing the maximum activation method, we visualized and analyzed the feature representation and invariance

收稿日期：2025-10-13。

learning characteristics of the fully connected (FC) layers and Average Pooling (AvgPool) layers in models trained with data augmentation. Additionally, t-SNE dimensionality reduction visualization was used to reveal the classification decision-making mechanisms of high-dimensional weight vectors between layers. The results demonstrate that: (1) the initial fully connected layer achieves translation invariance by gradually expanding the position-aware domain; (2) the fully connected layer effectively integrates distributed local features through a linear weighting mechanism, following a connectivity logic of positive excitation for intra-class features and negative inhibition for inter-class features; and (3) the AvgPool layer in the ResNet architecture exhibits superior translation invariance, while the cascaded fully connected structure demonstrates stronger rotation and scaling invariance.

Keywords: Deep Convolutional Neural Networks; data augmentation; translation invariance; feature visualization

在图像分类任务中，只有当输入图像经过平移、旋转、缩放等常见的几何变化后提取的特征依旧保持不变，才能保证复杂场景下物体分类识别的有效性。这是图像识别的理想情况[1]。

深度卷积网络（ConvNets）受生物模型启发引入了卷积与池化运算[2]。过去的观点普遍认为卷积网络通过卷积和池化使其在结构上具有平移不变性[3]。但然而，由于卷积中的步长未能考虑采样定理以及全连接层不具有空间推理能力等原因，ConvNets 在结构上并不具备完全的平移不变性[4]。虽然已有研究尝试构造专门的不变性网络结构[3,5]，但这些方法往往针对特定应用[6]，要么只是对部分旋转角度起作用，或者需要增加额外的模型参数来处理不变性，甚至有可能导致模型准确度下降[7]。

更通用且有效的方法是通过数据增强使 ConvNets学习不变性特征[8]。这种技术仅需在训练阶段对数据集图像进行平移、旋转、缩放等操作，生成更多样本，然后 ConvNets 就能够在训练中自动地学习到不变性特征。深入理解经此方式训练后的网络关于位置、方向、尺度等信息的表征方式和决策机理，具有重要的理论与实践价值。这不仅有助于为不变性模型的设计提供思路，更能揭示数据增强的潜在负面影响。例如，斯坦福大学的Hermann等发现，随机裁剪和平移增强是导致深度神经网络无法利用全局形状信息进行识别的主因[9]。

目前，针对数据增强与不变性的分析大多数是利用平移-敏感图对比网络层的整体不变性程度[4,10-11]，而忽视了位置、方向等信息的表征机制，且缺乏对特征综合机理的系统阐释。本文针对实际应用更广泛的AlexNet和ResNet的两种相关结构，全连接层（Fc）和全局平均池化层（AvgPool），从特征可视化（feature visualization）和特征综合（feature integration）两个方面研究它们是如何学习到关于位置、方向和尺度不变的特征，以及如何实现分类的决策机理。

为了规避图像本身复杂特征带来的

不必要干扰，使深入分析数据增强的不变性机理变得可能，本研究借鉴认知神经科学领域的实验范式[12-13]，构建了圆形/方形二分类简化场景，设计了一个包含两类几何形状的圆方数据集。本文主要在圆方数据集上进行了增强训练条件下网络的特征可视化分析，并在更加复杂一些的真实数据集MNIST数据集上进行了验证。本文的主要工作和贡献如下：

1）基于梯度下降生成神经元最大激活图的特征可视化方法是解释神经网络的有效手段[14-16]。本文利用该方法，对经过平移、缩放和旋转增强训练的AlexNet和ResNet的关键结构进行可视化。实验发现AlexNet全连接第一层（Fc1）提取的特征就与位置、方向和尺度无关了。通过对训练过程中神经元的可视化，展现了网络是如何在训练中逐渐扩大位置感知域的。

2）除了对全连接层单个神经元的特征可视化，还对把分布在各个神经元上的特征进行线性加权的特征综合机制进行了解释。线性模型虽然简单，但可解释性未必强。当成千上万的特征都对某次决策有贡献，用户很难直观理解决策的原因[17]。本文对两个全连接层中间的高维权重向量进行了*t*-SNE可视化[18]，直观展示了各神经元对图形分类的决策机理，并在特征较复杂的数据集MNIST[19]上进行了实验验证。

3）以ResNet为代表的网络采用全局平均池化层替换了AlexNet的全连接层。由于是对滤波响应图进行平均值的计算，相比于全连接层，AvgPool层表现出更强的平移不变性。但是，全连接层的两层级联结构的旋转和缩放不变性比AvgPool层强。

1　相关工作

1.1　深度卷积网络的变换不变性

Kayhan等[20]对不同网络层进行了实验，发现卷积层对物体的绝对位置进行编码，其对位置敏感。Myburgh[10]等人使用余弦相似度计算敏感度，发现全连接层在实现平移不变性起主要作用。Han[21]等发现经过ImageNet预训练的五层网络体现出了一定的平移不变性。顾等[6]使用卷积网络实现合成孔径雷达自动目标识别，采用了平移敏感图来展示模型的平移不变性。Laptev等[5]采用结构化数据增强方法，将旋转图像输入卷积网络，通过共享权重与全局最大池化提取关键特征以增强网络旋转不变性。

1.2　深度卷积网络的可解释性

可视化是解释深度神经网络内部机制的让人感兴趣且有效的方法。直接可视化滤波器权重值（visualizing filters）或响应图（activations）的方法，其结果往往只有第一层是可以解释的，高层特征由于变的更加抽象，即使可视化出来也很难理解，导致该类方法失去效果。

显著性图（saliency map）能够展示出对模型的预测贡献最大的图像像

素。Zeiler等人[22]通过反卷积网络将特征激活映射回输入像素空间，以揭示哪些输入模式能够引发特定的输出响应。类激活映射[23]通过使用全局平均池化和线性层以确定各个特征以及特征对分类结果的影响。Zeiler等人[22]使用灰色斑块遮挡图像，观察模型对目标类别的预测概率的变化来估计输入像素的重要性。

认知神经科学上理解大脑的一种途径是研究能够激活动物单个神经元细胞的输入刺激图像，也被称为首选刺激[16]。同样，机器学习中，可视化神经元的首选刺激可助于了解神经元所学特征。一种方式是从大型数据集中寻找使神经元响应最大的图像[15,22]。另外一种方式是生成视觉刺激图像[14]。这类方法通过梯度下降迭代优化输入，使目标神经元的激活值最大，从而揭示其学习的特征。这两种激活最大化方法（activation maximization）分别被称为以数据为中心和以网络为中心的方法[24]，后者生成图像的逼真性依赖于具体的优化算法，有时候不如第一种方法直观。该类方法在很多文献中也被称为特征可视化[16]。

除了可视化方法，还有文献试图从更加全局的角度研究模型整体的特征路径、决策逻辑等。局部可解释的模型无关解释(LIME）[17]通过局部扰动输入数据来训练线性模型，以分析特征对预测的贡献，形成单条样本预测结果的解释。Zhang[25]等通过决策树在语义层面上明确网络每一次预测的具体原因。宋明黎等通过可视化不同类别在网络中的贡献程度，从一定程度上揭示了模型处理不同类别时的决策路径[26]。

2 特征可视化算法描述

2.1 以网络为中心的特征可视化方法

该方法的核心原理是通过优化输入数据，使得网络中的某个特定神经元或滤波器的激活值达到最大。考虑一幅待优化图像$x\in\mathrm{R}^{C\times H\times W}$，其中$C$表示颜色通道数，$H$和$W$表示图像的高度和宽度。对于一个训练好的神经网络，假设其参数θ(包括权重和偏差) 是固定的；$a_{ij}(\theta, x)$表示网络第j层指定神经单元i的激活。此方法的目标是找到一个最优输入x^*，使得$a_{ij}(\theta,x)$达到最大值[14]，其优化目标函数如公式公式（1）所示：

$$x^* = \arg\max_x \ a_{ij}(\theta,x) \tag{1}$$

式（1）是一个非凸优化问题，采用梯度上升的方法找到一个最优解x^*。在可视化过程中，为了使生成的图像看起来更自然，通常会使用正则化技术对x进行约束，避免产生一些抽象的“痕纹”使其碰巧有比较高的响应值[14,27-28]。优化目标如公式（2）所示：

$$x^* = \arg\max_x \left(a_{ij}(x) - \mathrm{R}_\theta(x)\right) \tag{2}$$

式中：$R_\theta(x)$表示正则化函数。在具体的实现过程中，通过$r_\theta(\cdot)$算子定义正则化，该算子将x映射到自身的正则化版

本。本文采用了[15]中一种实施起来比较方便的梯度下降框架，该方法轮流进行$a_{ij}(x)$的梯度计算和r_θ函数的计算。用η表示梯度下降的学习率，每一步的更新公式如下所示：

$$x \leftarrow r_\theta(x+\eta\frac{\partial a_{ij}}{\partial x}) \tag{3}$$

r_θ包含L_2衰减与高斯模糊。采用L_2衰减阻止极端大的像素值的产生，并在迭代过程中周期性地对输入图像施加高斯模糊，以抑制梯度下降引入的高频噪声。在可视化阶段，将优化得到的图像反标准化回像素空间，并将像素值裁剪到[0,255]范围内输出。正则化参数通过随机超参数搜索确定[15]，以可视化结果的主观最优效果为选择依据。实验发现，增大高斯模糊强度可抑制可视化结果中的高频噪声，但会同时削弱边缘与细线结构，使结果更为模糊。本文设置学习率η=0.01，迭代次数T=200。L_2衰减系数θ_{decay}=0.1，高斯模糊半径θ_{b_width}=0.1，每隔θ_{b_every}=100次迭代施加一次高斯模糊。

2.2 以数据为中心的特征可视化方法

以网络为中心的特征可视化可以生成边缘、纹理等特征，有助于理解网络特征提取过程。然而该方法计算复杂度较高，并且某些情况下容易受到高频信息影响导致可视化效果不佳[29]。为了克服这些局限性，本研究同时采用了以数据为中心的可视化方法。该方法在数据集中寻找能够引起神经元产生最高激活的输入样本[22]。本实验中，对于每个神经元，寻找9幅使其响应最高的图像，称为top-9集。同时采用这两种方法对网络进行可视化分析，有助于结合两者的优点得到更加全面的结论。

3　平移增强网络的特征可视化

AlexNet、VGG、ResNet等雏形网络不仅在实际应用中仍然有很大的需求，它们的经典架构如卷积层、池化层、全连接层等，也在很多新的网络中被使用到了。本文针对AlexNet和ResNet18这两种卷积网络，首先采用特征可视化的方法对经过增强训练的网络关键层的神经元进行可视化分析，并在此基础上对特征的综合机理进行了探究。

3.1 实验数据集和参数设置

深度卷积网络通常都有比较复杂的模型结构和大规模的参数量，其输入和输出之间的映射关系往往也非常复杂，了解其特征表示和决策过程是非常有挑战性的。本文主要针对平移、缩放和旋转信息的可视化。为了便于研究，避免物体本身所含复杂特征带来的不必要干扰，需要一个物体特征比较简单的数据集。受文献[13]的启发，本文设计了一个只包含两类标志性几何形状——圆形和方形的平移增强数据集（简称为圆方数据集）。除了该数据集，还在MINST这个真实世界数据集上进行了实验验证。下面是对这两个数据集的详细介绍。

圆方数据集：每张图像包含一个圆形或正方形，其中圆的直径范围为[10,48]，正方形的边长范围为[10,56]。画布大小为224×224，目标可随机出现在画布的任意位置。该数据集一共包含8000张圆形和8000张方形。训练集、验证集和测试集的比例为7:2:1。

T-MINST数据集：将大小为28×28像素的MNIST图像放大至原来的4倍后嵌入至224×224像素的黑色背景画布中。每张MNIST图像可随机出现在画布的任意位置。该数据集一共包含70000张图像，训练集、验证集和测试集的比例为7:2:1。

为避免预训练模型参数对实验结果的影响，采用从头开始学习的方式对网络进行训练。网络参数采用PyTorch默认初始化：卷积层与全连接层权重采用Kaiming(He)初始化，偏置项为默认初始化。实验前将输入图像统一调整为224×224并进行归一化处理。模型训练采用Adam优化器和交叉熵损失函数[30]。通过实验对比了多个初始学习率，以验证集性能最优值作为训练参数。最终参数设置学习率为0.001，批次大小为64。为防止过拟合，引入基于验证集误差的早停策略，当验证集误差连续6个迭代周期上升时，停止训练并保留性能最优的模型[13]。

3.2 平移–敏感图实验

文献[11]利用平移-敏感图（Translation-Sensitivity Maps，TSM）量化网络的平移不变性，发现相比于卷积层，全连接层具有更强的平移不变性。本文进一步利用平移-敏感图对全连接层与AvgPool层两种结构的平移不变性进行比较。

平移-敏感图通过比较原始图像的网络响应（基准向量）与平移后图像的网络响应（平移向量）的相似度来计算。通常采用余弦相似度[11]，其值越大表示向量越相似，对应表明网络对平移不敏感。平移-敏感图S_T的计算公式为：

$$S_T(x,y)=\cos_{sim}(V,V'(x,y)) \tag{4}$$

式中：$S_T(x,y)$表示敏感图中在位置(x,y)处的像素；V表示基准图像响应；$V'(x,y)$表示测试图像在平移位置(x,y)处的响应；$cos_{sim}(\cdot)$表示余弦相似度。通过计算敏感图以图像中心为原点，在不同半径上的均值，可得到径向敏感度图[10]。

首先在圆方数据集上对AlexNet和ResNet18进行训练，并分别计算其网络层的平移-敏感图和径向敏感度图。对每一个类别，都生成了5张基准图像，大小范围正方形边长从10个像素到28个像素，圆形半径从5个像素到14个像素。对每一个基准图像，计算其与441个不同平移位置的测试图像的相似度。平移-敏感图大小为21×21。对得到的平移-敏感结果计算均值得到最终的平移-敏感图。

为了进行对比，本文还设计了第二种平移增强训练的方案，训练物体被限定在较小的图像区域内。以图像中心为原点，x方向和y方向的随机平移量均

限定在-35~35像素范围内。这两种训练方案分别称为：平移增强范围I和平移增强范围II。前者的训练物体覆盖区域大，后者的训练物体覆盖区域小。

图1分别展示了两种增强训练条件下AlexNet的最后一层卷积（CL5）、全连接第一层（Fc1）、第二层（Fc2），以及ResNet18的AvgPool层的平移-敏感图和径向敏感度图。平移-敏感图（图1第1、3行）的灰度从黑到白对应的余弦相似度从0到1，颜色越白对应相似度越高。径向敏感度图（图1第2、4行）中的横轴代表测试对象相对于基准对象的径向距离，以像素数为单位。纵轴则显示了在相应径向距离上余弦相似度的平均值。

无论哪种情况，CL5层的平移敏感图只在图像中心附近比较高（亮度高，相似度值接近1）。这表明经过平移的目标与未经过平移的目标，在CL5层的输出差异较大。其对应的径向敏感度图也表明位移越大，滤波器响应相似度越低。这说明了AlexNet的CL5层不具有平移不变性。

如图1中红色虚线框所示，增强范围I时，Fc2层在不同平移距离下的相似度变化都不大；增强范围II时，其相似度只在平移量为40像素以内较高，超过该值，相似度急剧下降。这说明了AlexNet的全连接层只在经过训练的区域具有平移不变性，且全连接层是使网络具有平移不变性的关键结构。

ResNet采用了全局平均池化层，因此无论是否进行平移增强训练，应该都具有一定的平移不变性。这个结论从图1中得到了验证，可以看出，在两种增强范围下，AvgPool层都有接近于1的相似度值。

尤其是在增强范围II时，如图中的蓝色虚线框对比所示，AvgPool层在处理位置变化时展现出比AlexNet更优的平移不变性，不管是否进行了平移增强训练。这主要是由于其计算的是二维特征图的平均值，从而对位置变化具有较强的鲁棒性。

3.3 特征可视化实验结果与分析

AlexNet的Fc1和Fc2层各包含4096个神经元，采用第2节介绍的方法对全部神经元进行可视化。如图2所示，为了便于对比，同一神经元的两种可视化结果被放在一起，分别位于上下两行。

对于经过圆方数据集训练的Fc1和Fc2层，可以看出，以网络为中心的可视化结果包含两类特征：一类呈现出明显的竖直和水平线条（图2（a）第1行第1、3、5幅图）；另一类则没有明显的横竖线条，而是呈现出一些随机分布的弧线或角（corner）（图2（a）第1行的第2、4幅图）。从它们对应的以数据为中心的可视化结果可知，这两类特征分别对应方形和圆形的局部特征。Fc1层的激活图均匀的分布在图像的各个位置，这说明从Fc1层开始，特征已经与位置无关了。

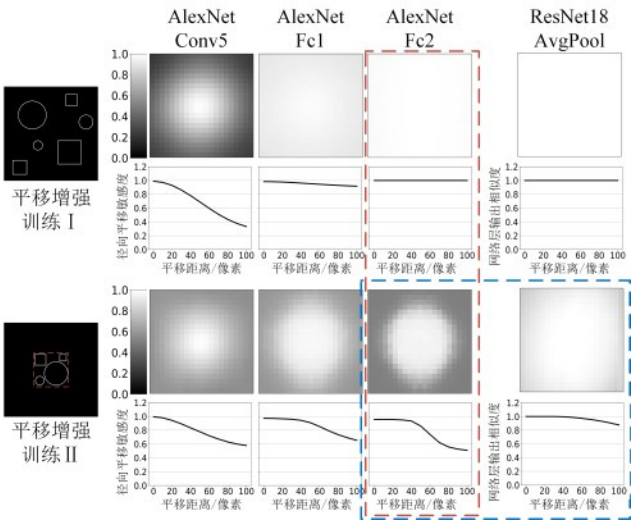


图 1 两种训练条件下，AlexNet的CL5、Fc1、Fc2层和ResNet的AvgPool层的平移-敏感图和径向敏感度图

Fig. 1 Under the two training conditions， the translation-sensitivity map and radial sensitivity map of the CL5， Fc1， Fc2 layers of AlexNet and the AvgPool layer of ResNet.

图 2 （c） 显 示 了 ResNet18 的 AvgPool层的可视化结果。图中呈现了类似AlexNet的结果，包含横竖线条或角这两类特征，并且均匀的分布在整个图像上，也就是说该层的神经元表征的特征与位置无关。

为了了解Fc1层神经元在训练过程中是如何学习关于位置的信息，本文对训练过程中不同batch时AlexNet的Fc1层神经元进行了可视化，结果如图3所示。可以观察到Fc1层的神经元对位置的敏感区域随着训练的进行逐渐扩大，直到完全覆盖到画布的所有位置。根据这种实验现象，推测是因为训练中不断地遇到新的位置的物体，加强了这种映射。总之，可以看到神经元是在训练中逐渐扩大对位置信息的感知的。

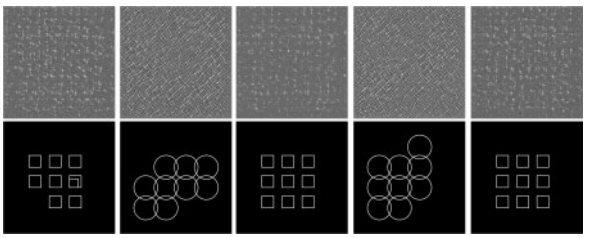
T-MNIST实验结果如图4所示。尽管数据集更复杂使得直接区分神经元学习的特征变得较为困难，但还是能看出Fc1和Fc2层神经元学习到的某些特征，表现出重复出现的局部特征互相重叠在一起。

以上实验表明Fc1层与Fc2层实现了完全的平移不变性，而ResNet的AvgPool层天然具有的对位置不敏感的特点。

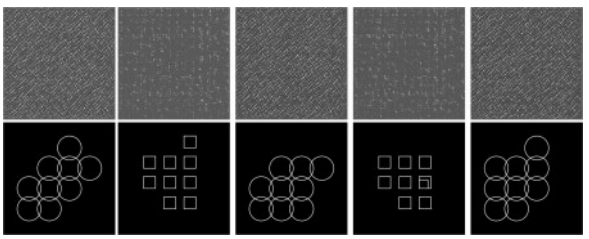
3.4 利用 *t*–SNE 探究全连接层特征综合机理

从上面的可视化结果可以看出来，AlexNet网络Fc1层已经可以形成与位

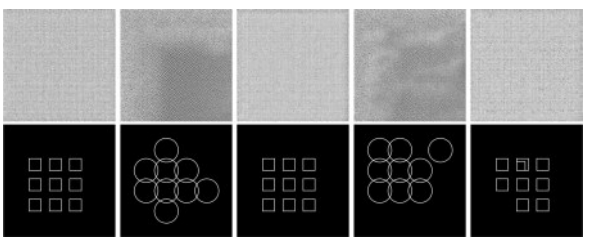
置无关的特征，接下来进一步分析Fc2层是如何综合这些信息并输出到softmax层进行分类的机理。主要通过对Fc1和Fc2层之间的连接权重矩阵$W_{FC}\in R^{4096\times4096}$，进行分析。



（a）　AlexNet网络Fc1层特征可视化



（b）　AlexNet网络Fc2层特征可视化



（c）　ResNet18网络AvgPool层特征可视化

图 2　平移增强训练I条件下，AlexNet网络Fc1、Fc2和ResNet18的AvgPool层在圆方数据集上以网络为中心的可视化和以数据为中心的可视化结果。

Fig. 2　Under the condition of translation-enhanced trainingI， the AvgPool layers of AlexNet networks Fc1， Fc2 and ResNet18 are network-centered visualization and data-centered visualization results on the round-square dataset.

以网络为中心的可视化结果中，特征均匀的分布在整个图像上，对应的以数据为中心的可视化结果中，圆或方形分布在图像的各个位置，均说明对应网络层的神经元可以捕获各个位置的特征，即具有平移不变性。

首先，对W_{Fc}矩阵中的所有数据进行一维直方图统计，发现其正负值数量比近似于1:1，且以y轴呈对称分布，如图5（a）所示。然后采用*t*-SNE降维技术对W_{FC}进行可视化分析，并对结果进行Kmeans聚类。如图5（b）所示，结果明显的可以划分成三部分：中间黄色对应的神经元在实际样本中的滤波器

响应值普遍很低，接近0。经过实验验证，剔除这些神经元后，分类准确率下降幅度低于0.001，说明这些神经元贡献有限。通过以数据为中心的可视化方法，发现红色神经元主要表征的是方形特征，蓝色神经元则表征的是圆形特征。

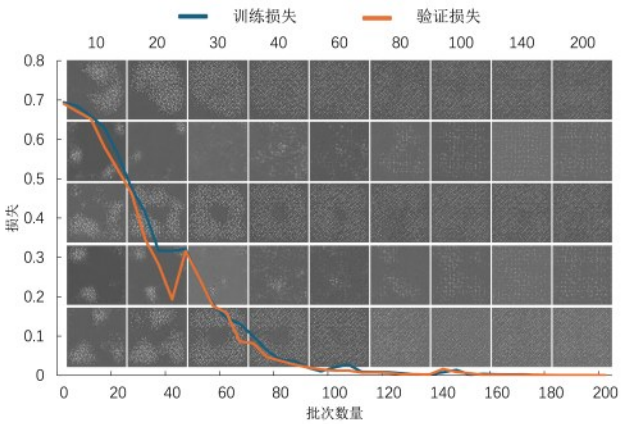


图3 平移增强训练I条件下，AlexNet的Fc1层神经元在训练过程中不同批次时的可视化结果。

Fig. 3 Under the condition of translation enhancement training I， the visualization results of AlexNet 's Fc1 layer neurons in different batches during the training process.

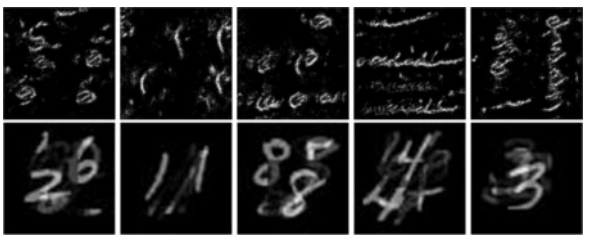
横轴上方的数字表示神经元特征可视化图像对应的批次数量，下方表示网络训练的批次数量。每一行展示了同一神经元在不同训练批次的可视化演变过程。可以观察到，神经元对位置的敏感区域随着训练的进行逐渐扩大，直到完全覆盖到画布的所有位置。

3.3节可以看出经过训练的AlexNet只需提取线段和角特征就可以实现对圆方数据集的二分类。该数据集特征的简单性使得更进一步了分析Fc1与Fc2层之间的神经元的连接路径及其物理意义成为可能。对Fc2层表示同一类别特征的所有神经元的权重向量进行平均，得到的两个平均权重向量分别记为$W_{squre}\in R^{1\times4096}$和$W_{circle}\in R^{1\times4096}$，对应图5（c）中的实线和虚线，对Fc1层的4096个神经元进行线性加权求和。进一步统计分析，可以发现W_{squre}与Fc1层的方形神经元的连接权重88%都是正值，与圆形神经元的连接权重99%都是负值。与之相反，W_{circle}与Fc1的方形神经元的连接权重值99%都是负值，与圆形神经元的连接权重值91%都是正值，其示意图如图5（c）所示。

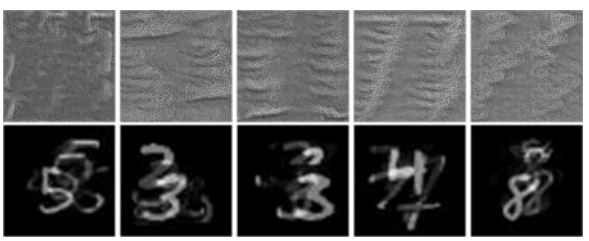
也就是说，若Fc1层神经元与Fc2层神经元描述的是同一类别的特征，则彼此之间的连接权重倾向于正连接；若不是，则倾向于负连接。当输入图像是正方形时，通过AlexNet卷积层的层层特征提取，到达Fc1层时，一部分神经元产生正激活，然后通过一些正连接传递到Fc2层的某些神经元，这些响应最

后通过softmax层实现了分类。

（a）　AlexNet网络Fc1层特征可视化



（b）　AlexNet网络Fc2层特征可视化



（c）　ResNet18网络AvgPool层特征可视化

图 4　平移增强训练I条件下，AlexNet网络Fc1、Fc2和ResNet18的AvgPool层在MNIST数据集上以网络为中心的可视化和以数据为中心的可视化结果

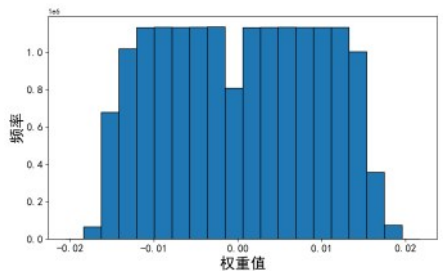
Fig. 4　Under the condition of translation enhancement training I， the AvgPool layer of AlexNet network Fc1， Fc2 and ResNet18 is network-centric visualization and data-centric visualization results on MNIST dataset.

以网络为中心的可视化结果中，特征均匀的分布在整个图像上，对应的以数据为中心的可视化结果中，目标分布在图像的各个位置，均说明对应网络层的神经元可以捕获各个位置的特征，即具有平移不变性。

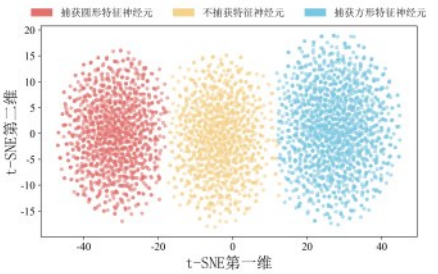
为了在更真实数据集上进行实验分析，选取了MNIST数据集1、2、3、4四个类别，生成数据集对网络进行平移增强训练。首先对W_{FC}进行t-SNE降维和Kmeans聚类分析。利用轮廓系数法[31]确定最佳聚类数目为9，如图6所示。对每个簇内簇心对应的神经元进行以数据为中心的可视化。结果显示，每个簇内的神经元主要对应一种或多种数字类别。可以发现，包含多个数字类别的聚类簇中的数字通常在形状上具有一定的局部相似性，例如数字“2”和数

字“3”；数字“1”和数字“4”等。这表明该神经元学习了这些数字类别的共有局部特征。

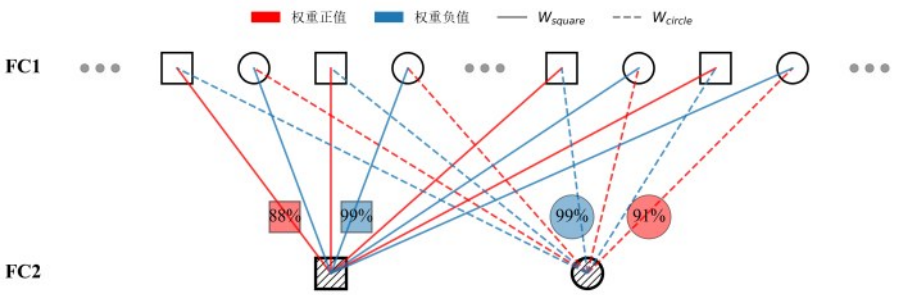
可以推测，随着数据集特征变得越加复杂，这种聚类将会越来越不明显，对神经元连接路径和特征综合机理的分析也会变得越加困难。



（a） Fc1和Fc2层连接权重 W_{FC} 权重分布统计



（b） W_{FC} 的 t-SNE降维可视化



（c） AlexNet Fc1层与Fc2层神经元连接路径示意图。图中的圆形和方形代表可视化特征分别为圆和方的神经元。

图 5 圆方数据集上平移增强训练I条件下，AlexNet 全连接层的特征综合机理分析。

Fig. 5 Analysis of feature synthesis mechanism of AlexNet fully connected layer under the condition of translation enhancement training I on square data set

4 旋转与缩放增强网络的特征可视化

4.1 实验数据集

R-MNIST：将 MNIST 图像随机旋转-180°~180°后调整至 224×224 像素大小。该数据集一共包含 70000 张图像，训练集、验证集和测试集的比例为 7:2:1。

S-MNIST：参考文献[32]的设计，将 MNIST 图像嵌入 224×224 像素的黑色背景画布中，每张 MNIST 图像随机缩放后置于画布中央，缩放倍数为 1~16。该数据集一共包含 70000 张图像，训练

集、验证集和测试集的比例为7:2:1。

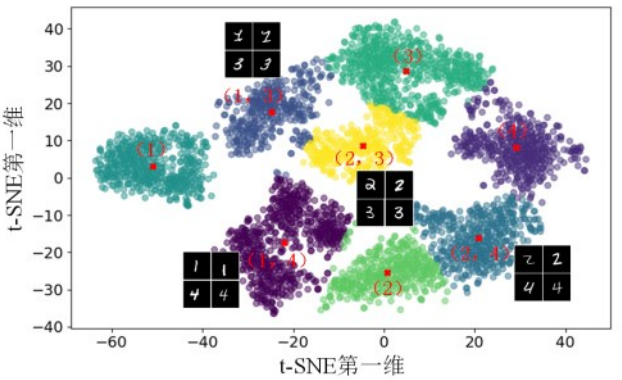


图 6　MNIST部分数据集增强训练AlexNet的W_{FC}的t-SNE可视化，以及各聚类中心神经元的以数据为中心的可视化结果对比。

Fig. 6　The t-SNE visualization of AlexNet 's WFC is enhanced by some MNIST datasets， and the data-centric visualization results of each cluster center neuron are compared.

每个聚类簇内簇心的神经元的可视化结果均包含一种或多种数字，且包含多个数字类别的聚类簇中的数字通常在形状上具有一定的局部相似性，例如数字“2”和“3”有共同的局部特征。

4.2 旋转与缩放–敏感图实验

在R-MNIST与S-MNIST上使用与3.2节类似的方法对AlexNet和ResNet18网络分别进行旋转-敏感图与缩放-敏感图实验。旋转-敏感图实验中，基准图像为未进行旋转的图像，测试图像通过对目标进行步长为5°，在-180°~180°范围内均匀旋转得到。缩放-敏感图实验中，基准图像为未进行缩放的图像，测试图像通过对目标进行步长为0.1，在1~16范围内进行均匀缩放得到。

与平移增强类似，为了便于对比，设计了部分数据增强方案。其中，部分旋转增强将旋转角度限定在-30°~30°范围内。部分尺度增强将缩放倍数限定在2~6范围内。

对数据增强和部分数据增强训练下的AlexNet网络的CL5、Fc1和Fc2层和ResNet18网络的AvgPool层绘制旋转-敏感图与缩放-敏感图，如图7、8所示。实验结果表明，经过特定的增强，从全连接层开始逐渐展现出相应的数据增强的不变性，Fc2层的不变性比Fc1层更强。在旋转与缩放变换上，AvgPool层并没有优势。无论是旋转还是缩放变换，AlexNet和ResNet18都仅仅在经过相应的数据增强的范围内展现出不变性。例如经过旋转增强的网络在各个旋转角度都有较高的不变性，而经过部分旋转增强的网络仅在学习过的旋

转范围–30°~30° 内展现出不变性。缩放变换同样仅在经过数据增强的缩放倍数内展现出不变性。

ResNet18的AvgPool层的旋转和缩放不变性的波动相比于Fc1和Fc2层要大，这一点可以在平移敏感试验中发现类似现象。当平移、旋转或缩放的程度比较大的时候，不变性就会有一定的减弱。而全连接层随着网络层级的加深，对这种变化的鲁棒性较强。

注：每个聚类簇内簇心的神经元的可视化结果均包含一种或多种数字，且包含多个数字类别的聚类簇中的数字通常在形状上具有一定的局部相似性，例如数字“2”和“3”有共同的局部特征。

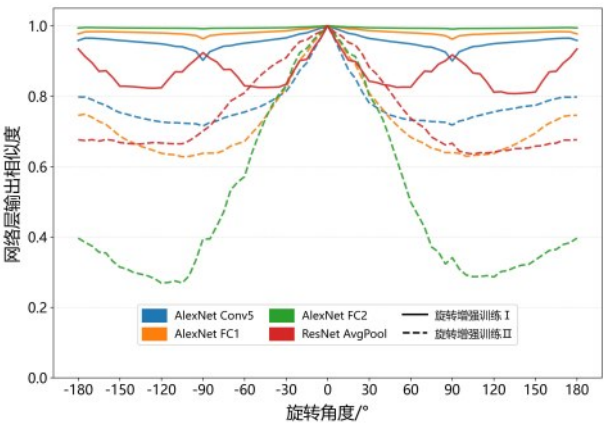


图 7 旋转增强训练I、II条件下，AlexNet网络CL5、Fc1、Fc2层和ResNet18网络AvgPool层的旋转-敏感图。

Fig. 7 The rotation-sensitive maps of CL5, Fc1, Fc2 layers of AlexNet network and AvgPool layer of ResNet18 network under the condition of rotation enhancement training I and II.

纵坐标表示旋转特定角度后的图像与未旋转图像在对应网络层输出的相似度，相似度越高，表明该网络层的旋转不变性越强。

4.3 特征可视化实验结果与分析

采用第2节的方法对AlexNet的Fc2层全部神经元进行可视化，图9（a）展示了旋转数据增强实验结果。为了便于对比，同一神经元的两种方法的可视化结果被放在一起，分别位于上下两行。以数据为中心的可视化的Top-9集的9张图像被放在一起。

结合两种可视化结果可发现，各神经元能够捕捉具有不同旋转角度的局部特征，例如，图9（a）的第3组结果中，以网络为中心的图像显示出该神经元能够提取类似字母“u”的局部结构特征，且能学习该特征的不同旋转角度的变体。而数字“2”和“4”中均包含

这种局部特征，该神经元以数据为中心的可视化中包含不同旋转角度的数字“2”和“4”，对应不同旋转角度的“u”形特征。

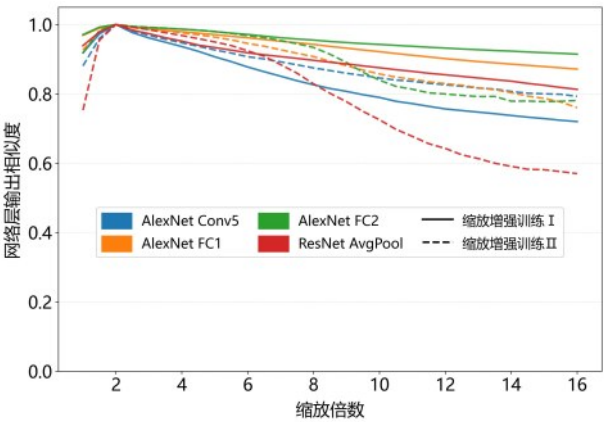


图 8　缩放增强训练I、II条件下，AlexNet网络CL5、Fc1、Fc2层和ResNet18网络AvgPool层的尺度-敏感图。

Fig. 8　The scale-sensitive maps of CL5， Fc1， Fc2 layers of AlexNet network and AvgPool layer of ResNet18 network under the condition of scaling enhancement training I and II.

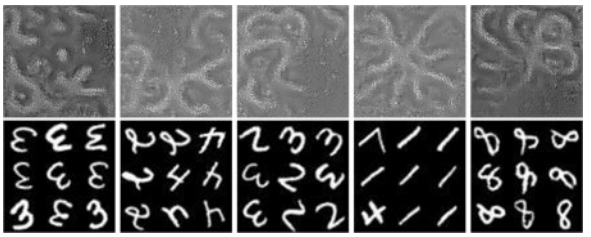
纵坐标表示缩放后的图像与未缩放图像在对应网络层输出的相似度，相似度越高，表明该网络层的缩放不变性越强。

同样对ResNet18网络的AvgPool层的全部512个神经元进行可视化，一部分实验结果如图9（b）所示。与AlexNet Fc2层的类似，可以观察到每个神经元捕捉不同局部特征的旋转变体。

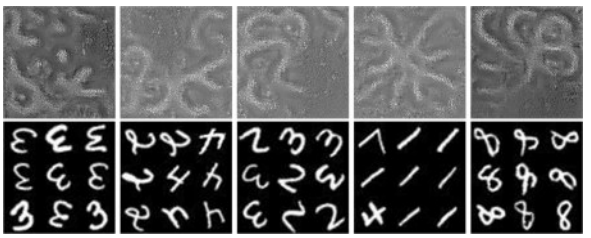
相比之下，在本文设计的缩放数据集上，以网络为中心的可视化方法，会对各个不同大小的数字特征混合重叠在一起，形成一团光斑，难以对应到清晰的几何形状或语义模式，也就失去了可视化的意义。因此只展示了以数据为中心的Fc2层与AvgPool层的可视化结果，如图10所示。可以看到，同一神经元对应的高响应样本往往包含不同缩放倍数、相同或不同类别的数字。这表明，在引入尺度数据增强后，网络深层神经元可形成对尺度不变的特征。

5　结论与讨论

本文通过不变性敏感图分析与最大激活特征可视化方法，研究了数据增强下深度卷积网络的不变性学习机理，并对比了卷积层、全连接层与全局平均池化层的不变性强度。通过对两个全连接层之间的高维权重向量进行了*t*-SNE可视化分析，进一步研究了全连接层的特征综合机理。主要结论如下：



（a） AlexNet网络Fc2层特征可视化

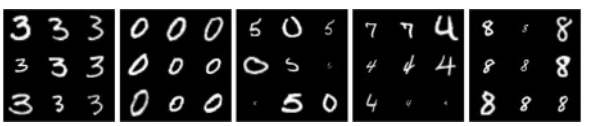


（b） ResNet18网络AvgPool层特征可视化

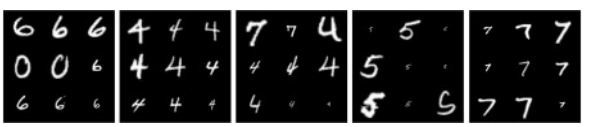
图 9 旋转增强训练I条件下AlexNet网络Fc2层和ResNet18网络AvgPool层的神经元以网络为中心的可视化和以数据为中心的可视化结果

Fig. 9 The network-centric visualization and data-centric visualization results of neurons in the Fc2 layer of AlexNet network and the AvgPool layer of ResNet18 network under rotation-enhanced training I conditions

以网络为中心的可视化结果中，包含同一种局部特征的各个旋转角度，对应的以数据为中心的可视化结果中，也包含各个旋转角度的目标，均说明对应网络层的神经元可以捕获各个旋转角度的局部特征，即具有旋转不变性。



（a） AlexNet网络Fc2层特征可视化



（b） ResNet18网络AvgPool层特征可视化

图 10 缩放增强训练I条件下AlexNet网络Fc2层和ResNet18网络AvgPool层的神经元以数据为中心的可视化结果。

Fig. 10 The data-centric visualization results of neurons in the Fc2 layer of AlexNet network and the AvgPool layer of ResNet18 network under scaling-enhanced training I conditions

（1）全连接层第一层已经能够真正的学习到平移不变的特征，该层神经元对位置的敏感区域随着训练的进行逐渐扩大，直到完全覆盖到画布所有位置。而AvgPool层具有更强的平移不变性，

（2）特征整合遵循同类特征正向激发、异类特征负向抑制的连接逻辑。此外，在MNIST数据集上的分析表明，网络能够提取并整合跨类别的共有局部特征。

（3）而全连接层的两层级联结构的旋转和缩放不变性比AvgPool层结构更强。这对实际应用中不变性结构的选择可以提供一定的指导。

参考文献

[1] BLYTHING R, BISCIONE V, VANKOV I I, et al. The human visual system and CNNs can both support robust online translation tolerance following extreme displacements[J]. Journal of Vision, 2021, 21(2): 9-9.

[2] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 2002, 86(11): 2278-2324.

[3] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[C]//Advances in Neural Information Processing Systems. 2015.

[4] BISCIONE V, BOWERS J. Learning translation invariance in CNNs[J]. arXiv preprint arXiv:2011.11757, 2020.

[5] LAPTEV D, SAVINOV N, BUHMANN J M, et al. Ti-pooling: Transformation-invariant pooling for feature learning in convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 289-297.

[6] 顾正强, 张严, 张冰尘 . 一种基于模糊滤波提高SAR自动目标识别平移不变性的方法[J]. 系统工程与电子技术, 2020.
GU ZhengQiang, ZHANG Yan, ZHANG Bingchen. Improving translation invariance of SAR automatic target recognition based on blur filtering method[J]. Systems Engineering and Electronics, 2020, 42 (11): 2488-2496.

[7] 李俊英 . 深度卷积神经网络的旋转等变性研究[D]. 浙江大学, 2019.
LI Junying. Research on Rotation Equivariance of Deep Convolutional Neural Networks[D]. Zhejiang University, 2019.

[8] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25.

[9] HERMANN K, CHEN T, KORNBLITH S. The origins and prevalence of texture bias in convolutional neural networks[C]// Proceedings of the Advances in Neural Information Processing Systems. 2020.

[10] MYBURGH J C, MOUTON C, DAVEL M H. Tracking translation invariance in CNNs[C]//Proceedings of the Southern African Conference for Artificial Intelligence Research. Springer, 2020: 282 – 295.

[11] KAUDERER-ABRAMS E. Quantifying translation-invariance in convolutional neural networks[J]. arXiv preprint arXiv: 1801.01450, 2017

[12] POSPISIL D A, PASUPATHY A, BAIR W. ‘Artiphysiology’ reveals V4-like shape tuning in a deep network trained for

image classification[J]. Elife, 2018, 7: e38242.

[13] BAKER N, LU H, ERLIKHMAN G, et al. Local features and global shape information in object classification by deep convolutional neural networks[J]. Vision Research, 2020, 172: 46-61.

[14] ERHAN D, BENGIO Y, COURVILLE A, et al. Visualizing higher-layer features of a deep network[J]. University of Montreal, 2009, 1341(3): 1.

[15] YOSINSKI J, CLUNE J, NGUYEN A, ET al. Understanding neural networks through deep visualization[J]. arXiv preprint arXiv:1506.06579, 2015.

[16] NGUYEN A, YOSINSKI J, CLUNE J. Understanding neural networks via feature visualization: A survey[M]//Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, 2019: 55-76.

[17] RIBEIRO M T, SINGH S, GUESTRIN C. " Why should i trust you?" explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1135-1144.

[18] MAATEN L VAN DER, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9 (Nov): 2579-2605.

[19] DENG L. The MNIST database of handwritten digit images for machine learning research [best of the web][J]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142.

[20] KAYHAN O S, GEMERT J C VAN. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14274-14285.

[21] HAN Y, ROIG G, GEIGER G, et al. Scale and translation-invariance for novel objects in human vision[J]. Scientific Reports, 2020, 10(1): 1411.

[22] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks [C]//Proceedings of the European Conference on Computer Vision. Springer, 2014: 818–833.

[23] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2921-2929.

[24] 尚骏远, 杨乐涵, 何琨. 基于特征可视化分析深度神经网络的内部表征[J]. 计算机科学, 2020, 47(5): 190-197.
SHANG Junyuan, YANG Lehan, HE Kun. Analyzing Internal Representations of Deep Neural Networks Based on Feature Visualization[J]. Computer Science, 2020, 47(5): 190-197.

[25] ZHANG Q, YANG Y, MA H, et al. Interpreting cnns via decision trees[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6261-6270.

[26] HU J, GAO J, YE J, et al. Model LEGO: Creating models like disassembling and assembling building blocks[C]// Proceedings of the Advances in Neural Information Processing Systems. 2024.

[27] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images[C]//Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition. 2015: 427-436.

[28] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.

[29] WANG F, LIU H, CHENG J. Visualizing deep neural networks by alternately image blurring and deblurring[J]. Neural Networks, 2018, 97: 162-172.

[30] 王芳珍, 张小丽, 赵琦武, 等. 一维卷积神经网络在机械故障特征提取中的可解释性研究[J]. 西安交通大学学报, 2025,59(7):24-35.
WANG Fangzhen, ZHAGN Xiaoli, ZHAO Qiwu, et al. Study on the Interpretability of One-Dimensional Convolutional Neural Networks in Mechanical Fault Feature Extraction[J]. Journal of Xi’an Jiaotong University, 2025, 59(7): 24-35.

[31] ROUSSEEUW P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20: 53-65.

[32] JANSSON Y, LINDEBERG T. Scale-invariant scale-channel networks: Deep networks that generalise to previously unseen scales[J]. Journal of Mathematical Imaging and Vision, 2022, 64(5): 506-536.