数据增强下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]，邢振榕[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：针对目前数据增强提升深度卷积网络不变性的研究缺乏对其实现不变性的机理进行阐释的问题，对其位置、方向等信息的表征机制和特征综合机理展开研究。首先，采用圆形和方形二分类任务构建简化实验场景，使用其增强数据训练网络；然后，通过最大激活方法对全连接层和全局平均池化层进行特征可视化，并采用 t-SNE 降维揭示了全连接层间高维权重向量的分类决策机理。结果表明：全连接层首层通过逐步扩大位置感知域的方式实现平移不变性；在圆方数据集平移增强实验中，Fc2 层平移不变性得分为 0.99，较 Conv5 层提升了 45%；全连接层通过线性加权机制综合圆形和方形的局部特征，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；ResNet 架构中平均池化层展现更优的平移不变性，在圆方数据集增强实验中，平均池化层得分为 0.93，高于全连接层的 0.79，而级联全连接结构则表现出更强的旋转和尺度不变性。以上结论在 MNIST 数据集的扩展实验上也得到了验证。

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：TP391.4 **文献标志码**：A
DOI：10.7652/xjtuxb202609021 **文章编号**：0253-987X(2026)09-0218-11

Interpretability Study of the Invariance Mechanism of Deep Convolutional Networks under Data Augmentation

LIU Ni[1], LEI Jiaojiao[1], XING Zhenrong[1], HAN Wenjing[1], HE Lihuo[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: Addressing the issue that current research on enhancing the invariance of deep convolutional networks through data augmentation lacks an explanation of the mechanisms underlying its implementation, this paper explores the representation and pooling mechanisms of position, orientation, and other information. First, a simplified experimental scenario is constructed using a binary classification task of circles and squares, which is employed to augment data for network training. Then, feature visualization is conducted for the fully connected layer and global average pooling layer using the maximum activation method, and t-SNE dimensionality reduction is adopted to reveal the classification decision-making mechanism of high-dimensional weight vectors within the fully connected layers. It is shown that translation invariance is achieved by the first

收稿日期：2025-10-13。 作者简介：刘妮(1985—)，女，讲师；何立火(通信作者)，男，教授，博士生导师。 基金项目：国家自然科学基金区域创新发展联合基金资助项目(U23A20682，U22A20247)；陕西省自然科学基础研究计划资助项目(2024JC-YBMS-566)。
网络出版时间：2026-04-29 网络出版地址：https://link.cnki.net/urlid/61.1069.T.20260428.2001.002

fully connected layer through the gradual expansion of the position-aware receptive field. In the translation augmentation experiment on the circle-and-square dataset, a translation invariance score of 0.99 is obtained by the Fc2 layer, representing a 45% improvement over that of the Conv5 layer. The local features of circles and squares are synthesized by the fully connected layers through a linear weighting mechanism, following a connection logic in which same-category features are positively activated and different-category features are negatively suppressed. Better translation invariance is exhibited by the average pooling layer in the ResNet architecture. In the augmentation experiment on the circle-and-square dataset, a score of 0.93 is obtained by the AvgPool layer, higher than the 0.79 obtained by the fully connected layer, whereas stronger rotation and scale invariance is exhibited by the cascaded fully connected structure. These conclusions are also verified through extended experiments on the MNIST dataset.

Keywords: deep convolutional network; 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]。数据增强作为提升深度卷积网络分类鲁棒性的常用手段，其作用机理值得深入研究。

受生物模型启发，深度卷积网络(ConvNets)引入了卷积与池化运算。传统观点普遍认为，卷积和池化在结构上赋予了网络平移不变性[2]。然而，由于卷积步长的设计未遵循采样定理，且全连接层不具备空间推理能力，因此 ConvNets 实际上并不具备完全的平移不变性[3]。针对这一问题，已有研究尝试构造专门的不变性网络结构[2,4]，但此类方法往往面临以下局限：仅对特定应用有效[5]；只能处理部分变换角度；需引入额外参数；可能导致模型精度下降[6]。

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[7]。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，网络便能在训练过程中自动习得对应的不变性特征。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，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。Hermann 等[8]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要原因。

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度对比网络层的整体不变性程度[3,9-10]，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。对此，本文以 AlexNet 和 ResNet 模型中应用较为广泛的两类结构：全连接层(Fc)和平均池化层(AvgPool)为研究对象，从特征可视化和特征综合两个角度，分析这两类结构如何习得对位置、方向和尺度变换具有不变性的特征，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[11]，构建了圆形与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增强训练条件下网络的特征可视化分析，并在更加复杂的 MNIST 数据集上进行验证，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[12-14]，本文利用该方法，对经过平移、旋转和尺度增强训练的 AlexNet 和 ResNet 模型的关键结构进行可视化。实验发现 AlexNet 全连接第一层提取的特征与位置、方向和尺度无关。训练过程中神经元的可视化结果展现了网络是如何在训练中逐渐扩大位置感知域的。

(2)对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[15]。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[16]，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 MNIST 数据集[17]上进行了实验验证。

(3)以 ResNet 为代表的网络以平均池化层取代

了 AlexNet 中的全连接层。由于 AvgPool 层对特征图上的滤波响应取全局平均值，因此其在平移不变性上强于全连接层。然而，在旋转和尺度不变性上，全连接层的两层级联结构则强于 AvgPool 层。

1 相关工作

1.1 深度卷积网络的变换不变性

Semih 等[18]对不同网络层的实验发现，卷积层对物体的绝对位置敏感。Myburgh 等[9]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 等[19]发现，经过 ImageNet 模型预训练的 5 层网络体现出了一定的平移不变性。顾正强等[5]使用卷积网络实现合成孔径雷达自动目标识别，采用平移敏感度展示模型的平移不变性。Laptev 等[4]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有效方法。直接可视化滤波器权重值或响应图的方法，其结果往往只有第一层是可以解释的，高层特征由于变得更加抽象，即使可视化出来也很难理解，导致该类方法失去效果。

显著性图能够展示出对模型的预测贡献最大的图像像素。Zeiler 等[20]通过反卷积网络将特征激活映射回输入像素空间，揭示了哪些输入模式能够引发特定的输出响应。类激活映射[21]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[20]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[14]。在机器学习中，可视化神经元的首选刺激有助于了解神经元所学特征。一种方法是从数据集中寻找使神经元响应最大的图像[13,20]；另外一种方法是生成视觉刺激图像[12]。后者通过梯度下降迭代优化输入，使目标神经元的激活值最大，从而揭示其学习特征。这两种激活最大化方法分别被称为以数据为中心和以网络为中心的方法，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[14]。

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。一种局部可解释且模型无关的方法(LIME)[15]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 等[22]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。Hu 等[23]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。

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 $\boldsymbol{x} \in \mathbf{R}^{C\times H\times W}$，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，假设固定参数 θ(包括权重和偏差)，$a_{ij}(\theta, \boldsymbol{x})$ 表示网络第 j 层指定神经单元 i 的激活响应。此方法的目标是找到一个最优输入 $\boldsymbol{x}^*$，使得 $a_{ij}(\theta,\boldsymbol{x})$ 达到最大值[12]，其优化目标函数为

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

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 $\boldsymbol{x}^*$。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 $\boldsymbol{x}$ 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[12,24-25]。优化目标为

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}}(a_{ij}(\boldsymbol{x}) - R_\theta(\boldsymbol{x})) \tag{2}$$

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 $\boldsymbol{x}$ 映射到自身的正则化版本。本文采用了文献[13]的梯度下降框架，该方法轮流进行 $a_{ij}(\boldsymbol{x})$ 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，每一步的更新公式为

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

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[13]，以可视化结果的主观最优效果为选择依据。本文设置学习率 $\eta = 0.01$，迭代次数 $T = 200$。L_2 衰减系数 $\theta_{\text{decay}} = 0.1$，高斯模糊半径 $\theta_{\text{b\_width}} = 0.1$，每隔 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[26]。为了克服这些局限性，本研究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[20]。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。

3　平移增强网络的特征可视化

AlexNet、VGG、ResNet 等经典网络不仅在实际应用中仍有广泛需求，而且其卷积层、池化层、全连接层等核心模块也被大量新型网络所沿用。基于此，本文以 AlexNet 和 ResNet18 为研究对象，通过特征可视化方法对经数据增强训练后的网络关键层神经元进行分析，并在此基础上探究全连接层特征的综合机理。

3.1　实验数据集和参数设置

受文献[11]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 数据集上进行了实验验证。

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 像素，正方形的边长范围为 10～56 像素。画布尺寸为 224×224 像素，目标可随机出现在画布的任意位置。该数据集一共包含8 000 张圆形和 8 000 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：将尺寸为 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 70 000 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。实验前将输入图像像素统一调整为 224×224 像素并进行归一化处理。模型训练采用 Adam 优化器和交叉熵损失函数[27]。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[11]。

3.2　平移敏感度实验

文献[10]利用平移敏感度量化网络的平移不变性，发现相比于卷积层，Fc 层具有更强的平移不变性。本文进一步利用平移敏感度对 Fc 层与 AvgPool 层两种结构的平移不变性进行比较。

平移敏感度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。采用余弦相似度[10]，值越大表示向量越相似，对应表明网络对平移不敏感。平移敏感度 S_T 的计算公式为

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

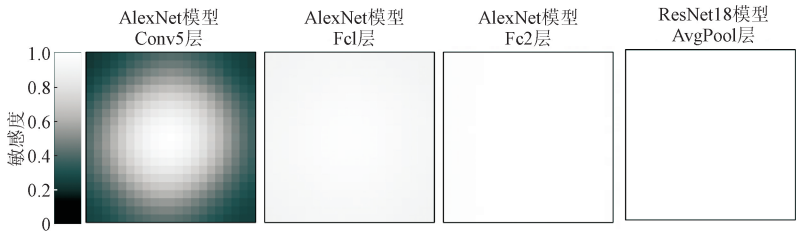
式中：$S_T(x,y)$ 为位置 (x,y) 处的像素敏感度；V 为基准图像响应；$V'(x,y)$ 为测试图像在平移位置 (x,y) 处的响应；$\cos_{sim}(\cdot)$ 为余弦相似度。通过计算敏感度，以图像中心为原点，在不同半径上可得到径向敏感度[9]；以平移敏感度图中的全范围敏感度均值作为不变性得分，衡量网络层的整体平移不变性程度。

首先，在圆方数据集上对 AlexNet 和 ResNet18 进行训练，分别计算其网络层的平移敏感度和径向敏感度。对每一个类别，都生成 5 张基准图像，正方形边长为 10～28 像素，圆形半径为 5～14 像素。对每一个基准图像，计算其与 441 个平移位置的测试图像的相似度。平移敏感度大小为 21×21。对得到的平移敏感度计算均值得到最终的平移敏感度。

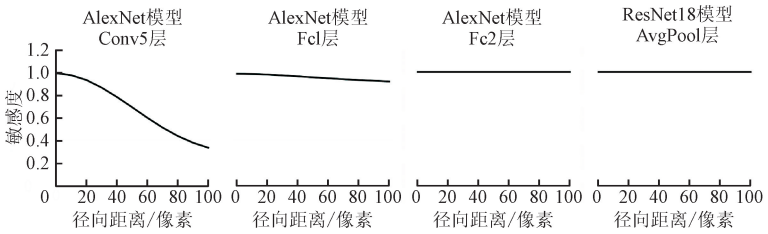
为了进行对比，本文还设计了第 2 种平移增强训练的方案，训练物体被限定在较小的图像区域内。以图像中心为原点，x 方向和 y 方向的随机平移量均限定在－35～35 像素范围内。这两种训练方案分别称为平移增强范围Ⅰ和平移增强范围Ⅱ。前者的训练物体覆盖区域大，后者的训练物体覆盖区域小。

图 1 分别展示了两种增强训练条件下 AlexNet 的最后一层卷积（Conv5）、全连接第 1 层（Fc1）、全连接层第 2 层（Fc2），以及 ResNet18 的 AvgPool 层的平移敏感度和径向敏感度。平移敏感度图的灰度从黑到白对应的余弦相似度从 0～1，颜色越白对应相似度越高。径向敏感度图中的横轴代表测试对象相对于基准对象的径向距离。

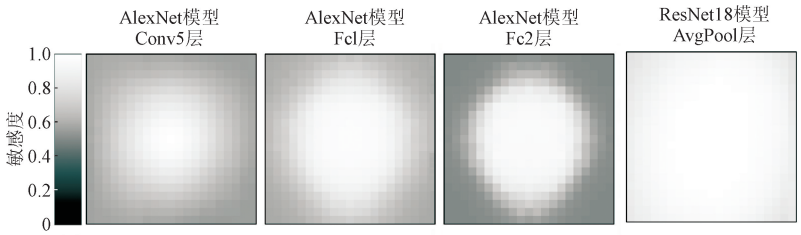
(a)平移增强训练Ⅰ和Ⅱ的示意图



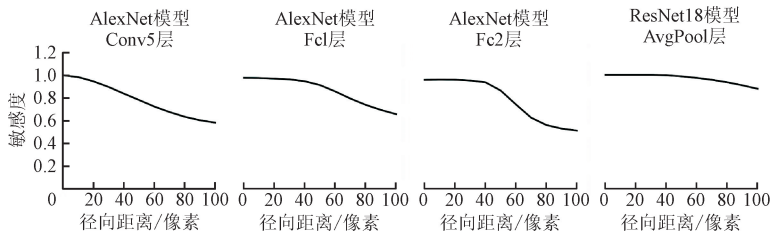
(b)平移增强训练Ⅰ时的平移敏感度



(c)平移增强训练Ⅰ时的径向敏感度



(d)平移增强训练Ⅱ时的平移敏感度



(e)平移增强训练Ⅱ时的径向敏感度

图 1 AlexNet 的 Conv5、Fc1、Fc2 层和 ResNet 的 AvgPool 层的平移敏感度和径向敏感度

Fig. 1 Translation sensitivity and radial sensitivity of the Conv5, Fc1, Fc2 layers of AlexNet and the AvgPool layer of ResNet

无论哪种情况，Conv5 层的平移敏感度只在图像中心附近比较高，表明经过平移的目标与未经过平移的目标在 Conv5 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 AlexNet 的 Conv5 层不具有平移不变性。以平移敏感度图中的全范围敏感度均值作为不变性得分，Conv5 层在本文设置的增强范围Ⅰ和Ⅱ下的平移不变性得分分别为 0.68 和 0.78，两种方案都较低。

当为增强范围Ⅰ时，Fc2 层在不同平移距离下

的相似度变化都不大；当为增强范围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 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。增强范围Ⅰ下，Fc2 层的平移不变性得分为 0.99，较 Conv5 层提升了 45%；增强范围Ⅱ下 Fc2 层得分为 0.79，而增强范围Ⅰ下为 0.99，说明增强范围越大，全连接层获得的平移不变性越强。

ResNet18 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。可见，在两种增强范围下，AvgPool 层都有接近于 1 的相似度值。

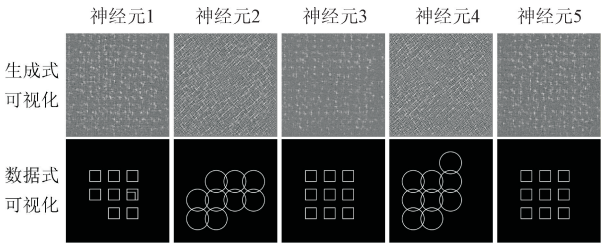
尤其是在增强范围Ⅱ时，AvgPool 层在处理位置变化时展现出比 AlexNet 更优的平移不变性，不管是否进行了平移增强训练。这主要是由于其计算的是二维特征图的平均值，从而对位置变化具有较强的鲁棒性。在增强范围Ⅱ条件下，AvgPool 层的平移不变性得分为 0.93，而同条件下 AlexNet 的 Fc2 层为 0.79，AvgPool 层比 Fc2 层高 0.14，进一步确认了全局平均池化在平移不变性上的结构性优势。

3.3　特征可视化实验结果与分析

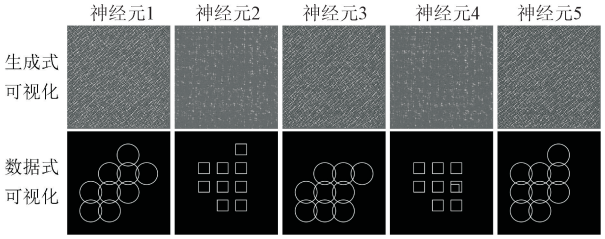
AlexNet 的 Fc1 和 Fc2 层各包含 4 096 个神经元。图 2 展示了部分神经元（随机选取了 5 个）的可视化结果。同一神经元的两种可视化结果被放在一起，分别位于上、下两行。以数据为中心（简称数据式）的可视化选取 Top-9 样本进行展示，将 Top-9 集内的样本绘制在同一张图上。

对于经过圆方数据集训练的 Fc1 和 Fc2 层，以网络为中心（简称生成式）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)神经元 1、3、5）；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（图 2(a)神经元 2、4）。从它们对应的数据式可视化结果可知，这两类特征分别对应方形和圆形的局部特征。Fc1 层的激活图均匀地分布在图像的各个位置，说明从 Fc1 层开始，特征已经与位置无关了。

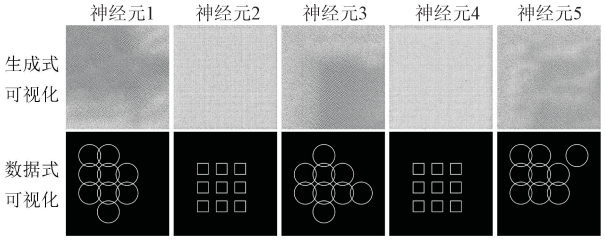
图 2(c)显示了 ResNet18 的 AvgPool 层的可视化结果。图中呈现了类似 AlexNet 的结果，包含横竖线条或角这两类特征，并且均匀地分布在整个图像上，也就是说该层的神经元表征的特征与位置无关。



(a)AlexNet 模型 Fc1 层



(b)AlexNet 模型 Fc2 层



(c)ResNet18 模型 AvgPool 层

图 2　平移增强Ⅰ的 AlexNet 和 ResNet18 模型对圆方数据集分类的部分可视化结果

Fig. 2　Partial visualization results of the classfication of circular square dataset by AlexNet and ResNet18 with translation augmentation Ⅰ

为了解 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批次时 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。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。

在 T-MNIST 数据集上的实验结果如图 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 层本身对位置不敏感。

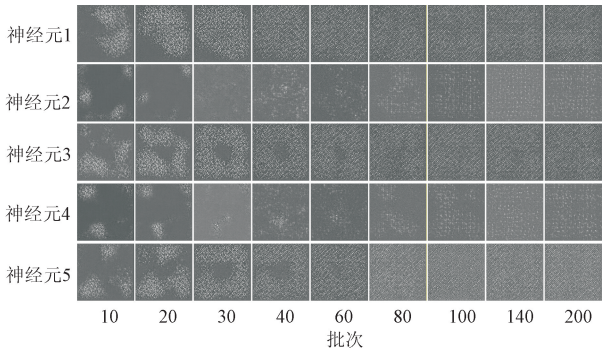
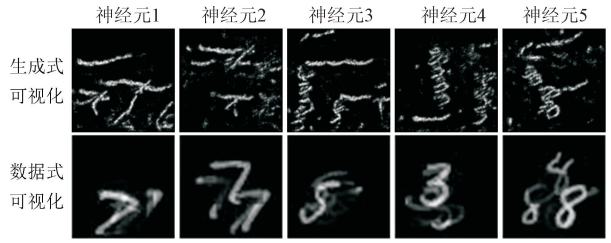
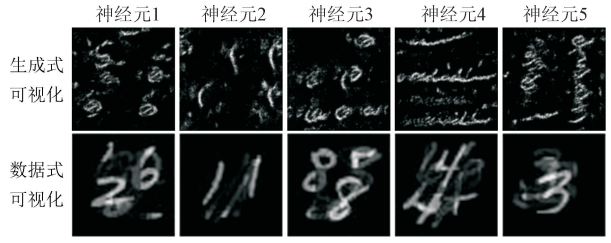


图 3 平移增强Ⅰ的 AlexNet Fc1 层神经元训练中不同批次的可视化结果

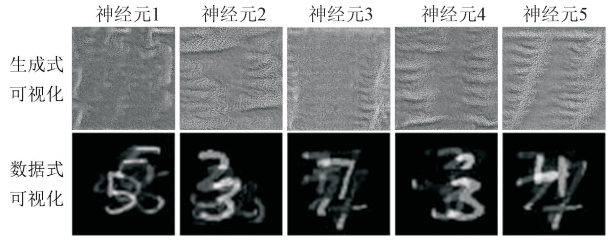
Fig. 3 Feature visualization results of different batches in AlexNet Fc1 layer neuron training of translation augmentation Ⅰ



(a)AlexNet 模型 Fc1 层



(b)AlexNet 模型 Fc2 层



(c)ResNet18 模型 AvgPool 层

图 4 平移增强Ⅰ的 AlexNet 和 ResNet18 模型对 MNIST 数据集分类的部分可视化结果

Fig. 4 Partial visualization results of the classfication of the MNIST dataset by AlexNet and ResNet18 with translation augmentation Ⅰ

3.4 利用 *t*-SNE 探究全连接层特征综合机理

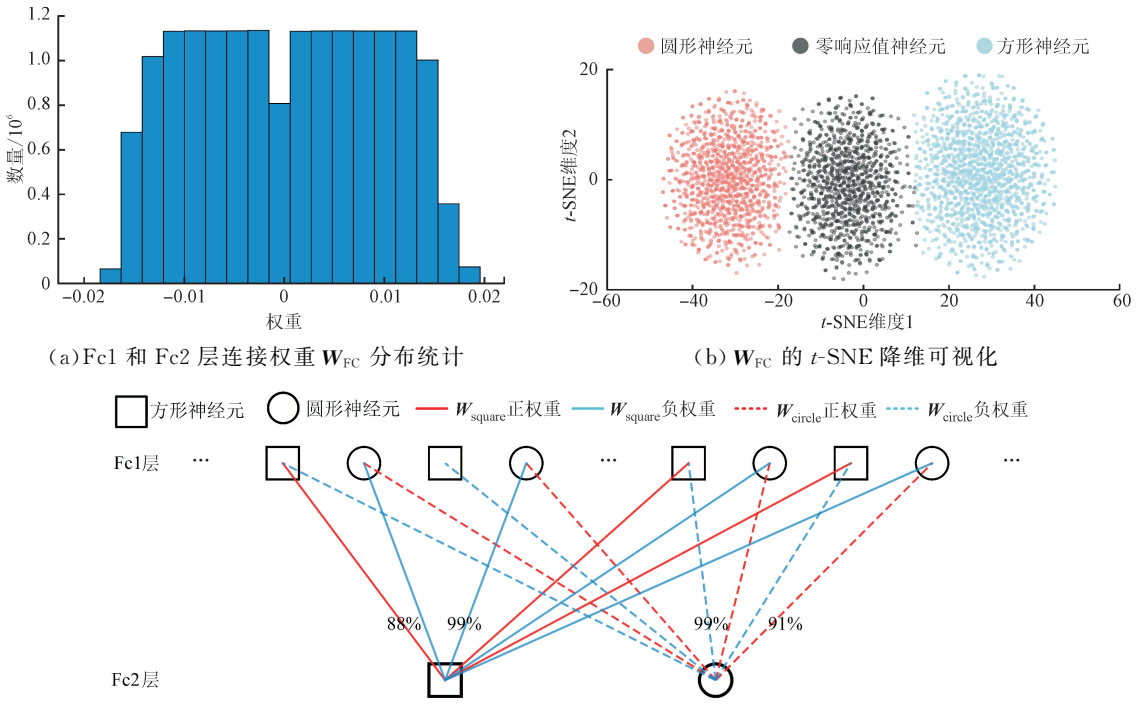
综上,AlexNet 网络 Fc1 层已经可以形成与位置无关的特征,接下来进一步分析 Fc2 层是如何综合这些信息并输出到 softmax 层进行分类的机理。本节主要通过对 Fc1p 和 Fc2 层之间的连接权重矩阵 $\boldsymbol{W}_{\mathrm{FC}} \in \mathbf{R}^{4\,096\times 4\,096}$ 进行分析。

首先,对 $\boldsymbol{W}_{\mathrm{Fc}}$ 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 轴对称分布,如图 5(a) 所示。然后,采用 t-SNE 降维技术对 $\boldsymbol{W}_{\mathrm{FC}}$ 的列向量进行降维,并对降维结果进行 K-means聚类,如图 5(b)所示,降维后每个点对应一个 Fc2 层神经元。结果可以明显地划分成三部分:中间黑色区域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,其以网络为中心的特征可视化结果为全黑图像,不包含任何特征,说明对应神经元不学习任何特征;红色区域对应的神经元主要表征方形特征;蓝色区域对应的神经元主要表征圆形特征,二者均通过以数据为中心的可视化方法验证。

从 3.3 节可知,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 $\boldsymbol{W}_{\mathrm{square}} \in \mathbf{R}^{1\times 4\,096}$ 和 $\boldsymbol{W}_{\mathrm{circle}} \in \mathbf{R}^{1\times 4\,096}$,对应图 5(c)中的实线和虚线。进一步分析发现,$\boldsymbol{W}_{\mathrm{square}}$ 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,$\boldsymbol{W}_{\mathrm{circle}}$ 与 Fc1 层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,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。

为了在更真实的数据集上进行实验分析,选取 MNIST 数据集 1、2、3、4 共 4 个类别,生成数据集对网络进行平移增强训练。对 $\boldsymbol{W}_{\mathrm{FC}}$ 进行 t-SNE 降维和 K-means 聚类分析。利用轮廓系数法[28]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,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,如数字“2”和“3”,数字“1”和“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。



(a)Fc1 和 Fc2 层连接权重 $\boldsymbol{W}_{FC}$ 分布统计　　(b) $\boldsymbol{W}_{FC}$ 的 t-SNE 降维可视化

(c)Fc2 层两类神经元各自权重均值与 F1 层神经元的连接路径示意图

图 5　圆方数据集平移增强Ⅰ的 AlexNet 全连接层特征综合机理分析

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

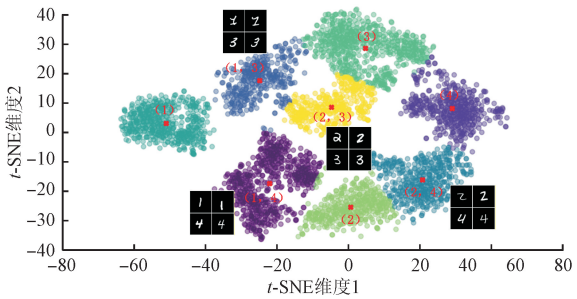


图 6　AlexNet 的 $\boldsymbol{W}_{FC}$ 的 t-SNE 降维及聚类中心神经元的数据式可视化结果

Fig. 6　t-SNE dimensionality reduction and clustering of $\boldsymbol{W}_{FC}$ for AlexNet and data-centric visualization results of clustering center neurons

可以推测，随着数据集特征变得越加复杂，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。

4　旋转与尺度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。该数据集一共包含 70 000 张图像，训练集、验证集和测试集的比例为 7∶2∶1。使用此数据集称为旋转增强训练Ⅰ。

S-MNIST：参考文献[29]，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 70 000 张图像，训练集、验证集和测试集的比例为 7∶2∶1。使用此数据集训练称为尺度增强训练Ⅰ。

4.2　旋转与尺度敏感度实验

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度与尺度敏感度实验。旋转敏感度实验中，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。尺度敏感度实验中，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，设计了数据增强方案Ⅱ。其中，旋转增强训练Ⅱ将旋转角度限定在－30°～30°。方案Ⅱ将缩放倍数限定在 2～6 范围内。

对数据增强Ⅰ、Ⅱ下的 AlexNet 网络的 Conv5、Fc1 和 Fc2 层和 ResNet18 网络的 AvgPool 层绘制旋转敏感度与尺度敏感度，分别如图 7 和图 8 所示。

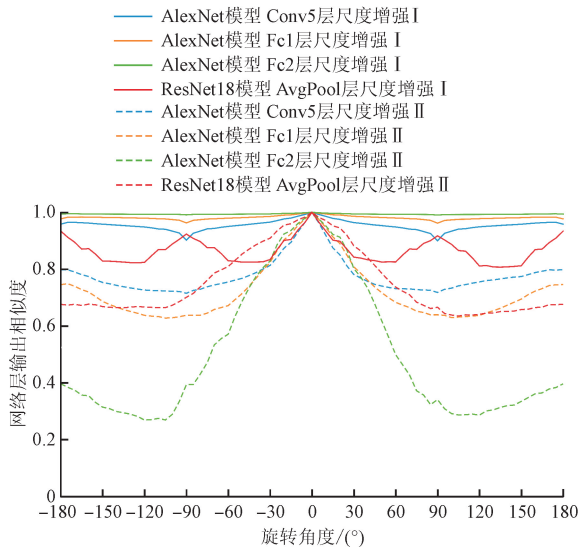


图 7 旋转增强Ⅰ、Ⅱ的 AlexNet 和 ResNet18 部分网络层的旋转敏感度

Fig. 7 Rotation augmentation rotation sensitive maps of AlexNet and ResNet18 partial network layers of Ⅰ and Ⅱ

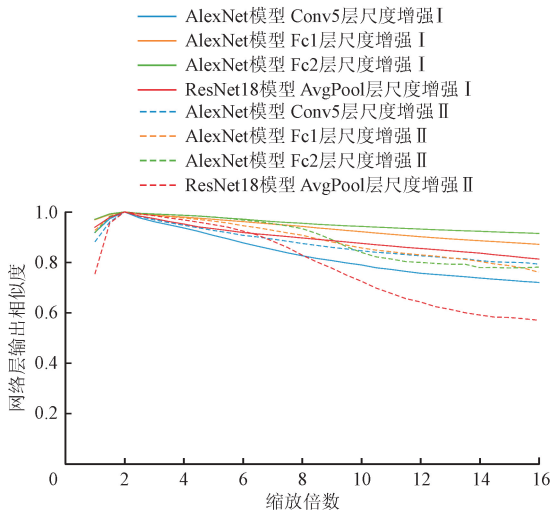


图 8 尺度增强Ⅰ、Ⅱ的 AlexNet 和 ResNet18 部分网络层的尺度敏感度

Fig. 8 Scale augmentation scale sensitive maps of AlexNet and ResNet18 partial network layers of Ⅰ and Ⅱ

以旋转和尺度敏感度的全范围余弦相似度均值作为不变性得分，衡量网络层的整体不变性程度。可以看出，经过特定的数据增强训练后，网络从全连接层开始逐渐表现出对应变换的不变性，且 Fc2 层的不变性强于 Fc1 层。以旋转不变性得分为例，增强范围Ⅰ下 Fc1、Fc2 层得分分别为 0.98 和0.99，表明全连接层具有较好的旋转不变性。对于旋转与尺度变换，AvgPool 层相较于全连接层并无优势，旋转增强Ⅰ下 Fc2 层得分为 0.99，AvgPool 层为0.87；尺度增强Ⅰ下 Fc2 层得分为 0.95，AvgPool 层为0.90，两类变换下全连接层均优于 AvgPool 层。

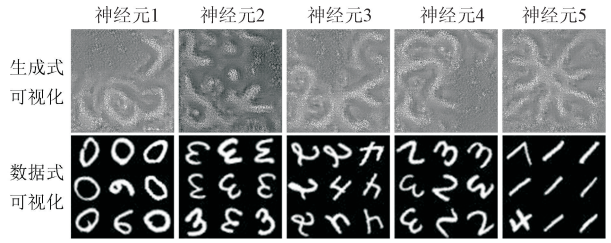
进一步观察发现，无论是旋转还是尺度变换，AlexNet 与 ResNet18 均只在训练时所覆盖的数据增强范围内表现出不变性。以旋转变换为例，经旋转增强Ⅰ训练的网络在各个旋转角度下均表现出较高的不变性，而经旋转增强Ⅱ训练的网络则仅在其学习过的角度范围(－30°～30°)内表现出不变性。尺度变换的情况与此一致，网络同样只在训练所涵盖的缩放倍数范围内表现出不变性。

与 Fc1 和 Fc2 层相比，ResNet18 的 AvgPool 层在旋转和尺度不变性上的波动更大，这一现象在平移敏感性实验中同样存在。总之，当平移、旋转或尺度的幅度较大时，各层的不变性均有所减弱，而全连接层随着网络深度的增加，对这种幅度变化的鲁棒性相对更强。

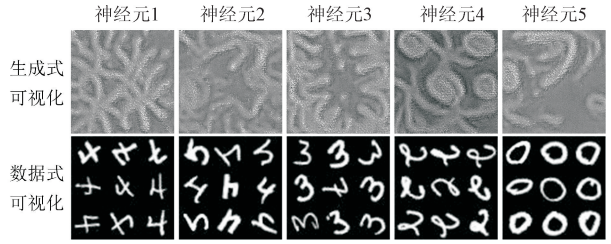
4.3 特征可视化实验结果与分析

图 9(a)为旋转数据增强实验结果，每个网络层展示 5 个神经元的可视化结果。为了便于对比，同一神经元的以网络为中心和以数据为中心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。可以看出，各神经元能够捕捉具有不同旋转角度的局部特征。在图 9(a)神经元 3 的可视化结果中，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；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。ResNet18 网络 AvgPool 层的全部 512 个神经元的部分可视化实验结果如图 9(b)所示。与 AlexNet 模型 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，在本文尺度数据集上，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。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 Top-9 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。



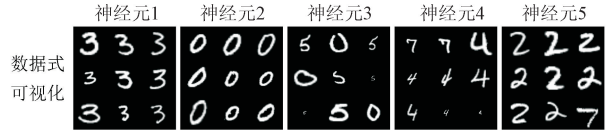
(a)AlexNet 模型 Fc2 层可视化



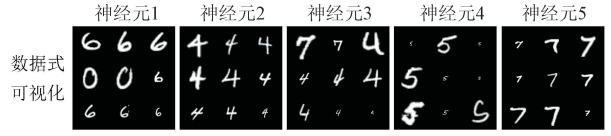
(b)ResNet18 模型 AvgPool 层可视化

图 9　旋转增强Ⅰ的 AlexNet 和 ResNet18 部分可视化结果

Fig. 9　Partial visualization results of AlexNet and ResNet18 of rotation augmentation Ⅰ



(a)AlexNet 模型 Fc2 层可视化



(b)ResNet18 模型 AvgPool 层可视化

图 10　尺度增强Ⅰ AlexNet 和 ResNet18 部分可视化结果

Fig. 10　Partial visualization results of AlexNet and ResNet18 of scale augmentation Ⅰ

5　结论与讨论

(1)全连接层第一层已经能够学习到平移不变的特征，该层神经元对位置的感知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。在本文圆方数据集实验中，平移增强Ⅰ下 Fc2 层平移不变性得分为 0.99，较 Conv5 层提升了 0.31；平移增强Ⅱ下 Fc2 层得分为 0.79，低于增强Ⅰ的 0.99，说明增强范围越大，全连接层获得的不变性越强。

(2)AvgPool 层比全连接层具有更强的平移不变性。圆方数据集平移增强Ⅱ平移不变性得分为 0.93，高于同条件下 Fc2 层的 0.79，但是全连接层的两层级联结构的旋转和尺度不变性更强。这对实际应用中不变性结构的选择可以提供指导。

(3)本文设计的圆形和方形数据集，能够直观展示全连接层权重所构成的线性机对图形进行分类的特征综合和决策机理。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。在圆方数据集实验中，同类别特征神经元间正连接权重占比大于 88%，异类神经元间负连接权重占比达 99%。

本文关于特征综合机理的分析方法对增加深度模型的可解释性做出了一定贡献，但是很难直接扩展到特征复杂的真实数据集。未来将持续研究更加复杂数据特征下模型的更加一般的可解释性方法。

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.

[2] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2015: 2017-2025.

[3] Biscione V, Bowers J. Learning translation invariance in CNNs [PP/OL]. arXiv(2020-11-06) [2025-10-13]. https://arxiv.org/abs/2011.11757.

[4] Laptev D, Savinov N, Buhmann J M, et al. TI-POOLING: transformation-invariant pooling for feature learning in convolutional neural networks [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 289-297.

[5] 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，2020，42(11)：2488-2496.
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.

[6] 李俊英. 深度卷积神经网络的旋转等变性研究 [D]. 杭州：浙江大学，2019.

[7] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2012: 1097-1105.

[8] Hermann K L, Chen Ting, Kornblith S. The origins and prevalence of texture bias in convolutional neural networks [C]//Proceedings of the 34th International

Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2020: 19000-19015.

[9] Myburgh J C, Mouton C, Davel M H. Tracking translation invariance in CNNs [C]//Artificial Intelligence Research. Cham, Switzerland: Springer International Publishing, 2020: 282-295.

[10] Kauderer-Abrams E. Quantifying translation-invariance in convolutional neural networks [PP/OL]. arXiv (2017-12-10) [2025-10-13]. https://arxiv.org/abs/1801.01450.

[11] Baker N, Lu Hongjing, 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.

[12] Erhan D, Bengio Y, Courville A, et al. Visualizing higher-layer features of a deep network [J]. University of Montreal, 2009, 1341(3): 1.

[13] Yosinski J, Clune J, Nguyen A, et al. Understanding neural networks through deep visualization [PP/OL]. arXiv(2015-06-22) [2025-10-13]. https://arxiv.org/abs/1506.06579.

[14] Nguyen A, Yosinski J, Clune J. Understanding neural networks via feature visualization: a survey [M]// Samek W, Montavon G, Vedaldi A, et al. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham, Switzerland: Springer International Publishing, 2019: 55-76.

[15] 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. New York, USA: ACM, 2016: 1135-1144.

[16] van der Maaten L, Hinton G. Visualizing data using *t*-SNE [J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605.

[17] Deng Li. The MNIST database of handwritten digit images for machine learning research [J]. IEEE Signal Processing Magazine, 2012, 29(6): 141-142.

[18] Semih K O, van Gemert J C. On translation invariance in CNNs: convolutional layers can exploit absolute spatial location [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 14262-14273.

[19] Han Yena, Roig G, Geiger G, et al. Scale and translation-invariance for novel objects in human vision [J]. Scientific Reports, 2020, 10(1): 1411.

[20] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C]//Computer Vision-ECCV 2014. Cham, Switzerland: Springer International Publishing, 2014: 818-833.

[21] Zhou Bolei, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 2921-2929.

[22] Zhang Quanshi, Yang Yu, Ma Haotian, et al. Interpreting CNNs via decision trees [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 6254-6263.

[23] Hu Jiacong, Gao Jing, Ye Jingwen, et al. Model LEGO: creating models like disassembling and assembling building blocks [C]//Proceedings of the 38th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2024: 127711-127738.

[24] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2015: 427-436.

[25] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [PP/OL]. arXiv(2014-02-19) [2025-10-13]. https://arxiv.org/abs/1312.6199.

[26] Wang Feng, Liu Haijun, Cheng Jian. Visualizing deep neural network by alternately image blurring and deblurring [J]. Neural Networks, 2018, 97: 162-172.

[27] 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, Zhang 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.

[28] 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.

[29] 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.

（编辑　亢列梅）