学校代码：10284
分 类 号：TP183
密　　级：公开
U D C：004.8
学　　号：DG20330004

南京大学

博 士 学 位 论 文

论 文 题 目　从神经元出发的卷积神经网络可解释性研究

作 者 姓 名　窦慧
专 业 名 称　计算机科学与技术
研 究 方 向　人工智能
导 师 姓 名　申富饶　教授

2025 年 12 月 4 日

答辩委员会主席 武港山 教授

评　　阅　　人 武港山 教授

路通 教授

戴新宇 教授

张道强 教授

魏秀参 教授

论文答辩日期　2025 年 11 月 19 日

研究生签名：

导师签名：

Research on the Interpretability of Convolutional Neural Networks from the Perspective of Neurons

by

Hui Dou

Supervised by

Professor Furao Shen

A dissertation submitted to

the graduate school of Nanjing University

in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science and Technology

School of Computer Science

Nanjing University

December 4, 2025

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：从神经元出发的卷积神经网络可解释性研究

计算机科学与技术 专业 2020 级博士生姓名：窦慧

指导教师（姓名、职称）：申富饶 教授

摘 要

神经网络是当今最广泛使用的人工智能技术之一，目前神经网络在各领域中展现出卓越的性能，并得到了广泛应用。然而，由于神经网络的复杂性和高维性，其运行机制是不透明的，导致潜在的信任危机。因此，理解神经网络的决策过程变得尤为重要。当前的神经网络解释方法通常与固定的神经网络模型架构紧密相关，无法广泛应用于不同类型的神经网络模块。这种局限性使得对复杂模型的全面理解变得困难，也影响了跨模块的系统性分析和解释。

本文提出一种基于神经元的可解释性方法，称为神经网络扫描仪（*Neural Network Scanner, NNS*）。神经元是神经网络模型的基本组成单元，通过神经元灵活地结合成不同模块，实现以一种统一的方式解释模型。此外，本文基于神经网络扫描仪实现从不同颗粒度定量解释神经网络模型的目标，包括：神经元级解释、滤波器级解释和模式级解释。针对不同解释目标，探索模型的运行原理和决策结果。总结来说，本文主要工作包括：

一种从神经元层面定性解释神经网络的解释方法。本文提出了一种通过可视化神经元学习过程来解释神经网络的方法，神经网络扫描仪。对于指定神经网络，神经网络扫描仪能够提取神经元所学习到的特征，并以人类可理解的形式进行展示。通过整合不同神经元所学习到的特征，可以对不同神经网络模块的工作机制进行分析。该方法适用于各类神经网络模型，且无需对模型架构进行任何修改。本文将神经网络扫描仪应用于不同网络模块中，在图像分类任务的模型上进行实验，验证该方法的有效性。通过这些实验，本文深入分析了不同神经网络模块的工作机制。

基于 NNS 的神经元层面解释方法。目前大多神经网络可解释性的研究都集中在特征图或整个模型层面，通常与特定的神经网络架构联系在一起，缺乏对神

经元定量解释的相关研究。尽管神经网络扫描仪能够有效地可视化神经元学习到的特征，但在定量衡量这些特征方面存在不足。针对这一问题，基于神经网络扫描仪的可视化结果，本文提出了一种从神经元层面出发，分析卷积神经网络中卷积层工作机制的方法。本文引入了特征量的概念，通过衡量学习图像从输入样本中获取的特征，量化每个神经元编码的信息。基于该概念，本文量化和比较不同神经元编码的特征，从三个方面对卷积层进行了系统性地分析，并在多种模型上进行了实验。

基于 NNS 的滤波器层面解释方法。目前可解释性方法大多依赖于经验性启发，并缺乏严格的数学基础。且当前的解释方法通常使用单一度量解释算法，缺乏全面解释能力。针对这一问题，本文引入了可解释特性的概念，从不同角度对滤波器学习到的特征进行有效量化，从而提取多种可解释特性。在卷积神经网络中，神经元是滤波器的基本单元。本文通过组合神经元的特征，进一步扩展神经网络扫描仪，将其应用于滤波器的分析，以揭示滤波器的行为与特性。在此基础上，本文提出了一个统一且具有良好扩展性的可解释性框架，称为模型可解释特性测量框架，旨在利用不同的可解释特性对模型进行多角度解释。基于该框架，本文设计了三种具有代表性的评估指标，以提供从不同视角出发的模型解释。本文通过实验展示了如何通过模型可解释特性测量框架探索模型特性，并结合提出的指标对模型进行多角度解释。

基于 NNS 的模式层面解释方法。现有的可解释性方法在局部或特征层面提供了一定程度的模型解释，但这类解释缺乏全局结构性，难以揭示神经网络模型在更高层次上从输入样本中所学习到的模式。针对这一问题，本文提出了一种方法，通过神经网络扫描仪对滤波器进行扫描，并对学习图像进行聚类，进而探究聚类结果与模型过拟合之间的关系。本文设计了量化指标，用以评估滤波器的特性并识别异常滤波器。本文从不同角度设计并开展实验分析异常滤波器，揭示其对模型性能的影响。

关键词：人工智能；卷积神经网络；可解释性

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Research on the Interpretability of Convolutional Neural Networks from the Perspective of Neurons

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: Hui Dou

MENTOR: Professor Furao Shen

ABSTRACT

Neural networks are among the most widely used artificial intelligence technologies today. Currently, neural networks demonstrate exceptional performance across various domains and have been widely adopted. However, due to the complexity and high dimensionality of neural networks, their operational mechanisms remain opaque, leading to potential trust issues. Therefore, understanding the decision-making processes of neural networks becomes particularly important.

Existing interpretability methods for neural networks are often tightly coupled with fixed model architectures, making them difficult to generalize across different types of network modules. This limitation hinders a comprehensive understanding of complex models and affects systematic, cross-module analysis and interpretation.

This paper proposes a neuron-based interpretability method, referred to as *Neural Network Scanner (NNS)*. Neurons are the fundamental building blocks of neural network models, and by flexibly combining them into various modules, model interpretation can be achieved in a unified way. Furthermore, based on the NNS, this paper aims to provide quantitative interpretations of neural networks at different granularities, including: neuron-level interpretation, filter-level interpretation, and pattern-level interpretation. For each interpretability objective, the operational principles and decision outcomes of the model are explored. In summary, the main contributions of this paper include:

A qualitative interpretability method for neural networks at the neuron level. This paper proposes a novel method for interpreting neural networks by visualizing

the learning process of individual neurons, referred to as the Neural Network Scanner (NNS). For a given neural network, NNS can extract features learned by neurons and present them in a human-understandable format. By integrating the features learned by different neurons, the working mechanisms of different network modules are analyzed. This method is applicable to various types of neural network models and does not require any modification to the model architecture. NNS is applied to different network architectures and experiments are conducted on image classification models to validate the effectiveness of the proposed method. Through these experiments, in-depth analysis of the operational mechanisms of different neural network models is provided.

A neuron-level interpretability method based on NNS. Most existing research on the interpretability of neural networks focuses on feature maps or the model level as a whole, and is typically tied to specific neural network architectures. In contrast, relatively little work has examined quantitative interpretations at the neuron level. Although NNS effectively visualizes the features learned by individual neurons, they fall short in providing quantitative assessments of these features. To address this limitation, this paper proposes a neuron-level analytical method for understanding the working mechanisms of convolutional layers in convolutional neural networks, based on the visualizations produced by NNS. The concept of Feature Quantity is introduced to quantify the information encoded by each neuron by measuring the extent to which learned representations capture features from input samples. Building on this concept, we quantitatively evaluate and compare the features encoded by different neurons, conduct a systematic analysis of convolutional layers from three perspectives, and validate the approach through experiments on multiple models.

A filter-level interpretability method based on NNS. Most current interpretability methods rely on empirical heuristics and lack rigorous mathematical foundations. Additionally, a single-metric explanation strategy is often used, lacking comprehensive explanatory capability. To address this issue, the concept of Interpretable Properties (IPs) is introduced to effectively quantify the features learned by filters from multiple perspectives, enabling the extraction of diverse IPs. In CNNs, neurons are the fundamental units of filters. By aggregating the features of neurons, NNS is extended to

analyze filters, revealing their behaviors and characteristics. On this basis, a unified and extensible interpretability framework called the Test with Interpretable Properties (TIP) is proposed, aiming to interpret models from multiple perspectives using different interpretability properties. Based on this framework, three representative evaluation metrics are designed to provide model interpretation from diverse angles. Through experiments, the exploration of model properties using this framework is demonstrated, along with interpretation of models from multiple perspectives using the proposed metrics.

A pattern-level interpretability method based on NNS. Existing interpretability methods provide some degree of explanation at the local or feature level, but such interpretations lack global structure and struggle to reveal higher-level patterns that neural networks learn from input samples. To address this issue, a method is proposed that uses NNS to scan filters and cluster learned images, thereby exploring the relationship between clustering results and model overfitting. Quantitative metrics are designed to evaluate filter characteristics and identify anomalous filters. Experiments and analyses are conducted from different angles to examine these anomalous filters and reveal their impact on model performance.

KEYWORDS: Artificial Intelligence; Convolutional Neural Networks; Interpretability

目　录

插图目录

表格目录

第一章　绪论

1.1　研究背景与意义

人工智能（Artificial Intelligence，AI）已经成为科学研究领域的核心主题之一，带来了深远的社会影响，且其应用已渗透到各个行业[1-2]。随着高效、可扩展的基础设施不断发展，AI 系统已经成为众多领域的关键工具，甚至在多个复杂任务上，其表现超过了人类能力[3-4]。尽管 AI 系统在预测、推荐以及决策支持等方面展现出非凡的能力，但这一成果通常依赖于高度复杂的神经网络模型。这些模型的内部逻辑难以被外界理解，因此通常被称为“黑盒”模型[5-7]。神经网络通过非线性、非单调、非多项式的函数来拟合数据中的变量关系，这使得它们的运作原理极其不透明。神经网络在训练数据集上往往能取得良好的预测结果，但有时这种表现并非源于正确的学习，而是由于模型在错误的规律下得出正确答案，这也导致了模型在实践中的不稳定性[8-11]。因此，神经网络的“黑盒”特性使得其决策过程对人类而言难以完全被信任。

近年来，针对神经网络模型的可视化和解释研究日益增多。2018 年，欧洲议会在《通用数据保护条例》（General Data Protection Regulation，GDPR）中新增了有关自动化决策的条款，明确规定了数据主体应有权了解自动化决策的相关信息。此外，2019 年，人工智能高级专家组提出了关于可信赖 AI 的伦理准则。美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）在 2020 年 8 月发布的 XAI 四大原则进一步强化了这一观点[12]：可证明性（即解释结果应有证据支持）、可用性（解释结果应易于理解且对用户有实际意义）、准确性（解释应真实反映模型的运作机制）和限制性（解释能够识别模型运作的适用范围）。

可解释人工智能（Explainable AI，XAI）[13]的研究，旨在将人工智能的复杂过程以人类能够理解的方式呈现，从而使人类能够更清楚地把握模型的决策逻辑和内部工作机制，这为模型的维护和广泛应用提供了基础[14]。神经网络的可解释

算法是通过将模型运作过程以人类可理解的方式呈现来进行解释[15]。该过程的结果会受到所选模型和问题类型的影响。由于神经网络具有非线性结构，其内部运作机制对外部观察者而言非常不透明，因此很多研究将神经网络视为“黑盒”进行分析，关注模型可解释性的重要性、需要解释的模型与任务，以及如何有效地进行解释等问题[16-17]。目前对神经网络的可解释性的具体定义仍未达成统一，且各研究对其定义存在差异，甚至在某些情况下有所冲突[18]。例如，Doshi-Velez et al.[19]将可解释性定义为“通过人类可理解的术语解释模型行为的能力”。这种表述直接使用“可理解的术语”这一概念，使得解释本身不再需要额外阐述。实际上，解释的核心在于架设人类与自动化决策之间的桥梁，使决策过程既能准确地反映系统逻辑，又能被人类轻松理解。在这一基础上，Zhang et al.[20]进一步细化了“解释”的含义。依据该定义，理想情况下，解释应能归结为逻辑决策规则，或者至少能转化为逻辑规则。然而，实际操作中并不要求每个解释都以逻辑规则形式呈现，更重要的是能够从模型中提取出一些关键元素，用于构建更加直观的解释。至于“可理解的术语”，应当与特定任务相关的领域知识（或常识）相结合，以便让使用者能够轻松理解。对于“解释”这一概念，英语中有多个对应词汇，如 interpretation 和 explanation 等。Montavon et al.[21]对这些词汇做了细致区分，其中 interpretation 被定义为将抽象概念（如分类结果）映射到人类能够理解的领域，而 explanation 则指的是在可解释领域内，对某个特定决策（例如分类或回归）的影响因素进行解读。相关研究[22-24]进一步探讨了网络可解释性与其他特性之间的相互关系，如解释性与算法复杂度、准确度及保真度等，这些内容为深入理解神经网络的工作机制提供了更多的视角。

人类对于更好理解神经网络模型的需求不断增加。对于那些决策效果低于人类水平的模型，期望通过深入分析其内部结构和运作机制，找出问题并加以改进，以提高其性能。对于决策能力接近人类的模型，提供透明的解释能够帮助人类理解其决策，从而增强信任，促使模型广泛应用。对于超越人类决策能力的模型，分析其决策过程不仅能够帮助理解问题，还能为未来的决策提供启发。随着越来越多的公司将神经网络模型应用到实际业务中，其不透明性问题逐渐暴露出来，并成为一个不可忽视的缺陷[25]。这种不透明性使得人类无法充分理解网络决策背后的原因[26]，从而无法合理预测和评估这些决策带来的风险，进而引发潜在的信任危机和安全隐患[27-28]。尤其在一些对决策结果要求极高的高风

险场景中，比如医学诊断[29-30]、自动驾驶[31-32]以及司法审判[33]等领域，神经网络的可解释性问题变得尤为重要。广泛认为，提升模型的可解释性是确保其可被信任的关键因素之一[34-36]，并且这种可解释性对于增强系统的可靠性至关重要[37]。神经网络模型在训练过程中，可能会无意间引入偏见，这些偏见往往被隐藏在模型的决策规则中，难以察觉并加以修正[38]。这些潜在的偏见可能被误认为是普遍适用的规则，且由于神经网络的黑箱特性，模型的不透明性导致很难判断其在处理诸如性别、种族等问题时是否公平[39]。不仅如此，模型的不透明性还会对责任追究[40]、产品安全[41]及行业规范等产生深远影响[42]。

可解释性研究不仅有助于解决系统的可靠性和公平性问题，还能在优化神经网络性能方面发挥重要作用。许多研究专注于提升神经网络在复杂任务中的表现，但较少关注模型为何能够取得良好表现的原因[43-45]。虽然神经网络的架构被普遍认为与其性能密切相关，但其架构与性能之间的具体关系仍然缺乏系统性的解释。这种认知的缺失意味着，尽管神经网络在一些任务中表现优秀，但仍不完全清楚为何某些网络结构能有效提升性能。若能深入理解不同网络部分的作用及其相互关系，并掌握性能变化的根本原因，将为进一步优化神经网络、设计出更高效的模型提供宝贵的指导[46-47]。

1.2 国内外研究现状及面临的挑战

1.2.1 可解释性研究现状

近年来，许多研究致力于可视化、解释和理解神经网络模型的内部机制[34]。神经网络的可解释性问题得到了广泛关注[35]，并提出了多种技术以提高模型的透明度。一个广泛应用的技术是显著性图，它通过突出输入变量来展示模型在给定样本上的预测结果。显著性图的概念最早由 Simonyan et al.[48] 提出，他们通过激活最大化生成类别图像的方法和对给定图像生成类显著图（Saliency Map）的方法。这两类方法为后续可解释性研究提供两种可行的思路：其一为解释网络单元学习到的视觉模式；其二为解释图像中对网络而言的重要区域。用后一种思路解释网络的研究得到广泛关注，Zhou et al.[49]提出类激活映射方法（Class Activation Mapping，CAM），在卷积神经网络（Convolutional Neural Networks, CNNs）中使用全局平均池化层代替全连接层以保持网络的定位能力。该方法可

以将输入图像中对 CNN 决策影响大的区域突出显示，并通过热力图表示。因为该方法需要修改网络结构，因此应用受限。在此基础上，研究者们提出了进一步强调图像中特定概念预测的重要区域的方法[50][51]。梯度加权类激活映射（Grad-CAM）[52] 和 Grad-CAM++[53] 不需要修改模型架构，即可解释模型。它们通过使用目标类输出分数的梯度作为显著性图的加权成分来进行解释。Muhammad et al.[54] 将此类方法扩展，通过使用卷积层学习到的表示的主成分来创建视觉解释。类似地，Zeiler et al.[55] 使用去卷积模型揭示了神经网络在低层次主要学习简单的边缘特征，在高层次学习更复杂的物体特征。其他方法，如 Ren et al.[56] 和 Zhang et al.[57] 提出的，专注于生成高质量的视觉解释，而不依赖于非盲去卷积，或通过提取图形模型来描述特定滤波器的内容。然而，显著性图和类似的基于梯度的方法也存在局限性。由于在大规模视觉模型中梯度的噪声特性[58]，这些方法往往仅反映模型在狭窄输入范围内的行为，这可能导致误导性的特征重要性估计[59]。为了解决这个问题，一些方法，如 Rise[60]、Sobol[61] 和 HSIC[62]，采用了对输入图像进行扰动的方式，以生成更可靠的显著图。

基于显著图的解释方法没有明确的文字解释信息，需要人通过经验进行再加工进一步进行解释，针对这一问题，Bau et al.[63]提出一种量化视觉表征可解释性的方法，文中使用像素级别语义标注信息的数据集 Broden，在网络中评估隐藏单元与数据中语义概念的关系，从而实现有语义信息的解释。同样基于语义信息的解释方法还有文献[64]，文章提出一种通过将人类可理解的特征映射到网络提取的高级特征从而解释神经网络内部状态的算法（Testing with Concept Activation Vectors，TCAV）。这类基于语义信息的解释方法需要克服人为选择语义信息可能加强人类偏见的问题。Kim et al.[64] 提出基于显著性图度量预先选择概念影响的方法，尽管该方法需要人工标注的数据库，这可能导致较高的成本。为了解决这个问题，Ghorbani et al.[65] 提出了自动化提取数据集中的概念的方法，识别出那些适用于整个数据集而非单一样本的概念。类似地，Cheng et al.[66] 开发了一种方法，能够量化深度神经网络模型中间层编码的视觉概念，而无需显式标签。

上述解释方法都需要关于网络结构的先验知识，模型无关的解释方法提供了解释网络的另一种视角。这类解释方法与网络模型无关，可以用来解释黑盒模型。Ribeiro et al.[67]提出局部可解释模型无关解释方法（Local Interpretable Model-agnostic Explanations，LIME），通过训练一个可解释的代理模型来解释模型预测

结果的局部行为。基于 LIME 算法，Ramamurthy et al.[68]提出一种模型未知的多层次解释方法（Model Agnostic Multilevel Explanations，MAME），将 LIME 应用于模型未知的全局信息解释。

1.2.2 可解释性研究面临的挑战

尽管已有众多方法试图从不同角度揭示神经网络的内部机制，但当前可解释性研究仍面临诸多关键挑战，制约了其理论发展与实际应用的深入推进。

缺乏具有普适性的可解释性方法

如前文所述，神经网络可解释性研究的研究重点具有显著差异，不同研究往往聚焦于特定的模型架构、任务背景或解释目标，导致现有方法在适用范围和通用性方面存在局限。目前主流的可解释性方法往往与特定的神经网络结构紧密耦合，这些方法依赖于模型内部特定模块的行为特征进行解释。然而，不同神经网络结构在功能模块设计上差异显著，这种高度依赖结构的解释方式限制了可解释性方法在不同模型间的可迁移性和可比较性。此外，这种结构绑定的解释方式在对比分析多种模型机制、探索其泛化能力、诊断模型行为等任务中表现出明显不足，难以在统一的视角下理解和评估不同模型的工作原理。这不仅阻碍了神经网络解释研究的系统性推进，也对实际应用中的模型选择与部署带来了挑战。因此，亟需提出一种通用性强、结构无关的解释框架，能够跨模型结构对神经网络的决策过程和内部机制进行有效分析与对比，从而推动神经网络可解释性研究走向更加统一和系统的方向。

解释结果的主观性与缺乏量化标准

当前可解释性方法普遍依赖可视化技术展示神经网络内部特征，如特征图、热力图等形式，虽在一定程度上提升了模型输出的可理解性，但高度依赖人工直觉的判断使解释结果具有显著的主观性。这种主观性导致了可解释性研究在实践中缺乏严谨性与一致性，难以实现不同模型之间的横向比较，甚至同一模型在不同实验条件下的解释也难以复现。此外，这种缺乏量化标准的解释手段无法有效判断解释是否真正揭示了模型的决策依据，也难以形成科学有效的解释体系。因此，亟需发展客观、统一、可重复的量化评价指标，以提升模型解释的可信度与实用性。

缺乏可扩展的可解释性评估框架

当前可解释性研究中，解释指标缺乏可扩展性已成为制约方法评估一致性与通用性的核心问题之一。当前研究缺乏具备可扩展性的解释性评价指标体系，使得不同模型能够在统一框架下适配多种解释粒度。此外，该体系还应具有可组合性，能够灵活引入新维度的度量方式，从而提升解释评价在实际应用中的表现力与适应性。从而为模型选择、调试与优化提供更加系统和可靠的理论基础与实证支撑。

异常行为识别能力不足

此外，当前可解释性研究在识别和理解模型的异常行为方面仍显薄弱，尤其是在发现无效结构和过拟合模式等问题上缺乏有效手段。神经网络在训练过程中可能引入大量冗余或功能异常的单元，导致模型在性能下降。部分异常结构可能会对模型输出产生误导性影响，影响其在关键任务中的可靠性和安全性。因此，发展具备异常检测与解释能力的可解释性工具，能够帮助研究者深入分析模型内部潜在缺陷，为模型诊断、结构修正以及鲁棒性提升提供关键支持，具有重要的理论与应用价值。

可解释性粒度单一，难以形成系统性认知

现有可解释性研究主要集中在模型的局部行为分析，解释粒度较为单一，常以输入样本的响应或单个神经元的行为为核心展开。这类方法虽然能够提供模型在特定情况下的细节性洞察，但往往忽略了神经网络在不同层次结构之间的协同关系，难以揭示模型整体的运行机制。这种碎片化的解释方式限制对模型系统性理解的能力，也使得解释难以扩展到复杂任务。因此，可解释性研究需要从多粒度视角出发，构建层次分明、结构清晰的解释路径，以全面呈现神经网络从底层特征提取到高层语义抽象的过程。

1.3 本文主要工作

针对当前可解释性研究面临的问题，本文以神经网络扫描仪为基础进行研究。首先，本文提出一种基于神经元的可解释性研究方法，神经网络扫描仪。神经网络模型本质上都是由神经元组成的，通过神经元灵活地结合成不同模块，可以以一种统一的方式解释不同模型，以神经元作为基础单元探索不同神经网络模型运行原理的共性和特性，从而分析和比较神经网络不同模块的工作机制。基

于神经网络扫描仪，本文实现从不同颗粒度定量解释神经网络模型的目标，包括：神经元级解释、滤波器级解释和模式级解释。针对不同解释目标，探索模型的运行原理和决策结果。根据解释目标的颗粒度不同，图1-1列举出本文各章节间的相互关系，具体内容如下：

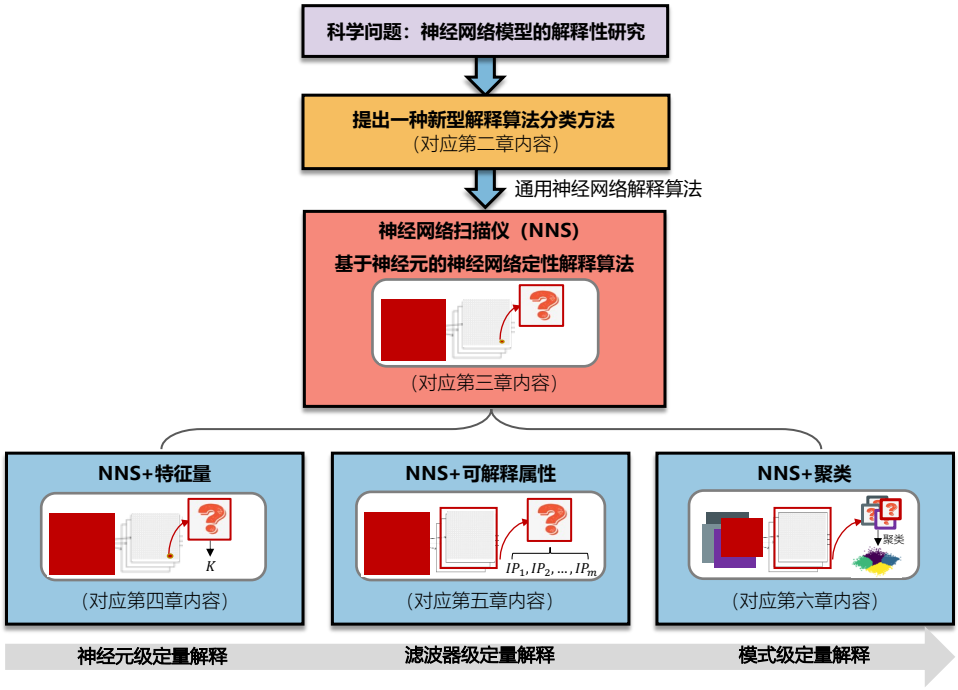


图 1-1 本文各章节之间的逻辑结构图

第二章对 CNN 模型的解释方法进行整理，并提出一种针对解释算法的新型分类方法。根据模型可解释性研究的关注点不同，本文将解释研究方法分为两个主要类别：基于网络的解释方法和基于输入的解释方法。基于网络的解释方法侧重于解释神经网络各单元所学习到的特征，而基于输入的解释方法则侧重于解释某一特定输入样本是如何导致特定输出结果的。在这一分类方法中，每个类别下都可以进一步细分为子类。基于网络的解释方法依据网络单元感兴趣模式的生成方式，可分为理想样本和真实样例两种；基于输入的解释方法依据输入方式的不同进行划分，可以分为单一输入解释和多个输入解释两类。本章对每一分类进行了详细的梳理和总结。

第三章致力于神经元级的定性解释。本文提出了一种通过可视化神经元学习过程来解释神经网络的方法，称为神经网络扫描仪。对于指定神经网络，神经网络扫描仪能够提取每个神经元所学习到的特征，并以人类可理解的形式进行展示。通过整合不同神经元所学习到的特征，可以对不同神经网络模型的工作机

制进行分析。该方法适用于各类模型，且无需对模型架构进行任何修改。本文将神经网络扫描仪应用于不同网络结构中，在图像分类任务的模型上进行了实验，以验证该方法的有效性。通过这些实验，本文深入解释了不同神经网络模块的工作机制，并从多个角度评估了神经网络的可解释性。

在第四章、第五章和第六章中，本文在第三章提出的神经网络扫描仪基础上实现从不同颗粒度定量解释神经网络模型。

第四章探索了神经元级的解释。目前大多神经网络可解释性的研究都集中在特征图或整个模型，通常与特定的神经网络架构联系在一起。这些方法使比较不同的网络体系结构的功能变得具有挑战性。尽管神经网络扫描仪能够有效地可视化神经元学习到的特征，但在定量衡量这些特征方面存在不足。由于缺乏标准评估准则的解释结果往往具有主观性，通常需要额外的人工解释。针对这一问题，基于神经网络扫描仪的可视化结果，本文提出了一种从神经元层面出发，分析卷积神经网络中卷积层工作机制的方法。本文引入了特征量的概念，通过衡量学习图像从输入样本中获取的特征，量化每个神经元编码的信息。基于该概念，本文量化和比较不同神经元编码的特征，从三个方面对卷积层进行了系统性地分析，并在多种模型上进行了实验。

第五章专注于滤波器级的解释。目前可解释性方法大多依赖于经验性启发，并缺乏严格的数学基础。且当前的解释方法通常使用单一度量解释算法，缺乏全面解释能力。针对这一问题，本文引入了可解释特性的概念，从不同角度对滤波器学习到的特征进行有效量化，从而提取多种可解释特性。在 CNN 模型中，神经元是滤波器的基本单元。本文基于神经网络扫描仪这一针对神经元的解释方法，通过组合神经元的特征，进一步扩展其应用于滤波器的分析，以揭示滤波器的行为与特性。在此基础上，本文提出了一个统一且具有良好扩展性的可解释性框架，称为模型可解释特性测量框架，旨在利用不同的可解释特性对模型进行多角度解释。基于该框架，本文设计了三种具有代表性的评估指标，以提供从不同视角出发的模型解释。以图像分类任务为例，本文通过实验展示了如何通过模型可解释特性测量框架探索模型特性，并结合提出的指标对模型进行多角度解释。

第六章探索了模式级的解释。现有的可解释性方法在局部或特征层面提供了一定程度的模型解释，但这类解释缺乏全局结构性，难以揭示神经网络模型在更高层次上从输入样本中所学习到的模式。针对这一问题，本文提出了一种方

法，通过神经网络扫描仪对滤波器进行扫描，并对学习图像进行聚类，进而探究聚类结果与模型过拟合之间的关系。为此，本文设计了量化指标，用以评估滤波器的特性并识别异常滤波器。本文从不同角度设计并开展实验分析异常滤波器在模型中的作用。

在第七章，本文对提出的算法进行总结，梳理当前研究成果的区别和联系，并对可解释性研究这一任务进行深入探讨，展望未来可能的研究方向。

第二章　相关工作

本章对与后续章节研究相关的神经网络基础模型架构进行介绍。本章介绍的模型架构包括全连接神经网络模型和卷积神经网络模型。随后，针对解释方法不同的关注点，本章将可解释性方法分为基于理想样本的解释、基于真实样例的解释、基于单个输入的解释和基于多个输入的解释。本章对这些神经网络的可解释性方法进行梳理和介绍，以便更好地理解后续的研究工作。

2.1　神经网络基础模型

2.1.1　全连接神经网络模型

全连接网络（Fully Connected Nerual lNetwork，FCN）是最基础且广泛应用的人工神经网络结构之一，其核心特征是每一层的神经元都与前一层的所有神经元相连接。全连接网络的每个神经元通过接收来自上一层的输入信号，并通过加权求和和激活函数的非线性变换，逐层提取数据中的特征并最终进行预测。全连接网络由输入层、隐藏层和输出层组成。输入层负责接收输入数据并将数据传递至第一层隐藏层。隐藏层通常由多个神经元组成，每个隐藏层的神经元都与上一层的所有神经元连接。隐藏层的作用是通过激活函数对输入数据进行非线性变换，从而提取更高层次的特征信息。输出层用于生成网络的最终预测结果。

在每一层中，神经元接收来自上一层的输出信号，并对这些信号进行加权处理。假设第 l 层第 i 个神经元为 a_i^l，它接收来自第 $l-1$ 层神经元的输出结果，通过加权和的方式进行计算：

$$a_i^l = f\left(\sum_j w_{ij}^l a_j^{l-1} + b_i^l\right), \tag{2-1}$$

其中，a_j^{l-1} 是第 $l-1$ 层第 j 个神经元的输出，w_{ij}^l 是 a_j^{l-1} 到 a_i^l 的权重，b_i^l 是 a_i^l 的偏置项，$f(\cdot)$ 是激活函数。激活函数的作用是引入非线性，使得网络能够表示

复杂的模式和特征。常见的激活函数包括 ReLU、Sigmoid、Tanh 等。

全连接网络是不可或缺的神经网络基础模型，尤其在处理高维特征或结构化数据时，发挥着重要作用。随着 CNN 等架构的出现，全连接网络在特定任务中的使用逐渐被局部连接结构所替代，但其依然为神经网络模型的发展奠定了重要基础。

2.1.2　卷积神经网络模型

卷积神经网络（Convolutional Neural Network，CNN）是用于处理具有网格状拓扑结构数据的深度学习模型，尤其擅长处理图像数据。与全连接网络不同，CNN 通过局部连接和权重共享机制，极大地减少了参数数量，提高了模型的计算效率和泛化能力。

CNN 由输入层、卷积层、池化层、全连接层组成。输入层接收原始数据。对于灰色图像，输入层表示为单通道矩阵。对于彩色图像，输入层为包含 RGB 三个通道的矩阵。卷积层通过卷积操作对数据进行处理。卷积操作通过滑动卷积核对输入数据进行局部感知，从而提取局部特征，生成一个特征图。对于输入数据 I 和卷积核 K，卷积操作通过滑动窗口计算其局部区域的加权和：

$$(I * K)(x, y) = \sum_{m} \sum_{n} I(x + m, y + n) K(m, n), \tag{2-2}$$

其中，符号 $*$ 表示卷积运算。m 和 n 是卷积核的坐标，x 和 y 是输出特征图的坐标。卷积操作的主要特性包括局部连接和权重共享。局部连接是指在卷积层中，每个神经元只与输入数据的一个小区域（感受野）相连接，而非与整个输入连接。这种局部连接方式使得卷积层能够专注于输入图像中的局部模式和特征，降低了计算复杂度。权重共享是指每个卷积核在不同位置上共享相同的权重，这意味着卷积核在滑动过程中，使用相同的参数进行卷积操作。通过权重共享，CNN 能够减少参数数量，提高计算效率。池化层用于对特征图进行降采样，降低维度的同时保持重要的特征。常见的池化方法有最大池化（Max Pooling）和平均池化（Average Pooling），其基本操作是在特征图上应用一个窗口，选取该窗口内的最大值或平均值。池化层不仅减少了计算量，还在一定程度上增强了模型的平移不变性。CNN 通过卷积层和池化层提取特征后，通过全连接层进行决策。全连接

层与全连接网络类似，将特征图展平成一维向量，通过加权求和和激活函数生成最终的预测结果。全连接层负责将局部特征整合起来，最终实现任务。输出层根据任务的类型输出最终结果。在分类任务中，输出层通常使用 Softmax 函数将网络的输出转化为各类别的预测概率；在回归任务中，输出层则给出预测的数值。

CNN 通过卷积层提取局部特征，池化层减少维度，并通过全连接层整合特征进行最终预测。其在图像识别、目标检测和语音识别等领域效果显著，是深度学习中的主流模型。

2.2 新型可解释性算法的分类方法

当前可解释研究的方法繁杂多样，梳理不同的可解释方法，为其进行分类是一项必不可少的工作。许多文献针对神经网络可解释问题从不同角度提出了分类方法。Guidotti et al.[69]总结了定义解释性算法的维度：全局或局部可解释性：模型可能是完全可解释的或只有单个决策是可解释的；时间限制：用户有空或被允许花在理解解释上的时间；用户专业知识的性质：使用模型的用户可能具有不同的背景知识和经验。针对黑盒模型的解释方法，Guidotti et al.[69]提出的分类方式是根据需要解决问题的类型、解释方法的类型、黑盒模型的类型、输入的数据类型等特征对解释方法进行分类。Angelov et al.[70]将解释方法分为：面向特征的解释方法、基于全局特征的解释方法、概念模型解释方法、代理模型解释方法、局部的基于像素的解释方法和以人为中心的解释方法。Bodria et al.[1]根据解释方法返回的解释类型和正在分析的数据格式提出了建议的分类。并以解释模型的忠实度、稳定性、稳健性和运行时间作为评估指标，选取一部分解释方法进行了定量比较。Linardatos et al.[71]以创建解释方法的目的以及实现此目的的方式为重点，将可解释方法概括分为四大类：解释复杂黑盒模型的方法、创建白盒模型的方法、促进公平和限制歧视存在的方法以及分析模型预测敏感性的方法。Gilpin et al.[22]同样从解释算法的目的出发，将解释算法分为三类：模拟数据处理用于在系统的输入和输出之间建立联系；用于解释网络内部数据的表示；用于解释生成网络。当前对于神经网络可解释性进行研究的综述文献，通常存在以下问题：1) 对特定问题进行分析研究，不能对可解释方法进行完备的概括；2) 对可解释方法的分类较为简单，可解释方法不能被完全归纳涵盖；3) 划分的类别

间存在交集，同一可解释方法同时属于多个类别；4) 分类的等级不能保持一致，类别间具有相互包含的关系。

针对当前可解释研究分类问题中存在的问题，本章提出一种新型可解释算法的分类方法。新方法从两个维度进行分类，每个分类中具有相互独立的子分类。本章提出的分类方法多角度多维度地分析解释算法，实现对解释算法的全面分类。同时不同分类间彼此相互独立，无重叠关系，无等级问题，可以实现清晰、快速的分类效果。

新型分类方法针对不同的关注点，从两个维度对解释方法进行分类：基于网络的解释方法和基于输入的解释方法。基于网络的解释方法关注神经网络中的各单元本身学习到的特征，基于输入的解释方法关注指定输入样本得到特定输出结果的具体原因。在新分类方法中，每个类别下具有独立的子分类。其中，根据网络单元感兴趣模式的生成方式的不同，可以将基于网络的解释方法分为理想样本和真实样例两个子类。根据神经网络解释算法的输入方式的不同，可以将基于输入的解释方法分为单一输入的解释和多个输入的解释两个子类。不同类别的具体分类方法及概念总结见表2-1 。

表 2-1　新型可解释性算法的分类方法

方法	概念
基于网络的解释	针对网络自身单元的属性进行解释，不依赖于特定的输入输出。 （1）理想样本：特定的网络单元自发生成最感兴趣的输入样本。 （2）真实样例：网络单元从输入样本中寻找感兴趣样本的代表。
基于输入的解释	针对指定输入样本，对网络给出的输出结果进行解释。 （1）单一输入的解释：对特定输入样本与输出结果的关系进行解释。 （2）多个输入的解释：为一类相似的输入样本的输出提供统一解释。

基于网络的解释方法侧重于解释网络内部各单元（如特征图、神经元等）的属性，而不依赖于输入和输出的具体内容。这类方法关注神经网络所学习到的模式，而不是在特定输入下的网络表现。由于深度神经网络通常无法像线性模型那样直接找到可解释的线性关系，因此需要采用其他方法来揭示网络单元的功能。直观的解释方法之一是通过可视化网络中指定单元（如隐藏层神经元等）最感兴趣的模式，从而推测出网络的运作机制。除此之外，还可以通过主动设计的方法来提升网络单元的可解释性。例如，高层神经元往往学习到较为复杂的混合模式，因此可以通过特定的技术将不同的模式分开，使得网络单元仅在特定的模式下被激活。这些方法使得解释算法更容易理解，例如，通过观察滤波器在动物头

部区域的激活，可以更清晰地解释网络是如何作出决策的[72]。

基于网络的解释方法可以根据网络单元感兴趣模式的生成方式分为理想样本和真实样例两类。这两种子方法的共同目标是展示网络单元最感兴趣的输入，但它们生成这些输入的方式有所不同。理想样本解释指的是，网络单元基于其学习到的感兴趣模式和激活情况，自动生成最感兴趣的输入样本，这些样本在训练数据集中并不存在。而基于真实样例的解释则是通过从训练数据集中选择出一个或一组样本，使得特定的网络单元激活最强，这些输入样本中显著包含网络单元感兴趣的模式。因此，这些输入样本可以被视为该单元感兴趣样本的代表。

基于输入的解释方法侧重于解释单一或一组输入样本的输出结果，而无需涉及神经网络的内部工作原理。这些方法的核心在于揭示输入与输出之间的关系，特别是通过特定输入样本对输出结果产生的影响。例如，显著图就是一种常用的工具，它可以用于理解为什么某个输入样本会得到特定的预测结果。具体来说，显著图揭示了网络在做出预测时，哪些区域在输入样本中最为重要，最终对预测结果产生了最大影响。如果这些关键区域能够与人类易于理解的概念（如物体的部件、动物的特征等）匹配，那么解释的效果就更加直观和有意义。

根据神经网络解释算法的输入方式，基于输入的解释方法可以进一步分为两大类：单一输入解释和多个输入解释。单一输入解释专注于单个输入样本，并通过赋予样本的不同部分（如图像的像素或特征区域）一定的权重，来揭示哪些部分对输出结果有较大影响。这类方法通常依赖于输入样本的具体信息（例如，通过计算特征的梯度或敏感度）来进行分析。相对而言，多个输入解释则是为一组相似的输入样本提供整体的解释。这类方法分析输入样本集中普遍存在的特征，评估这些特征如何或在何种程度上影响网络输出，进而为整个模型的决策过程提供解释。与单一输入解释不同，多个输入解释试图通过分析多个输入样本的共性来提炼模型的行为规律。例如，通过规则学习中的顺序覆盖方法或特征重要性排名，解释算法能够揭示哪些特征对多个样本的输出结果有着显著的贡献。

综上所述，新型分类方法可以全面对解释算法进行分类，同时不同分类间彼此相互独立，无重叠关系，可以实现清晰、快速的分类效果。为了深入理解不同类别的分类方法，表2-2根据新型分类方法的定义对各子类别进行分析，并结合样例示意图对类别进行说明。

目前，大多数神经网络解释算法都集中在单一输入的分析上。对于指定的单

一输入，解释算法通常会识别图像中的哪些区域（或像素）对最终分类结果产生了最大的影响。例如，常见的做法是利用敏感性分析，结合输入特征的信息（如每个像素的值）来计算其对模型预测的影响。而对于多个输入样本的解释，如何有效地对不同样本中的属性进行分类和分析成为一个重要问题。通常会通过像素特征进行分类，也有一些算法通过定义概念来研究不同预测结果与概念之间的关系。文献[64]中提到的一种方法是通过平面法向量来表示概念（例如“条纹”），并将其应用于网络的隐藏层空间，将有条纹和无条纹的样本区分开。基于这一方法，可以分析某个预测结果（例如斑马）对条纹这一概念的敏感度，从而为网络的决策提供进一步的解释。

2.2.1　基于理想样本的解释方法

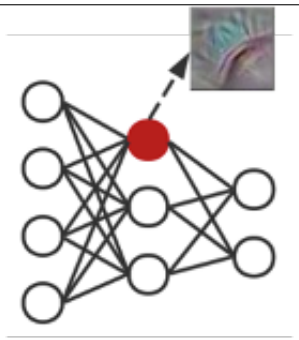
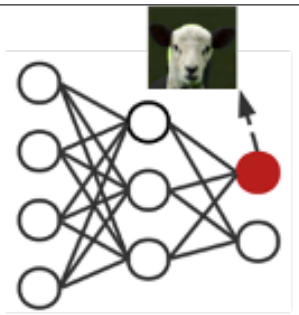
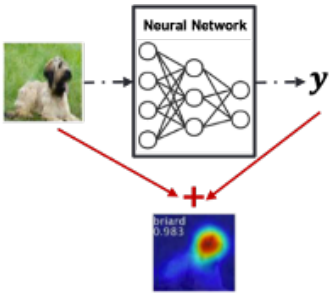
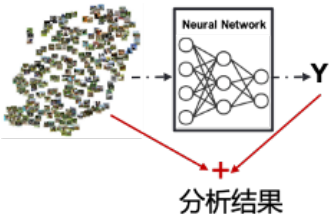
在基于理想样本的解释方法中，最典型的技术是激活最大化方法（Activation Maximization，AM）[73]。该方法通过最大化某个神经元、通道或层的激活值来找到一个代表性的输入样本。激活最大化作为一种优化技术，最初用于非监督学习模型，而文献[48]首次将其应用于 CNN，解决了图像分类任务中的深度网络可视化问题。文献[48]中介绍了一种方法，用于可视化 CNN 中指定类别的学习特征。具体来说，激活最大化方法通过对 CNN 全连接分类层中代表特定类别 c 的神经元进行优化来生成理想样本。在这一过程中，类别 c 的神经元激活值为 $S_c(I)$，其中 I 是随机初始化的输入样本。通过反向传播算法，在保持网络权重不变的情况下，输入样本 I 会被迭代优化，以最大化该类别神经元的激活值 $S_c(I)$，最终得到能够激活该类别神经元的最优输入样本。公式中的 λ 是正则化项的权重，用于防止过拟合。最终得到的可视化结果展示了网络在不同类别上学习到的特征：

$$\arg\max_{I} S_c(I) - \lambda \| I \|_2^2. \tag{2-3}$$

通过激活最大化方法，可以将这一过程推广到神经网络中任意神经元的理想样本生成。对于任意神经元 i，该方法通过优化输入样本 x^* 来最大化神经元的激活函数 $a_i(x)$：

$$x^* = \arg\max_{x} \left(a_i(x) - R_\theta(x) \right), \tag{2-4}$$

表 2-2 新型分类方法说明

分类	子类	分析	样例示意图
基于网络的解释	理想样本	旨在展示神经网络中的神经元学习到的特征，即找到能使指定神经元达到最大激活值的理想样本。示意图如图所示，生成当前选定神经元 (红色表示) 最感兴趣的样本。	
	真实样例	网络单元从输入样本中寻找一个或一组样本，使得网络单元的激活程度最高。示意图如图所示，找到使得选定神经元激活最大的输入图像。	
基于输入的解释	单一输入的解释	对特定输入进行解释，为输入的不同区域或像素分配重要度值或敏感度值解释其对输出结果的影响。示意图如图所示，解释方法结合网络给出的输出结果，对该样本的预测结果进行解释。	
	多个输入的解释	为一类相似的输入样本的输出结果提供统一的解释。示意图如图所示，结合网络的输出结果，解释方法对样本中的共性进行分析，给出解释。	

其中，$R_\theta(x)$ 代表正则化项，用于控制样本的属性。这个方法通过最大化神经元的激活值，能够有效地揭示神经网络中各个单元所学习到的特征。

研究发现，高频噪声是导致激活最大化方法效果不佳的主要原因[74-75]。为了解决这一问题，研究者们对输入样本施加约束，以使生成的样本更接近真实样本[76]。此外，为了提高可视化结果的清晰度，文献[77]对四种正则化方法进行

了分析，包括衰减、高斯模糊、小范数裁剪和小贡献裁剪，评估了它们对优化过程的影响。研究表明，不同正则化超参数的选择会显著影响生成样本的效果：一些超参数有助于突出高频信息，另一些则更有效地保留低频信息；有些超参数生成的图像包含密集的像素信息，而另一些则仅展示重要区域的稀疏轮廓。文献[76]提出了一种新的理想样本优化方法，该方法结合了图像模糊算子和去模糊算子。通过使用高斯低通滤波器进行卷积和去反卷积操作，该方法能有效去除高频噪声并恢复图像的细节。模糊操作主要用于去除噪声，而去模糊操作则帮助还原图像细节。此方法已被应用于不同网络的卷积层滤波器的可视化，证明了优化后的图像更具解释性。通过该方法，生成的感兴趣图像能够清晰地展现出网络在卷积层提取到的特征。例如，使用该算法对 VGG 网络的不同卷积层的滤波器生成感兴趣图像时，可以明显看到不同滤波器提取到的特征差异。为了进一步改善生成图像的质量，文献[78]提出了在生成过程中加入图像生成器网络的方案。该网络通过不断优化生成器输入，最大化网络中指定神经元的激活值，从而生成该神经元感兴趣的图像。通过这种方式，生成的图像更加接近真实图像，且能有效表达神经元关注的特征。

基于理想样本的解释方法的共同特点是，它们主要通过展示网络单元学习到的特征来实现解释。这些方法的目标是生成激活值最高的理想输入样本。虽然不同方法在可视化效果上存在差异，但它们通常致力于减少生成样本中的噪声，并提高样本的语义信息，使得生成的样本更易于理解。该方法的优点在于，其基本原理简单易懂，能够展示网络单元的学习特征，并且较为贴近人类理解神经网络的方式，因此易于被人类理解。然而，这类方法也存在一些局限性。首先，针对特定网络构建目标函数较为困难，且迭代优化过程可能会出现信息丢失的问题。其次，通过这种方法得到的理想样本，语义信息通常较为模糊，难以与人类的认知方式匹配，因此解释性较弱。尽管如此，这类方法在结构相对简单的网络中，往往能够提供较为有效的解释。

2.2.2　基于真实样例的解释方法

基于真实样例的解释方法通过从数据集中选择一个或多个样本，来激活目标网络单元，从而展示这些样本中包含该单元所关注的模式。这类方法的核心思想是，若某些样本能够显著激活网络中的特定单元，那么这些样本就能够代表该

单元学习到的特征。常见的相关方法包括多模态分析、特征提取和特征拆分等。

（1）多模态分析方法

多模态分析方法结合了多种技术，如图像可视化与文本说明，提供了对神经网络决策的多角度分析。为了解释 CNN 的内部工作机制，并探究训练过程中各种因素对模型的影响，文献[63]提出了网络分解框架。这一框架的主要目标是评估每个隐藏单元与一组语义概念之间的相关性，从而为网络提供解释。具体而言，网络分解方法通过为 CNN 每个卷积层中的隐藏单元打分，来评估该单元与语义信息的匹配程度。评分依据为该单元对具有语义信息的通用数据集的响应，语义信息包括物体、部件、场景、纹理、材料和颜色等。研究中使用的 Broden 数据集包含了像素级标记的视觉概念，该方法通过这些数据集中的信息为卷积单元打分。对于每个输入样本 x 和 Broden 数据集中的概念 c，会生成一个与样本相同尺度的二进制掩码 $L_c(x)$，表示该概念在样本中的存在。接着，定义 $A_k(x)$ 为样本 x 在卷积核 k 作用下的激活图，通过插值将其调整为与输入图像相同尺度的二进制掩码 $M_k(x)$，再通过交并比（IoU）来量化卷积核与概念之间的关联：

$$\mathrm{IoU}_{(k,c)} = \frac{\sum |M_k(x) \cap L_c(x)|}{\sum |M_k(x) \cup L_c(x)|}, \tag{2-5}$$

其中，$\mathrm{IoU}_{(k,c)}$ 代表卷积核 k 在检测概念 c 时的精确度。如果该值超过预设的阈值，则说明卷积核 k 对该概念具有较强的检测能力。基于这一理论，文献[79]进一步提出了一种方法，通过多个卷积核组合来共同捕捉同一概念的信息，从而增强模型的解释能力。文献[63]则通过假设网络单元的可解释性可以看作其随机线性组合的能力，探讨了不同参数初始化对网络单元语义匹配能力的影响，并通过实验验证了这一假设。

文献[80]提出了一种名为“解释基础分解”的框架，旨在为分类网络提供视觉上的解释，结合热力图和带标签的语义解释。该框架通过将输入图像中各神经元的激活值分解为预训练的语义可解释元素，来为分类任务提供清晰的解释。具体来说，类别的权重被分解为一组可解释的基向量，每个向量对应一个标签概念。然后，将这些概念的分量映射到热力图中，通过记录各个概念在热力图中的激活强度，来量化每个概念对最终预测结果的贡献。文献[80]采用该框架对常见的视觉识别网络进行了可解释性分析，并通过 AMT（Amazon Mechanical Turk）

平台对结果进行了评估，验证了解释的合理性和有效性。

（2）特征提取方法

特征提取方法旨在显性地展示网络中特定单元（对于 CNN 来说，通常是卷积核）所学习到的特征。为了增强 CNN 网络中单元表达信息的可解释性，文献[72]提出了一种方法，将传统的 CNN 结构改进为可解释的 CNN，从而揭示 CNN 高层卷积层中的知识表示。这种方法将每个高层卷积层的卷积核与一个特定的对象部分关联。在学习过程中，算法会自动为每个卷积核分配一个对象部分，使得网络学到的显式知识表示有助于理解 CNN 内部的运作方式，特别是网络是如何利用特定模式来做出预测的。传统 CNN 的高层卷积核可能学习到混合模式，导致网络的可解释性较差；而在可解释 CNN 中，每个卷积核仅会激活与特定对象部分相关的模式，从而使得网络学习到的特征在分类任务中变得更加明确和易于理解。文献[72]将该算法应用于四种不同结构的 CNN，并通过可视化可解释卷积层中的特征图，展示了不同卷积核关注的语义信息。此外，文章还通过两个指标——物体可解释性和位置不稳定性，定量评估了卷积核所提取的语义信息的准确性。对于给定的图像样本和对象类别，图像分割任务可被视为推断哪些像素属于该类别，这与特征提取方法的思路一致。因此，图像分割任务的结果可作为衡量特征提取有效性的标准。为进一步提高特征提取的精确度，文献[81]提出对 CNN 模型进行修改，使得在训练过程中，网络会受到约束，从而更加突出分类中重要像素的权重。该方法通过增强对影响分类的关键特征的关注，来改善模型的表现。文献[81]将这一新算法与传统的监督学习方法在分割任务中的结果进行了比较，证明了其在提取物体特征中的有效性。针对知识蒸馏的可解释性分析，文献[66]提出了一种方法，通过提取神经网络中间层编码的视觉概念，来解释知识蒸馏成功的原因。文章提出了关于知识蒸馏是否能帮助神经网络从原始数据中学习到更多视觉概念的三种假设，并通过实验量化网络中视觉概念的存在，从而逐一验证这些假设，最终得出结论。

（3）特征拆分方法

特征拆分方法旨在将神经网络所学习到的特征进行拆解，具体包括物体的颜色、组成部分、尺度、方向等各个方面。这些方法通过算法分析和检测神经网络所学习到的不同特征，帮助揭示网络内部如何处理各种信息。文献[82]提出，单个神经元能够检测多种类型的特征。通过对输入样本特征的分解，可以生成

不同特征，从而激活神经元，并且能够合成在这些特征下神经元感兴趣的图像。通过可视化选定神经元的激活结果，发现低层神经元并未能显著区分不同特征，而高层神经元则能够更复杂地处理特征，识别特征的不同维度。在输出层，神经元被训练来响应固定类别，学习到的特征则呈现出更大的多样性。为了深入揭示 CNN 中卷积层内编码的物体部分，文献[83]和文献[84]提出了基于图形模型的解释方法，称为解释图（Explanatory Graph）。这一方法考虑到每层中不同滤波器对图像特征的检测能力，自动从每个卷积核中提取图像的不同部分，并基于这些信息构建解释图。在解释图中，每个节点表示一个物体模式，而不同节点之间的连接则表示这些物体模式之间的协同激活关系及其空间关系。文献[83]通过解释图对四种不同结构的 CNN 模型进行了可解释性分析，并通过可视化解释结果、评估节点表示对象的一致性，以及在小样本局部定位任务中测试节点的可迁移性等实验，验证了该方法的有效性和准确性。实验结果表明，解释图中的每个节点能够表示不同输入图像中的相同物体部分，从而提升了模型的可解释性。

（4）基于真实样例的解释方法分析

基于真实样例的解释方法共同点在于：通过选择具有代表性的样本来分析和解释数据集，进而评估这些样本对模型的影响。这些样本能够帮助识别数据中的潜在偏差，使得模型在面对数据集变化时更具稳定性和适应性。此类方法尤其适用于 CNN，尤其是在数据集包含显式语义信息时。

根据具体的解释方式，基于真实样例的解释方法可分为多模态分析、特征提取和特征拆分等几种类型。多模态分析方法的优势在于它结合了图像与文本信息，从而提供了既直观又具有说服力的解释。然而，缺点在于这种方法通常要求对数据集进行严格的语义标注，因此训练过程较为复杂，并且数据集的筛选标准更为苛刻。基于特征提取的解释方法通过将网络中基本单元与输入样本的特征相对应，利用数据集中的信息来解释网络单元所学习的特征。这类方法提供了良好的可视化效果，并具备较强的可解释性。然而，它的问题在于很难对从单一网络单元提取的特征进行规律总结，且提取的特征可能缺乏足够的语义性，使得人类难以直观理解，因此难以对所有网络单元进行解释。基于特征拆分的解释方法通过将输入数据拆解为不同特征，并对这些特征进行独立解释。拆解后的特征易于理解，因此能够直观地呈现网络的学习过程。但这类方法通常面临建模复杂、计算量大的问题。

2.2.3　基于单一输入的解释方法

基于单一输入的解释方法针对特定的输入样本进行分析，主要通过为输入的不同区域或像素分配重要性值或敏感度值，来解释这些区域对最终输出结果的影响。通常，这类方法会利用输入样本的相关信息，如特征值、梯度等。常见的基于单一输入的解释方法包括类激活映射、基于梯度和反向传播的方法、模型未知方法和基于扰动的方法等。其他典型的相关研究方法将在后文章详细介绍。

（1）类激活映射方法

类激活映射方法通过生成类激活图，展示 CNN 在处理输入时感兴趣的区域。许多研究[85-87]已经证明，卷积层在 CNN 中具有显著的物体定位能力，但在使用全连接层进行分类时，这种能力会丧失。为了保持 CNN 的定位能力，文献[49]提出了类激活图（Class Activation Mapping，CAM）方法，通过调整网络结构，保持网络的定位功能。该方法通过在卷积层后加入全局平均池化层，替代原有的全连接层。通过这种结构设计，网络可以将输出层的权重投影回卷积特征图，形成类激活图，从而识别图像中各个区域的重要性。具体来说，类激活图表示模型在识别某个输入样本时，对该类别感兴趣的区域以及感兴趣程度。文章通过使用 CAM 方法修改了不同网络的结构，并评估了这些修改后网络的定位能力和分类能力。CAM 方法不仅能够实现物体定位，还能在保持定位能力的同时，保持网络的分类性能。

（2）基于梯度和反向传播的方法

基于梯度和反向传播的方法通过逐层将网络的输出结果反向传播到输入空间，生成与输入图像相同尺寸的特征图。这些特征图能够直观地展示网络在处理给定输入时对不同区域的关注程度和重要性[88]。具体而言，文献[48]提出了显著图方法，该方法为每个输入图像的像素分配一个重要度分数，从而为每个像素提供其在模型决策中的影响力量化。文献[75]通过对输入图像 I_0 和输出类别 c 之间的关系进行分析，使用泰勒展开的线性近似将类别得分 $S_c(I)$ 表示为输入图像 I 的线性函数：

$$S_c(I) \approx w_c I + b_c, \tag{2-6}$$

其中，w_c 表示类别 c 对输入图像 I_0 的梯度，b_c 是偏差项。这种线性近似能够通过梯度 $w_c = \frac{\partial S_c}{\partial I}\Big|_{I_0}$ 来生成显著图，揭示类别 c 在图像 I_0 中的关键区域。通过对不同

输入图像生成显著图，文献[75]验证了该方法的有效性，证明了其在解释网络决策过程中的重要作用。为了进一步优化显著图的质量，文献[58]提出了 SmoothGrad 方法，该方法通过在基于梯度的显著图中引入平滑操作，减少了噪声并提高了可视化效果。这一改进使得显著图更加平滑，降低了原始图像中可能出现的高频噪声，进而增强了图像分类任务中的可解释性。实验结果表明，SmoothGrad 显著图在清晰度和可读性方面相较传统方法有所提升。然而，尽管基于梯度的方法在许多任务中得到了应用，但它们往往依赖于两个假设：一是损失函数的近似采用了一阶泰勒展开，忽略了高阶项的影响；二是特征的重要性是孤立评估的，未考虑到不同特征之间的相互依赖性。文献[89]指出这些假设的局限性，并提出了一种改进的方法。该方法在泰勒展开中加入了 Hessian 项，即通过损失函数的二阶近似来增强显著图的准确性。这一改进使得显著图更加稳定，且能够减少噪声，实验表明，采用二阶近似的显著图与原图中的物体更加一致，具有更强的可解释性。在基于梯度的显著图方法中，偏差项 b 通常被忽略，而文献[90]提出了一种新的偏差反向传播算法，该方法通过从输出层开始，逐层将每一层的偏差归因到其输入节点，进而构建一个局部线性模型 $g(x) = wx + b$。该方法不仅提高了对输入图像的解释精度，还能生成互补的解释信息，增强了可理解性。

反卷积方法（Deconvolution）是通过将神经网络中间层的特征图映射回输入图像空间，用于理解模型在各层中所学到的特征。与传统的卷积方法不同，反卷积方法的核心思想是反向推断输入数据的结构，从而揭示网络在处理图像时的关注区域。文献[91]提出利用反卷积构建特征检测器，以提取图像中的低级和中级特征。与卷积模型类似，反卷积模型也包括卷积运算和池化操作，但其结构顺序与 CNN 模型正好相反，即反卷积将特征图映射回输入空间。文献[92]基于反卷积模型提出了一种方法，通过最小化每层反卷积输出和原始输入图像之间的误差来重构输入图像，进而揭示了每一层学习到的特征。该方法经过实验验证，证明了从四层分类网络中提取的特征优于其他特征学习方法。通过反卷积对每一层的特征图进行可视化，文献[55]进一步展示了网络在不同层级的特征学习：在低层，网络主要学习简单的边缘特征；在中层，网络则开始关注边缘连接的特征；而在高层，网络已经能够学习到物体的局部或整体结构。这一结果与人类对 CNN 的常识相吻合，证明了反卷积在解释模型中所起的重要作用。

然而，反卷积方法在处理池化操作时需要记录池化区域最大值的位置。为

了解决这个问题，文献[93]提出了 Guided Backprop 算法，通过结合梯度反向传播方法，提出了一种新的架构，避免了反卷积过程中常见的局限性。这一新架构提供了更清晰的可视化效果，并且在多个数据集上验证了其能够保持分类准确性的同时，提升了可解释性。与此类似，文献[94]在 CAM 基础上提出了 Guided Grad-CAM。与传统的 CAM 方法不同，Guided Grad-CAM 能够结合反向传播和特征图的梯度信息生成更精细的视觉解释。该方法不仅能够识别出感兴趣区域，还能够提供高分辨率的图像解释，突出了具体的类区分性。实验表明，Guided Grad-CAM 方法比 CAM 和 Guided Backprop 具有更好的解释能力：它能在图像分类任务中揭示模型的弱点，在弱监督的定位任务中表现良好，同时保证底层模型的忠实性。此外，该方法还能为图像字幕生成和视觉问答任务提供有效的物体定位支持。

（3）模型未知方法

顾名思义，模型未知方法将待解释的神经网络视为一个“黑盒”，无需关注模型的具体形式或参数。这类方法的主要目标是在不了解模型内部工作原理的情况下，研究模型对特定输入样本中特征的关注程度。由于神经网络通常是高度非线性的，直接解释其内部机制非常困难。因此，采用可以提供更好可解释性的线性模型来近似拟合这些非线性模型是一种常见的策略[95]。

其中，LIME（Local Interpretable Model-agnostic Explanations）是最具代表性的模型未知方法之一[67]。LIME 通过在输入样本周围采样，构建一个局部的线性模型，以解释网络对该输入样本的分类结果。其基本思想是，虽然深度神经网络本身可能是一个复杂且非线性的黑盒模型，但在局部区域内，可以通过一个线性模型来近似原始网络，从而使得该区域内的决策过程变得更加透明和可解释。LIME 中设待解释模型为 $f:\mathbb{R}^d\to\mathbb{R}$，输入样本为 x，并在其附近采样得到样本 $z\in\{0,1\}^d$（例如，对于图像任务，样本可以是图像中的超级像素块的存在与否）。通过衡量样本 z 与 x 的近似度 $\pi_x(z)$，LIME 通过最小化以下目标函数来寻找局部线性模型 g：

$$\xi(x)=\arg\min_{g\in G} L(f,g,\pi_x)+\Omega(g), \tag{2-7}$$

其中，$L(f,g,\pi_x)$ 衡量了线性模型 g 与原模型 f 之间的近似误差，$\Omega(g)$ 则是用来衡量线性模型的复杂度（即可解释性）。通过在 x 附近采样，并用线性模型 g

近似解释该样本，LIME 能够为模型的决策提供局部的、易于理解的解释。LIME 不仅可以应用于图像分类任务，还能够有效地应用于文本分类任务。文献中通过多个实验，展示了 LIME 在不同任务中的灵活性。特别是，LIME 还通过模拟用户实验评估了其在信任度相关任务中的有效性，揭示了在解释模型时，如何增加用户对模型结果的信任。

Shapley 值是博弈论中的经典概念，最早用于解决合作博弈中的份额分配问题。在模型解释领域，Shapley 值可以用来衡量不同输入特征对模型预测结果的贡献，将预测值与基准值之间的差异归因到各个特征。一般情况下，Shapley 值能够提供明确的解释，尤其在特征独立的情况下效果尤为突出。然而，当特征之间存在相互依赖时，独立性假设的违背可能导致解释效果下降。为了克服这一问题，文献[96]引入了“因果 Shapley 值”，它不仅考虑了特征对预测结果的直接贡献，还能够分解并分析特征间的因果关系。文章展示了如何为一般因果图计算因果 Shapley 值，并通过实践证明，即使在仅有部分信息的情况下，该方法也能有效地进行因果推理和分析。然而，精确计算 Shapley 值的计算量非常庞大，尤其在输入特征数量较多时，计算复杂度会呈指数级增长。为此，文献[97]提出了一种高效的近似计算方法，通过多项式时间来逼近 Shapley 值，显著降低了计算成本，并且在效率和准确性上优于传统方法。此外，文献[98]则将这一模型未知方法应用于反事实推理领域。该方法通过使用编码器或基于 k-d 树获得的类原型，将扰动引导到可解释的反事实区域。通过这种方式，算法不仅在图像和表格数据集上证明了其有效性，还有效地解决了黑盒模型中数值梯度计算的计算瓶颈问题。

（4）基于扰动的方法

基于扰动的方法通过对输入样本施加扰动，进而观察这些扰动对模型输出的影响，从而揭示输入样本的哪些部分对预测结果有决定性贡献。具体而言，通过研究输入样本中的不同区域，判断哪些区域对于模型的输出值影响较大。在基于扰动的方法中，一个核心问题是如何选择扰动方式，或者如何定义哪些变体可以有效地用于分析[99]。通常会通过观察扰动对模型输出的影响来分析输入样本 x_0 的不同区域。为了解决这一问题，文献[100]提出了一种新方法，该方法使用三种不同的扰动方式——恒定值扰动、噪声扰动和模糊扰动，来生成有意义的扰动图像。这些方法可以帮助研究者更好地理解哪些部分对输出结果最为关键。设定掩码 $m:\Lambda\to[0,1]$，其中每个像素 $u\in\Lambda$ 与一个标量值 $m(u)$ 相关联。根据

扰动类型，扰动的方式可以定义为：恒定值扰动：用颜色均值 μ_0 来替代被扰动的区域。噪声扰动：使用高斯噪声 $\eta(u)$ 对区域进行扰动。模糊扰动：对区域进行高斯模糊，标准差为 σ_0。扰动公式表示为：

$$[\Phi(x_0;m)](u)=\begin{cases} m(u)x_0(u)+(1-m(u))\mu_0, & \text{constant} \\ m(u)x_0(u)+(1-m(u))\eta(u), & \text{noise} \\ \int g_{\sigma_0 m(u)}(v-u)x_0(v)dv, & \text{blur} \end{cases}. \tag{2-8}$$

为了得到最优掩码 m，定义优化函数如下：

$$m^*=\arg\min_{m\in[0,1]^\Lambda}\lambda\|1-m\|_1+f_c(\Phi(x_0;m)). \tag{2-9}$$

目标是通过最优化掩码 m，使得扰动后的样本在类别 c 上的准确度显著降低，即：$f_c(\Phi(x_0;m))\ll f_c(x_0)$。这一结果表明，掩码 m 遮挡了输入样本中对于类别判定最为重要的区域。文章展示了这一方法在图像分类任务中的应用，研究者通过移除最小掩码区域，阻止网络识别输入样本，从而能够有效地解释网络的决策过程。

另一类基于扰动的方法则侧重于使用生成模型来生成输入样本的扰动。这类扰动与传统的扰动方法相比，更加自然且不容易被人眼察觉。尽管如此，CNN 仍然容易受到这些几乎不可见的、像素级扰动生成的对抗样本的影响[101]。为了应对这一问题，文献[102]结合了对抗性防御技术，对文献[100]的方法进行了扩展，提出在计算扰动时，通过删除不相关或者最相关的像素，构建出一种更加细粒度的视觉解释图像。文献[102]通过对不同网络中的新方法与其他已有方法（如 BBMP[100]、Gradient[48]、Guided Backprop[93]、Contrastive Excitation Backprop[103]、Grad-CAM[94]、Occlusion[55]等）进行定性和定量的比较，证明该算法能够提供更细致的解释。通过这些比较，文章展示了该方法在生成更加精细化的扰动解释方面的优势。此外，文献[104]深入探讨了像素级扰动对 CNN 的影响，尤其是在对抗样本中的扰动如何影响图像分类的敏感性和功能。文献将扰动分为三类：抑制型扰动、促进型扰动和平衡型扰动。这些不同类型的扰动揭示了对抗性扰动的促进-抑制效应（PSE），并进一步探讨了这些扰动如何与类激活图（CAM）

的图像级可解释性相关联。通过使用特定类别的判别图像区域，文章解释了对抗样本中的模式。此外，文章还采用网络分解[63]方法来分析对抗效应，进而提供了对隐藏单元的概念级可解释性。

（5）其他方法

为了实现神经网络的分层解释，文献[105]通过 ACD（Agglomerative Contextual Decomposition）对给定输入样本及其预测结果进行分析，生成输入特征的分层聚类，并计算每个聚类对最终预测结果的贡献。ACD 方法能够揭示不同特征在决策过程中的重要性，有助于深入理解网络如何做出预测。文献中的实验表明，ACD 不仅能够有效诊断错误预测，还能识别数据集中的偏差，并从中提取不同长度的短语。通过这些实验，研究者证明 ACD 使得用户能够更准确地判断两个网络中的表现，进而提升对模型输出结果的信任度。文献[106]从信息论的视角对模型进行了解释，提出了一种基于实例特征选择器的解释方法。该方法通过优化特征选择器来选取输入中最有用的特征子集，最大化所选特征与函数响应之间的互信息。通过对特征子集进行条件分布的分析，该方法可以有效解释模型的决策过程。文献中的实验分别在合成数据集和真实数据集上进行了验证，并且应用于不同的网络模型中。通过对运行时间和可解释性等因素的定量分析，研究证实了该算法的有效性和实用性。

（6）基于单一输入的解释方法分析

基于单一输入的解释方法共同的特点是，它们通过提供单个预测结果的详细解释，揭示了网络决策中最关键的区域或像素。这些方法能够可视化与模型输出最相关的部分，从而提升模型的透明度和可信任度。根据解释策略的不同，基于单一输入的解释方法可分为类激活映射方法、基于梯度和反向传播的方法、模型未知方法和基于扰动的方法等。类激活映射方法的最大优势在于其具有良好的定位能力，能精确标记出 CNN 中某一类别对输入样本的关注区域，并可量化不同区域的重要性。然而，这种方法需要调整原始网络结构，因此它的解释结果往往缺乏细节，只能提供较为粗略的解释。类激活映射法特别适用于 CNN 的分类任务。基于梯度和反向传播的方法可以为网络提供更精细的解释，具有广泛的适应性，适用于不同类型的任务和网络结构。然而，它们的计算复杂度较高，且可能会在反向传播过程中发生梯度消失，导致解释失效。这些方法虽然提供了更多的细节，但也带来了较高的计算负担。模型未知方法的最大优点是原理简单

且不依赖于具体的模型架构，因而可以广泛应用于不同类型的网络。它们通过构建可解释的局部模型来提供直观的解释。然而，这类方法的主要缺点是计算过程较为耗时，并且生成的解释往往需要进一步的人工分析来确保其准确性。基于扰动的方法通过对输入样本的特定区域进行扰动，观察网络输出的变化来评估这些区域对预测结果的贡献。其直观性强，易于理解，并能清晰地指出哪些区域对模型决策至关重要。然而，这类方法的效果很大程度上依赖于网络的精度。在高精度的网络中，扰动带来的变化较为显著，能够有效揭示网络的决策过程；但在精度较低的网络中，扰动对输出结果的影响较小，解释效果也相应减弱。

2.2.4 基于多个输入的解释方法

基于多个输入的解释方法侧重于为一组相似的输入样本提供统一的解释，目标是揭示在输入样本中普遍存在的特征，并分析每个特征对网络预测结果的具体贡献。这些方法能够用于理解模型如何从全局上进行决策。一种直接的方式是通过对数据集进行分析，评估单个属性对预测结果的贡献。文献[107]提出的SpRAy （Spectral Relevance Analysis）算法旨在评估模型是否学会了真正有意义的特征，而非仅仅在不同样本和分类结果之间找到无意义的关联。通过计算输入样本与目标类别的相关图，并结合特征值进行聚类，算法能够识别出数据中的不同预测策略。例如，在图像分类任务中，算法通过分析“马”这一类别的图像，验证了模型是否学会了与该类别相关的正确特征。

另一种常见的基于多个输入的解释方法则通过结合多个单一输入样本的解释，提炼出一种全局规律。MAME（Model Agnostic Multilevel Explanations）[68]就是这样一种方法，它通过将已有的局部解释技术（如 LIME[67]）与多层次的解释树结合，构建了一个由局部到全局的解释框架。MAME 的目标是建立局部解释与全局解释之间的桥梁，并从底层逐步构建起更高层次的解释。通过对不同样本进行聚合，算法最终形成了一个统一的解释模型。文献[108]提出的 GIRP（Global Interpretation via Recursive Partitioning）方法则通过递归划分输入变量空间，构建决策规则树，用以解释模型的全局行为。该方法基于多样本的局部解释结果，分析不同输入变量对预测的贡献，从而有效地揭示了模型决策背后的规则。

基于“概念”这一人类易于理解的单元来解释神经网络的决策过程，提供了一种更具直观性的方法。文献[64]中的 TCAV（Testing with Concept Activation

Vectors）算法采用概念激活向量（CAVs）代替传统的输入特征，来衡量特定概念对模型预测的影响。通过收集包含或不包含某一目标概念（例如，有条纹或没有条纹的动物图像）的正负样本，算法计算出一个用于分隔这些样本的超平面分割向量。这个向量反映了概念在输入样本中的稳定性和重要性。文献[109]进一步扩展了这一方法，采用对输入样本进行聚类的方式自动发现概念，从而识别适用于整个数据集的概念，更全面地解释网络的决策机制。文献[65]提出了一种基于概念的解释方法，侧重于通过对输入样本中的局部特征进行聚类，形成可供人类理解的概念，并计算这些概念对网络预测结果的重要程度，从而为网络行为提供更具意义的解释。

基于多个输入的解释方法为理解神经网络的行为提供了更为全面的视角。通过分析全局特征或识别具有重要语义的区域，这些方法能够揭示模型学习到的决策模式和预测策略。虽然这些方法通常计算量较大，但它们能有效地理清模型的全局运行机制。这些方法不仅能够提供对个别输入样本的局部解释，还能够展示整个模型的行为逻辑，从而对模型的内部机制进行深刻剖析。

2.3 本章小结

本章介绍了神经网络的经典架构的基础知识，包括全连接网络模型、CNN模型等。同时，本章对神经网络的可解释性方法进行梳理和分类，针对解释方法不同的关注点，将可解释性方法分为基于理想样本的解释、基于真实样例的解释、基于单个输入的解释和基于多个输入的解释。本章的介绍是后续章节研究内容的基础。

第三章　神经网络扫描仪（NNS）

3.1　引言

可视化单个神经元学习到的特征是理解神经网络工作机制的重要方式。CNN模型的一些模块具有定位物体的能力，例如卷积层和池化层，而另一些则不具备定位能力，例如全连接层[49][110]。当前方法可以可视化具有定位能力的模块所学习到的特征[20]。然而，很难了解那些不具备定位能力的模块所学习到的信息，这使得比较不同类型模块的输出结果变得困难。

许多神经网络可解释性的研究通过可视化来解决这一问题[34]。一种直观的方法是直接可视化神经网络的值。目前已经提出了多种方法来展示神经网络中神经元的激活值[20]。Yosinski et al.[77]提出了一种工具，可以可视化 CNN 中每个特征图上的激活情况，激活值会根据输入样本动态变化。Wang et al.[111]提出了一种 CNN 可视化工具，称为 CNN Explainer。CNN Explainer 集成了一个模型，总结了 CNN 的结构，并动态展示了 CNN 各个模块的操作。该工具帮助用户理解每一层的作用以及相邻两层之间的相互作用。还有一种名为 3D 多层神经网络仿真工具[1]，可以生成一个 3D 形式的 FCN 可视化结果。这些工具的特点在于简单且易于理解。然而，这些工具缺乏深入解释，无法从中获得每个神经元所学习到的信息。

为了更深入地解释神经网络，一些研究通过可视化展示模型的关注点[112]。目前，已有两大类现有方法用于通过可视化解释神经网络。一类方法是激活最大化[48]，它试图找到能最大激活指定神经元的输入模式。该方法通过最大化某个输出神经元的激活值来生成样本。该样本展示了某个神经元感兴趣的内容。通过反向传播，这个样本会被迭代优化，最终得到具有最大激活值的分类神经元样本。Nguyen et al.[78]在此基础上添加一个图像生成网络，以合成更接近原始图

1　3D　多层神经网络仿真工具可以在https://tutorials.retopall.com/index.php/2019/02/16/3d-multilayer-neural-network-simulation访问

像的图像。为了避免高频噪声，Wang et al.[76]应用图像模糊和去模糊技术来生成图像。另一类解释神经网络的方法是显著图[48]。对于给定的输入样本，显著图尝试为每个像素分配一个重要性分数。类别激活映射方法生成一个图，显示CNN 中感兴趣的区域。Zhou et al.[49]提出了类别激活映射方法，通过使用全局平均池化代替 CNN 中的全连接层。通过这种结构，输出层的权重被投影回卷积特征图，从而识别图像中的重要区域。基于[49]的基本结构，Selvaraju et al.[52]提出了梯度加权类别激活映射，以突出显示图像中预测概念的关键区域。Muhammad et al.[54]利用卷积层学习到的表示的主成分来构建可视化解释。Zeiler et al.[55]通过去卷积模型解释 CNN，发现神经网络在低层学到的是简单的边缘特征，在高层学到的是一些或所有的物体。Ren et al.[56]提出一种不使用非盲反卷积的方法来生成具有优秀视觉效果的图像。Zhang et al.[57]提出通过提取图形模型来解释CNN。该方法自动提取图像的不同部分，以描述滤波器感兴趣的内容。

激活最大化方法专注于神经元学到的视觉模式。这些方法不适合解释单一输入样本的结果。显著图方法显示了输入样本的哪些区域对网络而言是重要的。这种方法与固定的神经网络模型密切相关，很难比较不同神经网络结构的工作机制，也难以分析模型中不同模块的功能。

本章提出一种名为神经网络扫描仪（*Neural Network Scanner, NNS*）的方法。该方法用于可视化神经网络中不同结构中神经元的学习结果。通过灵活地组合单个神经元学习到的特征，可以分析神经网络的不同模块。不同模块的结果以统一的方式进行可视化，从而比较不同模快的工作机制。在医学领域，研究人员使用外部设备（例如功能性磁共振成像（functional Magnetic Resonance Imaging, fMRI））来捕捉脑部图像并确定大脑神经元的激活情况。当人脑接收到信息后，fMRI 捕捉到由神经元活动引起的血流变化。正如 fMRI 可以扫描人类大脑，本章提出的 NNS 方法用于扫描神经网络模型。给定一个模型和一个输入样本，NNS 获取神经网络中每个神经元学习到的特征。从单个神经元学习到的特征分析对解释神经网络模型有用的信息。fMRI 和 NNS 的工作示意图如图3-1所示。

本章的贡献如下：

- 提出的 NNS 可以可视化神经元的学习过程，展示每个神经元学习到的特征，并以人类易于理解的形式呈现。在工作过程中，NNS 始终保持物体定位能力。

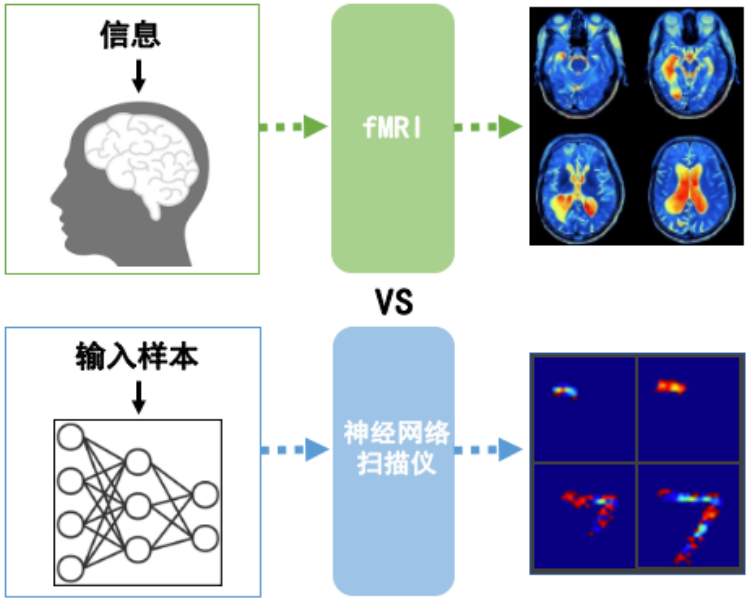


图 3-1 神经网络扫描仪与功能性磁共振成像（fMRI）的工作示意图

- NNS 是一种基于神经元的解释方法。通过灵活地结合单个神经元学习到的特征，可以分析神经网络不同模块的工作机制。不同模块的结果以统一方式可视化，NNS 适用于由人工神经元组成的神经网络，而无需改变其结构。本章对不同的神经网络结构进行了实验，展示了该方法的结果，并讨论了不同结构的工作机制差异。
- 本章进行了多项实验以评估 CNN 的可解释性。通过评估某一层获取的特征来验证该方法的有效性。随后评估了高度激活神经元的可解释性。最后将 NNS 的可视化结果与现有其他可视化解释方法进行了比较。

3.2 神经网络扫描仪方法

本节提出 NNS 方法，用于扫描神经网络并可视化模型中的每个神经元。对于一个输入样本，图像中物体的位置信息与数值信息同样重要。在全连接层中，仅使用数值信息，而位置信息被丢弃，这使得全连接层缺乏定位物体的能力。而在 NNS 中，位置信息得以保留。为每个神经元生成一个与输入样本分辨率相同的图像，称为该神经元的＂学习图像＂。神经元的学习图像是该神经元从输入样本中学到的特征的表示。不同神经元学习图像的生成过程如图3-2所示。

NNS 的工作流程如下：

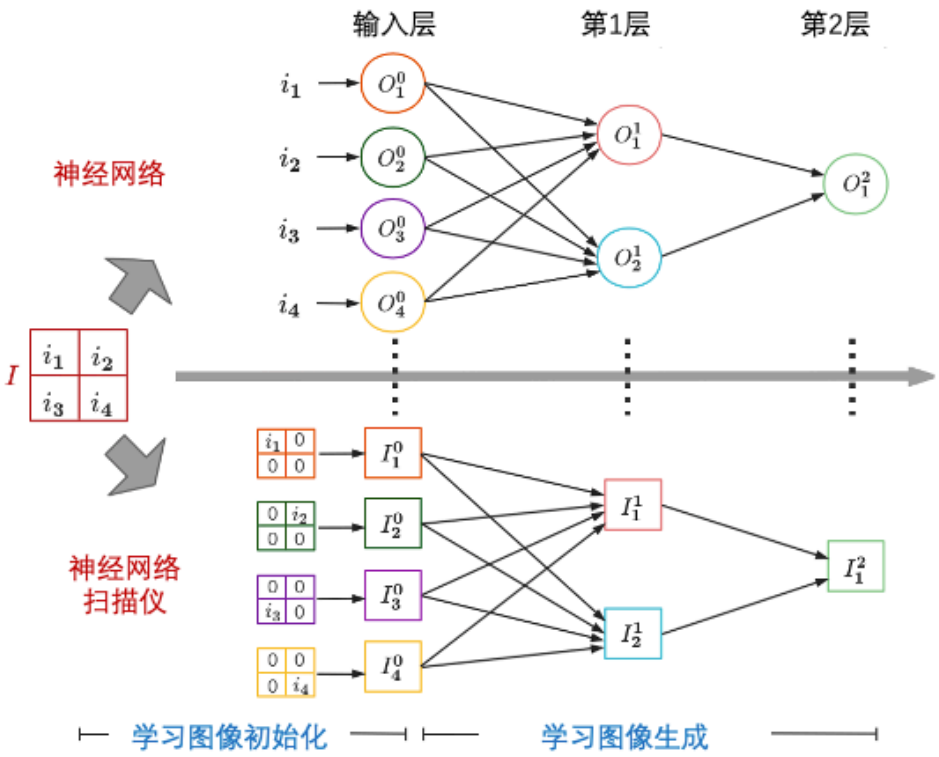


图 3-2　NNS 的工作流程

- 学习图像初始化。为输入层中的每个神经元基于输入样本构建学习图像。将输入图像分割成若干个小图像，每个分割后的图像保留输入样本的一部分特征，包括数值信息和位置信息。然后，将这些分割图像依次分配给输入神经元。
- 学习图像生成。为所有隐藏层和输出层的神经元生成学习图像。通过获取输入神经元对指定神经元的贡献，基于输入神经元的学习图像生成该神经元的学习图像。在整个过程中，NNS 始终保留物体定位的能力。

3.2.1　学习图像初始化

在 NNS 中，输入学习图像使用了输入图像中每个像素的数值信息和位置信息。如图3-3所示，为输入层中的神经元构建学习图像。

输入样本分割

首先，将输入样本视为向量 $I = [i_1\ i_2\ ...\ i_N]^T$，其中输入样本 I 的维度为 N，

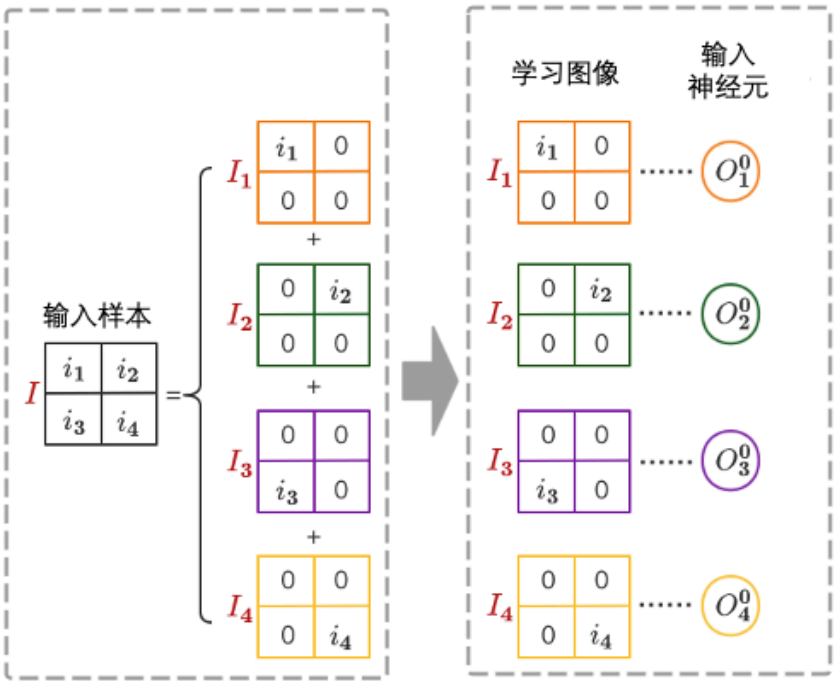


图 3-3 学习图像初始化过程

图3-3中设置 N 为 4。将输入样本 I 分割为 N 个图像 $I_1, I_2, ..., I_p, ..., I_N$：

$$I = \sum_{p \in [1,N]} I_p. \tag{3-1}$$

其中，第 p 个分割图像 I_p 是一个像素值全为 0 的图像，除了第 p 个像素。第 p 个像素的值等于输入样本 I 中的第 p 个像素值。第 p 个分割图像的表示如下：

$$I_p(x) = \begin{cases} i_p, & \text{if } x == p \\ 0, & \text{otherwise} \end{cases}, \tag{3-2}$$

其中 x 表示图像 I_p 中的像素编号，$x \in [1, N]$。每个分割图像包含了原始输入样本的一部分特征，并保留了对应特征的位置信息。

学习图像分配

NNS 为模型中的每个神经元计算一个学习图像。给定输入图像 I，将分割后的图像分配给输入层中的神经元。对于第 p 个输入神经元 O_p^0，其值为 x_p^0，其学习图像为 I_p^0。将分割图像依次分配给输入神经元。因此，第 p 个神经元 O_p^0 与

第 p 个分割图像 I_p 相关联：

$$I_p^0 = I_p, \qquad p \in [1, N]. \tag{3-3}$$

如果输入样本是 RGB 图像，则将该样本视为每个通道的独立图像。每个通道中的图像会被分割，得到 $3 \times N$ 个分割图像。它们会依次分配给输入神经元作为学习图像。

上述分析展示了学习图像初始化步骤，具体步骤见算法 3.1。

算法 3.1 学习图像初始化

输入： 输入样本 $I = [i_1\ i_2\ ...\ i_N]^T$
参数设置： $p = 1$，$I_p^0 \in \mathbb{R}^N$ 是输入层第 pth 神经元 O_p^0 的学习图像
repeat
　设置 $I_p \in \mathbb{R}^N$, $I_p = [0\ 0\ ...\ 0]^T$
　$I_p(p) = i_p$
　$I_p^0 = I_p$
　$p = p + 1$
until $p = N$
输出： 该层所有学习图像 $I_{sum}^0 = (I_1^0, I_2^0, ..., I_N^0)$

3.2.2　全连接层学习图像生成

本节为全连接层中的神经元生成学习图像。计算输入神经元对指定神经元的贡献，然后根据输入神经元的学习图像生成该神经元的学习图像。生成学习图像的过程如图3-4所示。

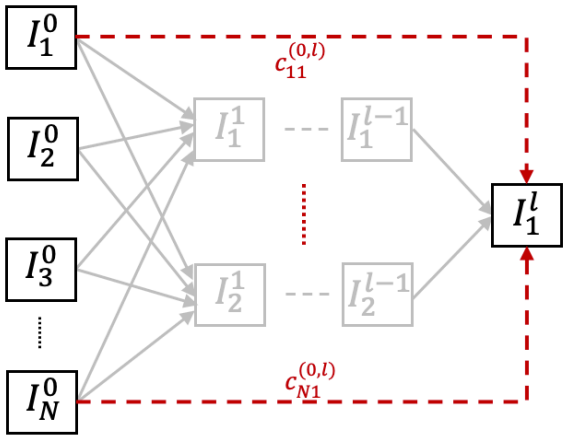


图 3-4　学习图像生成过程

指定神经元的学习图像生成

在计算第二层神经元的学习图像时，将权重值视为贡献值。以第二层的第 q 个神经元为例。

设置 $c_{pq}^{(0,1)}$ 为输入层第 p 个学习图像对第二层第 q 个学习图像的贡献值。贡献值 $c_{pq}^{(0,1)}$ 等于第 p 个神经元和第 q 个神经元之间的权重值。学习图像通过线性变换变为 $I' = [i_1'\ i_2'\ ...\ i_N']^T$：

$$I' = \sum_{p\in[1,N]} c_{pq}^{(0,1)} \times I_p^0. \tag{3-4}$$

I' 的第 t 个像素值计算如下：

$$i_t' = \sum_{p\in[1,N]} c_{pq}^{(0,1)} \times i_{pt}^0, \tag{3-5}$$

其中 i_{pt}^0 是 I_p^0 的第 t 个像素值。通过线性变换，神经元学到的特征与位置信息一同保存在 NNS 中。

然后，神经元通过非线性激活函数 f 进行非线性变换。本节设定 ReLU 作为激活函数。第二层第 q 个学习图像 I_q^1 定义如下：

$$I_q^1 = \begin{cases} I', & \text{if } x' > 0 \\ \mathbf{0}, & \text{otherwise} \end{cases}, \tag{3-6}$$

其中 x' 为神经元的值。如果 x' 大于 0，神经元被激活，对应的学习图像也会被激活。如果神经元经过 f 后值为 0，则该神经元未被激活，其对下一层神经元的贡献为零。因此，该神经元的学习图像为 $\mathbf{0}$，不会对下一层的学习图像产生贡献。

指定神经元的学习图像生成的算法见算法 3.2。

一般情况下的学习图像生成

到目前为止已经介绍了第二层学习图像的生成过程。通过获取输入学习图像对该学习图像的贡献从而习得其他层的学习图像。

算法 3.2 指定神经元的学习图像生成

输入：输入层神经元 p 的学习图像 I_p^0，层 1 神经元 q 的值 x_q^1，神经元 p 和神经元 q 之间的权重 w_{pq}，输入神经元的数量为 m，输出神经元的数量为 n

参数设置：第 p 个学习图像对第 q 个学习图像的贡献值 $c_{pq}^{(0,1)}$=w_{pq}，且 $p=1$, $q=1$

repeat
  $I' \in \mathbb{R}^N$
  repeat
    $I' = I' + c_{pq}^{(0,1)} \times I_p^0$
    $p = p + 1$
  until $p = m$
  if $x_q^1 > 0$ **then**
    $I_q^1 = I'$
  else
    $I_q^1 = 0$
  end if
until $q = n$

输出：该层所有学习图像 $I_{sum}^1 = (I_1^1, I_2^1, ..., I_n^1)$

在 FCN 中，第 $l-1$ 层和第 l 层之间的权重矩阵为 $W^l \in \mathbb{R}^{n \times m}$：

$$W^l = \begin{bmatrix} w_{11}^l & w_{12}^l & \cdots & w_{1m}^l \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1}^l & w_{n2}^l & \cdots & w_{nm}^l \end{bmatrix}, \tag{3-7}$$

其中 n 是第 l 层的神经元个数，m 是第 $l-1$ 层的神经元个数。第 $l-1$ 层的学习图像对第 l 层学习图像的贡献矩阵为 $C^{(l-1,l)} \in \mathbb{R}^{n \times m}$：

$$C^{(l-1,l)} = \begin{bmatrix} c_{11}^{(l-1,l)} & c_{12}^{(l-1,l)} & \cdots & c_{1m}^{(l-1,l)} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1}^{(l-1,l)} & c_{n2}^{(l-1,l)} & \cdots & c_{nm}^{(l-1,l)} \end{bmatrix}. \tag{3-8}$$

为了实现非线性变换，使用分段线性函数 S，如下所示：

$$S(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{3-9}$$

第 p 个学习图像在第 $l-1$ 层对第 q 个学习图像在第 l 层的贡献值 $c_{pq}^{(l-1,l)}$ 计算如

下：

$$c_{pq}^{(l-1,l)} = S(x_q^l)w_{pq}^l, \tag{3-10}$$

其中 x_q^l 是第 l 层第 q 个神经元的值。输入学习图像对第 q 个学习图像 I_q^l 在第 l 层的贡献矩阵表示为 $C_q^{(0,l)}$：

$$C_q^{(0,l)} = W_q^l \cdot \prod_{i=1}^{l-1} C^{(i-1,i)}, \tag{3-11}$$

其中 $W_q^l \in \mathbb{R}^{1\times m}$ 是第 $l-1$ 层和第 l 层第 q 个神经元之间的权重矩阵。

所有输入学习图像表示为矩阵 $I_{sum}^0 \in \mathbb{R}^{N\times N}$：

$$\begin{aligned} I_{sum}^0 &= (I_1^0, I_2^0, ..., I_N^0) \\ &= \begin{bmatrix} i_{11}^0 & i_{21}^0 & \cdots & i_{N1}^0 \\ \vdots & \vdots & \ddots & \vdots \\ i_{1N}^0 & i_{2N}^0 & \cdots & i_{NN}^0 \end{bmatrix}, \end{aligned} \tag{3-12}$$

其中 N 是输入学习图像的个数。第 q 个学习图像在第 l 层为 I_q^l：

$$I_q^l = I_{sum}^0 \cdot C_q^{(0,l)}. \tag{3-13}$$

至此可以得到所有神经元的学习图像。

3.2.3 卷积层学习图像生成

在全连接层中，输出神经元通过权重矩阵对输入神经元进行线性变换。以人工神经元为基本单元，各种类型的神经网络层可以转化为全连接层。以 CNN 为例，卷积层与全连接层之间的等价关系如图 3-5 所示。卷积层的输出是特征图，特征图中的神经元对应于相同的卷积核，即这些神经元共享相同的权重。卷积核的数值矩阵可以转化为两层之间的权重矩阵。卷积操作可以视为线性操作。在池化层中，特征图中的神经元被筛选出来，这一过程也可以视为一种特殊的卷积过程。因此，池化层也可以价为全连接层。将卷积层和池化层中的每个单元（像素）视为一个神经元，并为每个神经元构建学习图像。随着神经元值的变化，神

经元的学习图像也发生变化。在池化层中，从前一层筛选出的神经元值不发生变化，这些被选中的神经元的学习图像从前一层收集并保持不变。

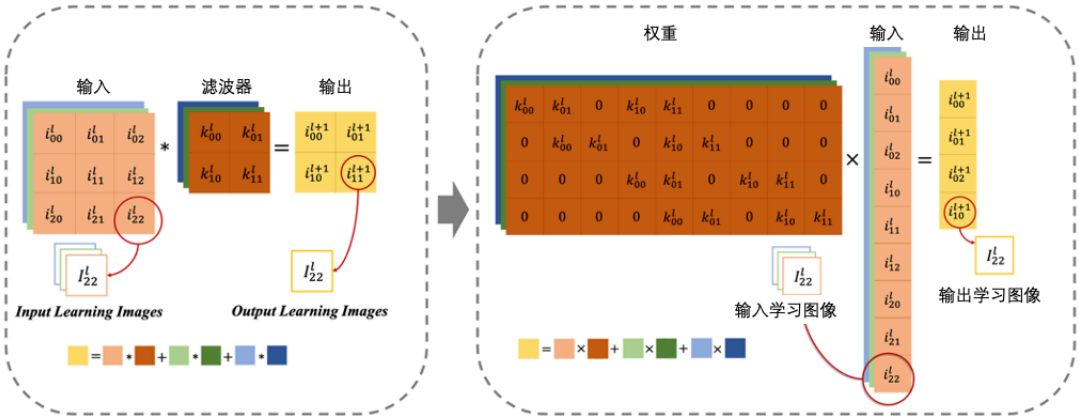


图 3-5　卷积运算与线性运算的等价关系

输入图像的大小为 $M \times M \times D$。首先，将输入图像划分为 $M \times M$ 个图像。对于一个单元 (x, y)，其对应的分割图像为 $I_{(x,y)} \in \mathbb{R}^{M \times M \times D}$。在每个通道中，分割图像的生成过程参见第 3.2.1 节。然后，根据第 3.2.1 节所述，将分割图像分配给输入层中的每个神经元：

$$I^0_{(x,y)} = I_{(x,y)}, \qquad x, y \in [1, M]. \tag{3-14}$$

如图 3-5 所示，将卷积和池化操作转化为线性操作，然后通过与全连接层计算学习图像相同的方式，获得卷积层和池化层中的学习图像。在整个过程中，学习图像的大小为 $M \times M \times d$，其中 d 是对应神经元的通道数。通过这种方式，可以获得 CNN 中神经元的学习图像。

3.3　不同模块扫描实验

在 CNN 中，将特征图中的每一个单元视为一个神经元。为了清晰地区分图像中不同像素的强度，采用热力图的形式来表示图像。热力图使用不同的颜色来表示不同的像素强度。热力图的颜色映射如图 3-6 所示。图像中的像素值越大，像素的颜色越接近“红色”；像素值越小，像素的颜色越接近“蓝色”。为了展示方法的适用性，将该方法应用于一组具有不同神经网络结构的模型。神经网络的

详细描述如表 3-1 所示。首先，训练了具有和不具有跳跃连接的神经网络，并在 MNIST 数据集上进行实验。随后，将模型应用于 ILSVRC-2013 DET、CIFAR-10 和 CIFAR-100 数据集的训练。所有模型均使用反向传播（BP）学习算法进行训练。

图 3-6 热力图的颜色映射

表 3-1 模型详细信息

模型结构	数据集	学习算法
FCN CNN	MNIST	BP
FCN with skip connections ResNet10	MNIST	BP
AlexNet VGG-16	ILSVRC-2013 DET CIFAR-10, CIFAR-100	BP

为了分析不同结构模型的操作机制差异，对一个 FCN 模型和一个简单的 CNN 模型进行了实验。这里使用的 FCN 模型包含一个输入层（784 个神经元）、两个隐藏层（每个隐藏层 500 个神经元）和一个输出层（10 个神经元）。简单的 CNN 由三个卷积层（10 个滤波器）、两个池化层、一个展平层、一个全连接层（64 个神经元）和一个输出层（10 个神经元）组成。选择一个简单的数据集，以确保不同模型的训练效果良好。这两个模型均在 MNIST 手写数字训练集上进行训练。

在以下实验中，首先，验证了所提出方法在不同神经网络上的有效性，并分析了它们的操作机制。其次明确了不同模型的神经元学习规则。接下来探讨了获胜神经元学习到的特征的性质。最后比较了具有跳跃连接的不同结构之间的学习过程。

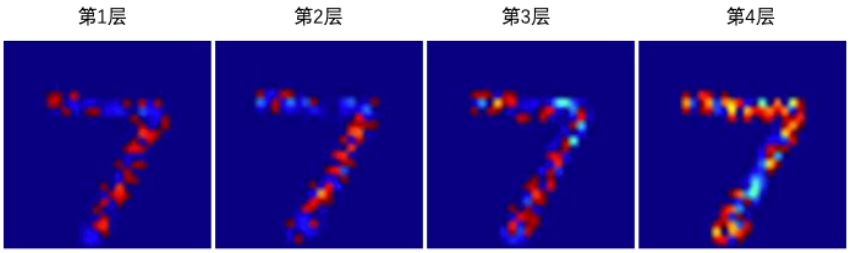
3.3.1 运行机制分析

本节比较了不同模型的操作机制。通过可视化训练后的神经网络中的所有神经元，可以深入了解模型的运行机制。给定相同的输入样本，在每一层选择一

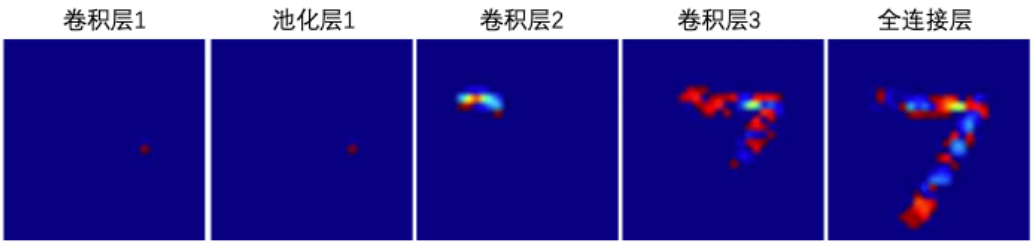
个激活的神经元，展示其学习图像。来自每一层的扫描结果展示了神经网络学习的分层特性。

在 FCN 中，从图 3-7a 中可以观察到，网络在每一层始终学习到全局特征。对于每个神经元，可以看到输入样本的边缘轮廓。在第一层，FCN 学习到的特征仅通过学习图像中的两种颜色表示。在第三层，学习到的特征则通过多种颜色表示。低层学习到的特征较为单一，学习图像中的颜色变化较小；而高层学习到的特征则更为丰富，学习图像中的颜色变化较大。

在 CNN 中，观察到网络始终在每一层学习到丰富的特征，并且学习图像中的颜色变化较大。如图 3-7b 所示，在低层（第一层和第二层卷积层），只有少数神经元被激活。低层仅对图像中的小部分特征做出响应。第三层卷积层学习到的特征覆盖了更大的区域。池化层的神经元从前一卷积层的神经元中选择出来，其值保持不变，因此这些选中神经元的学习图像保持一致。CNN 中的全连接层同样学习到了全局特征。



(a) FCN



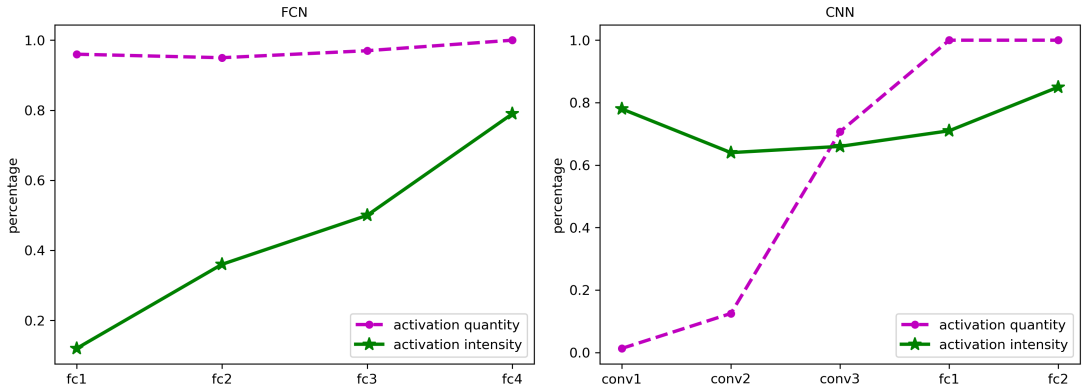
(b) CNN

图 3-7　神经元学习过程可视化

通过测量每一层神经元所学习的知识量来分析学习图像。知识量从两个方面进行评估：激活数量和激活强度。对于一个学习图像，激活数量是指激活像素的数量，激活强度是指像素强度的总和。为了便于比较，一层的激活数量被定义为该层所有神经元的平均激活数量与输入图像激活数量的百分比。同样，激活强

度被定义为该层所有神经元的平均激活强度与输入图像激活强度的百分比。激活数量和激活强度越大，神经网络从输入样本中学习到的知识越多。

图 3-8 展示了 FCN 和 CNN 的知识量。正如图 3-8a 所示，在 FCN 中，每一层的激活数量始终保持在非常高的水平。激活强度随着层数的增加显著增加。如图 3-8b 所示，卷积层的激活数量随着层数的增加而增加。与其他卷积层的激活强度相比，第一层卷积层的激活强度最大。其他卷积层的激活强度变化不大。在 CNN 中，全连接层的知识量变化趋势与 FCN 中的全连接层相似。



(a) FCN (b) CNN

图 3-8 FCN 和 CNN 每层知识量的比较

通过实验，本节分析了 FCN 和 CNN 的工作机制。在 FCN 中，知识量的变化主要通过激活强度的变化实现。尽管在第一层已经学习到全局特征，但这些特征仅仅是初步的特征。随着层数的增加，学习图像的激活强度变大。FCN 在图像相同位置上学习到的特征复杂性随着层数的增加而增加。在 CNN 中，知识量的变化主要通过激活数量的变化实现。低层神经元学习简单的局部特征，而高层神经元学习复杂的全局特征。

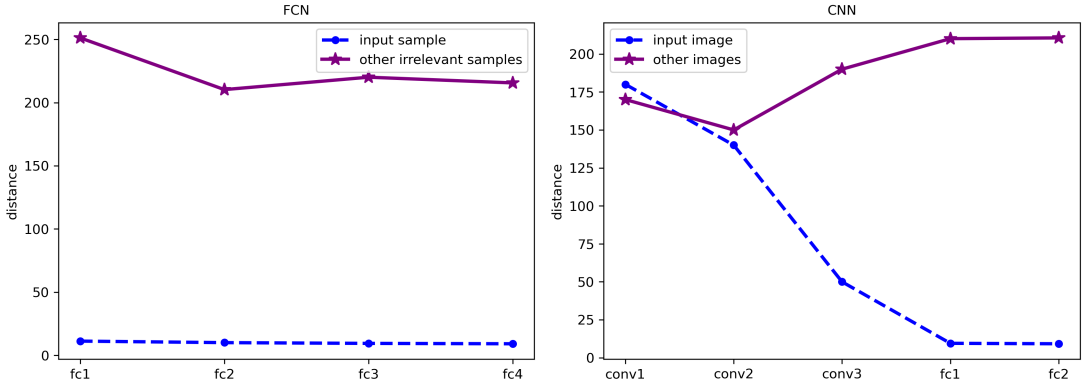
3.3.2 神经元学习规则

为了探究神经元关注的特征，本节计算学习图像与数据集样本之间的相似度。比较学习图像与对应输入样本的相似度，以及学习图像与其他无关样本的相似度。通过计算两幅图像之间的距离来衡量相似度。输入样本 I 和神经元 o 的学习图像 I_o 之间的距离计算如下：

$$dist_o^I = \left\| I_o - I \right\|. \tag{3-15}$$

距离越大，图像之间的相似度越低，反之亦然。图3-9计算了学习图像与对应输入样本之间的距离。为了进行比较和分析，同样计算了相同学习图像与其他无关样本之间的距离。

在 FCN 中，如图3-9a所示，不论在哪一层，学习图像与其他无关样本之间的距离都远大于学习图像与其输入样本之间的距离。全连接层中的神经元没有固定的学习特征。神经元学习的特征与输入样本始终有很大的关系。在 CNN 中，由于池化操作不改变神经元的学习图像，因此仅展示卷积层和全连接层中神经元的学习图像。如图3-9b所示，在卷积层 1 和卷积层 2 中，学习图像与其他无关样本之间的距离和学习图像与输入图像之间的距离相似。低层卷积层的神经元倾向于学习固定的特征。在卷积层 3 中，学习图像与输入图像之间的距离小于学习图像与其他无关样本之间的距离。与低层卷积层的学习图像相比，高层卷积层的学习图像与输入图像更加相似。全连接层中神经元的学习图像变化趋势与 FCN 中的全连接层相似。



(a) FCN　　(b) CNN

图 3-9　学习图像与样本之间的相似性

通过实验分析了不同神经网络结构的学习规则。全连接层中神经元学习到的特征与输入样本中的特征始终保持高度相似。全连接层神经元学习到的特征并不是固定的，而是随着输入样本的变化而变化。全连接层的神经元学习全局特征并获取全局轮廓信息，这是指定输入样本与其他无关样本之间的明显区别。卷积层中的神经元倾向于学习固定特征，但这种特性随着层数的上升而减弱。低层卷积层的神经元关注图像局部特征，这些特征与输入样本的全局轮廓关系较小。在高层卷积层，神经元学习更复杂的特征，且涉及更大的区域。学习到的特征与输入样本之间的相似度增大。

3.3.3 获胜神经元分析

在本节探讨了神经网络输出层中各神经元判别特征的能力。对于相同的输入样本，比较了不同输出神经元重建输入样本的能力。获胜神经元是对应于正确类别的输出神经元。输入样本 I 与指定输出神经元 o 的学习图像 I^o 之间的距离计算如公式3-15所示。

所有属于类别 c 的输入样本为 I_c。输入样本 I^c 与指定神经元 o 之间的距离表示为类别 c 中每个样本与神经元 o 之间的距离的平均值：

$$DIST_o^c = \frac{\sum_{I \in I^c} dist_o^I}{N}, \tag{3-16}$$

其中 N 是类别 c 的输入样本数量。

图3-10展示了输入样本与输出神经元学习图像之间的距离。图3-10中有 10 个子图。在第 c 个子图中，展示了类别 c 的输入样本与不同输出神经元之间的距离。当 $o = c$ 时，距离最小。对于 FCN 和 CNN，在每个子图中，与其他输出神经元的学习图像相比，获胜神经元的学习图像更接近输入样本。获胜神经元学习到的特征与输入样本最为相似。换句话说，和输出层中的其他神经元相比，获胜神经元能够更好地重建输入样本。

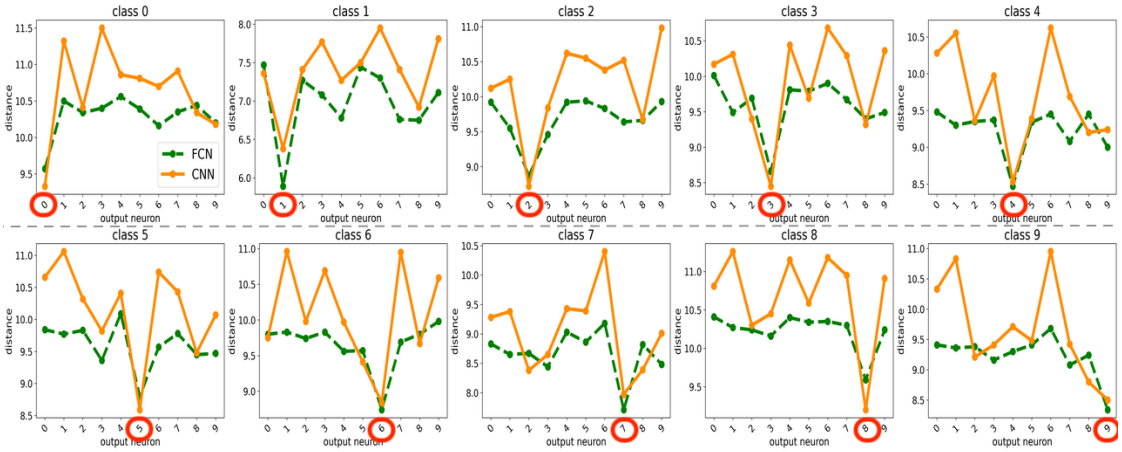


图 3-10 输入样本与输出神经元学习图像之间的距离

为了分析获胜神经元的特性，表3-2对三种不同准确率的神经网络进行分析，涉及 FCN 和 CNN 两种模型。首先，分析了距离和准确率之间的关系。展示了获胜神经元的学习图像与输入样本之间的距离。对于 FCN 和 CNN，模型的准确率越高，距离越小。即，准确率更高的神经网络学习到的特征与输入样本更相似。

表 3-2　获胜神经元的结果

	FCN			CNN		
准确率	40	70	90	40	70	90
距离	9.10	8.84	8.47	8.91	8.52	8.35
激活强度	0.589	0.648	0.665	0.427	0.453	0.678
激活数量	1.0	1.0	1.0	0.995	0.996	0.998

随后，本节度量了获胜神经元学习图像的激活强度和激活数量。如表3-2所示，对于 FCN 和 CNN 的获胜神经元，激活强度随着模型准确率的提高而增加。对于 FCN，激活数量在不同的神经网络中始终为 1。而对于 CNN，激活数量随着准确率的提高而增加，并始终保持较高的值。无论是 FCN 还是 CNN，具有不同准确率的神经网络中的获胜神经元都获得了全局特征信息。准确率更高的神经网络中的获胜神经元学习到的特征具有更高的激活强度。

3.3.4　跳跃连接结构分析

本节对具有跳跃连接的不同结构模型的学习过程进行比较。设计了一个包含一个残差块的全连接结构和一个 ResNet10 模型的实验。比较了残差块前、残差块内和残差块后的神经元学习图像。图3-11展示了同一位置的神经元学习图像的变化。残差块中的神经元学习到的特征与残差块前的神经元学习到的特征不同。

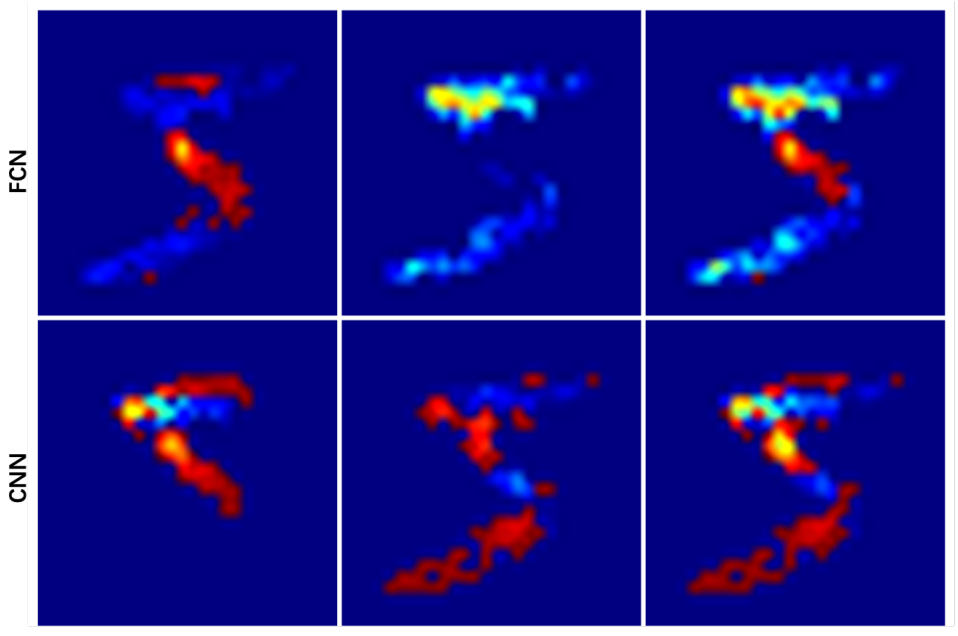
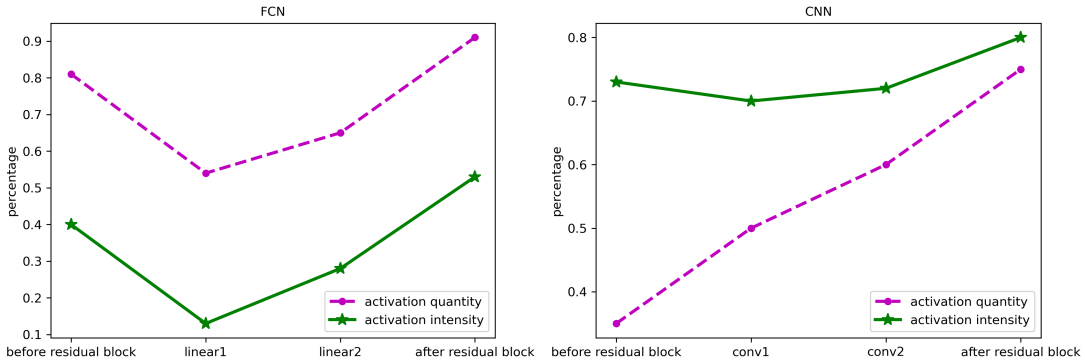


图 3-11　具有跳跃连接的全连接结构（上）与具有跳跃连接的卷积结构（下）的可视化比较



(a) FCN (b) CNN

图 3-12 具有跳跃连接的不同结构的知识量比较

图3-12显示了不同结构中具有跳跃连接的知识量。对于 FCN，残差块中神经元的激活数量和激活强度低于残差块前的神经元。加入残差块后，神经元的激活数量和激活强度显著增加。对于 CNN，激活数量始终保持增长趋势。残差块中的激活强度相较于残差块前有所下降。加入残差块后，激活数量显著增加，激活强度略有上升。无论是 FCN 还是 CNN，残差块都增加了神经网络的知识量。

3.4 CNN 解释评估实验

本节解释评估 CNN 模型。使用 AlexNet 和 VGG-16 作为代表模型来分析 CNN。这些模型在 ILSVRC-2013 DET 数据集上进行训练。首先通过评估模型在某一层的可解释性来验证 NNS 的有效性。接着，分析在同一层中高激活神经元的学习能力。最后，将 NNS 的可视化结果与现有可解释性方法进行比较。

3.4.1 层级解释评估

本节通过实验分析生成单个神经元学习图像是否具有意义。特征图具有定位物体的能力，因此将特征图作为基准。通过测量 CNN 模型单层中学习图像与特征图之间的相关性来评估 NNS 的可解释性。某一层的学习图像是该层所有神经元学习图像的总和。实验发现，某一层的学习图像与相应的特征图具有高度的相关性。换句话说，某一层学到的特征是该层所有神经元所学特征的总和。

对于输入图像，$F^{k,l}$ 是第 l 层滤波器 k 的特征图。$f^{k,l}_{(x,y)}$ 表示 $F^{k,l}$ 在空间位置 (x,y) 处的激活值。$I^{k,l}_{(x,y)}$ 是 $F^{k,l}$ 中 (x,y) 位置神经元的学习图像。使用 $I^{k,l} = \sum_{x,y} I^{k,l}_{(x,y)} \times f^{k,l}_{(x,y)}$ 来表示第 l 层第 k 个滤波器的学习图像。

为了将低分辨率的特征图与输入分辨率的学习图像进行比较，特征图通过双线性插值从 $F^{k,l}$ 上采样到掩模分辨率 $S^{k,l}$。特征图与学习图像之间的相关性表示为：

$$C^{k,l} = \frac{S^{k,l} \cap I^{k,l}}{I^{k,l}}. \tag{3-17}$$

公式 (3-17) 是交集与并集比率的变体。由于学习图像中包含的细节特征多于特征图，因此将 $S^{k,l} \cup I^{k,l}$ 替换为 $I^{k,l}$，使得该公式对变化更加敏感。

某一层特征图与学习图像之间的相关性值为：

$$C^{l} = \frac{\sum_{k \in K} C^{k,l}}{K}, \tag{3-18}$$

其中 K 是该层滤波器的数量。在第 l 层中，C^l 值越大，特征图与学习图像之间的相关性越高，反之亦然。

根据模型的不同，选择不同的层来衡量特征图与学习图像之间的相关性。对于 AlexNet，选择池化层 1、池化层 2、卷积层 3、卷积层 4 和卷积层 5。对于 VGG-16，选择池化层 1、池化层 2、池化层 3、池化层 4 和池化层 5。

相关性性能如表 3-3 所示。为方便起见，将这些层分别描述为第 1 层、第 2 层、第 3 层、第 4 层和第 5 层。对于 AlexNet 和 VGG-16，这两种模型的特征图与学习图像在不同层之间始终保持较高的相关性。两者的表现是可比的。在低层（第 1 层和第 2 层），特征图与学习图像之间的相似度非常高。由于特征图通过双线性插值被放大到输入分辨率，相较于学习图像，特征图缺乏细节。因此，在高层（第 3 层、第 4 层和第 5 层），一致性有所下降。

表 3-3　特征图与学习图像的相关性

	第 1 层	第 2 层	第 3 层	第 4 层	第 5 层
AlexNet	0.998	0.872	0.641	0.634	0.623
VGG-16	0.952	0.923	0.688	0.620	0.538

图 3-13 展示了 AlexNet 中每一层特征图与其对应学习图像的图像比较结果。学习图像与其对应的特征图之间高度相关。学习图像中的信息比对应特征图中的信息更为详细。因此，输入分辨率的学习图像展示了模型学习到的特征。每一层学习到的特征是该层所有神经元学习到的特征的总和。

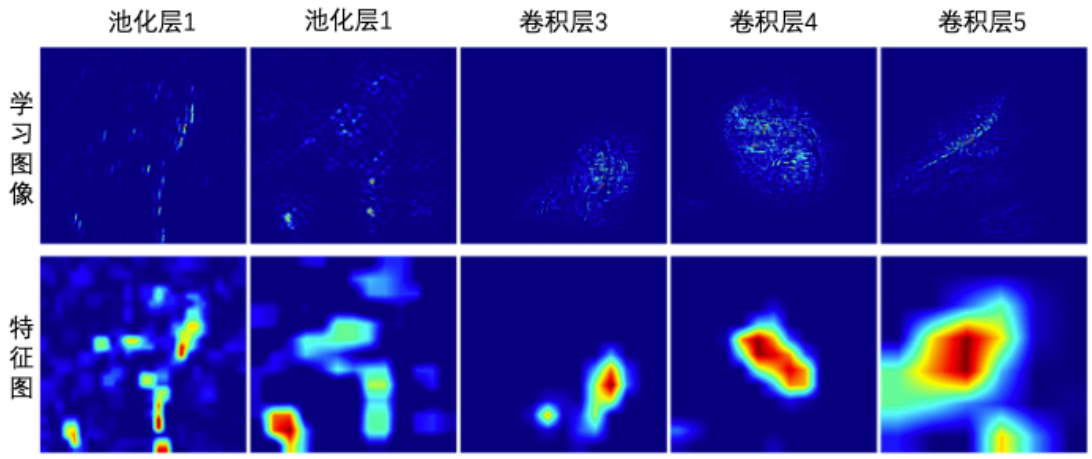


图 3-13 AlexNet 中学习图像及其对应的特征图

3.4.2 神经元级解释评估

通过 NNS，本节分析并比较同一卷积层中神经元的学习图像。对于输入图像，选择每个卷积结果中激活值前 2 的神经元，并展示它们的学习图像。这些神经元是该层最具代表性的神经元。实验发现，激活值前 2 的两个神经元的学习图像高度相似，这些神经元往往在图像的相似位置学习特征。

激活值前 2 的两个神经元的学习图像分别是 $I_1^{k,l}$ 和 $I_2^{k,l}$。使用交并比（IoU）来衡量这两个神经元之间的相关性：

$$IoU^{k,l} = \frac{I_1^{k,l} \cap I_2^{k,l}}{I_1^{k,l} \cup I_2^{k,l}}. \tag{3-19}$$

这些神经元在某一层的相关性为：

$$IoU^{l} = \frac{\sum_{k \in K} IoU^{k,l}}{K}, \tag{3-20}$$

其中 K 是该层中的滤波器数量。

激活值前 2 的神经元的相关性总结如表3-4所示。与之前的实验类似，本节选择不同的层来衡量不同模型中的相关性。在低层（第 1 层和第 2 层），相关性较高。激活值较高的神经元趋向于在相似的位置学习相同的特征。尽管在高层（第 3 层、第 4 层和第 5 层），相关性有所下降，但激活值较高的神经元学习的特征仍具有一致性。在 VGG-16 中，激活值较高的神经元在不同层之间始终保持较

高的相关性。与 AlexNet 相比，VGG-16 中的神经元相关性更强。低层中激活值较高的神经元趋向于学习相似的特征。高层神经元的学习能力在不同模型中差异显著。

表 3-4　高度激活神经元间的学习图像相关性

	第 1 层	第 2 层	第 3 层	第 4 层	第 5 层
AlexNet	0.999	0.953	0.761	0.676	0.615
VGG-16	0.997	0.990	0.873	0.835	0.804

3.4.3　与其他可解释方法比较分析

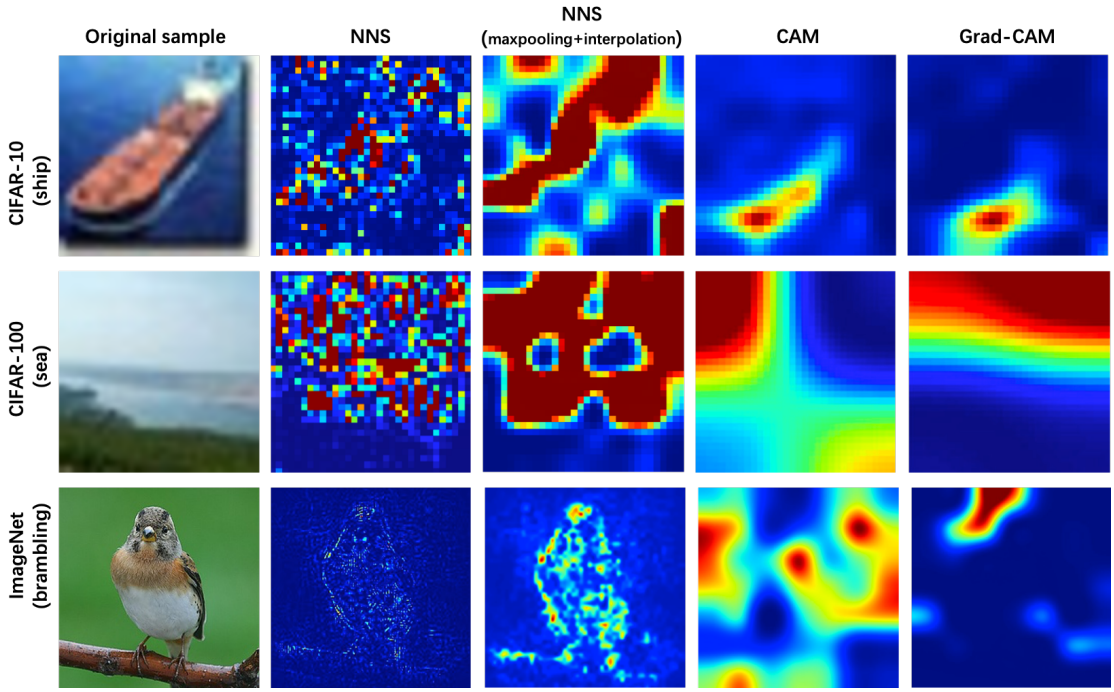


图 3-14　不同解释方法的可视化结果比较

本节将 NNS 与 CAM 和 Grad-CAM 的可视化结果进行比较。在图3-14中，展示了在 CIFAR-10、CIFAR-100 和 ILSVRC-2013 DET 上的视觉比较结果。对于 NNS，结果展示了输出层获胜神经元的学习图像。NNS 生成的热图与原始图像大小相同。而 CAM 和 Grad-CAM 生成的是小尺寸的热图，之后通过插值放大到原始图像大小，因此它们的热图比 NNS 的热图更加平滑。与大尺寸的 ILSVRC-2013 DET 数据集相比，这一现象在小尺寸的 CIFAR-10 和 CIFAR-100 上尤为明显。为了减少这一现象对比较结果的影响，对 NNS 的结果进行了最大池化和插

值处理。NNS 展示了模型从输入样本中学到的内容，以及模型如何描述样本。而其他算法则展示了模型的兴趣区域。在 CIFAR-10 和 CIFAR-100 上，NNS 展示了物体特征信息。在 ILSVRC-2013 DET 数据集上，可以看到 NNS 描述了输入样本的轮廓和重要细节。

3.5 本章小结

本章提出了一种基于神经元定性解释神经网络工作过程的方法。所提出的神经网络扫描仪能够在不改变模型结构的情况下，直观地展示各个神经元学习到的特征。通过统一的可视化结果展示不同模块，能够直接比较不同模型的工作机制。本章对 FCN 和 CNN 结构进行了实验，并通过这些实验深入理解神经网络的工作原理。

首先，可视化训练后的网络中各个神经元的学习情况，以验证 NNS 的有效性，并分析不同模型的工作机制。接着，通过分析学习图像与不同样本之间的相似性，探讨了不同模型中神经元的学习规则。然后，分析了输出层中获胜神经元学习的特征，并深入理解其在分类任务中的作用。实验还比较了不同结构中带有跳跃连接的学习过程，观察这些连接如何影响特征的学习。

实验从不同角度评估了 CNN 模型的可解释性，首先验证了 NNS 在每一层中评估可解释性的有效性。然后分析了高度激活神经元的学习能力，并将 NNS 的可视化结果与现有解释方法进行了比较。实验结果表明，NNS 显著提高了探索神经网络特征的能力。

第四章　基于 NNS 的神经元层面解释

4.1　引言

当前关于神经网络可解释性定量评估的研究大多集中在特征图或整个模型上，这些方法通常与特定的神经网络结构相关[113-114]。这些方法使得比较不同结构的功能变得具有挑战性[115]。由于神经元是神经网络模型的基本单元，定量分析它们的作用有助于对模型行为的深入理解，从而构建定量解释模型的框架。通过分析单个神经元的影响，能够从微观角度解释神经网络模型，观察每个神经元学习到的特征如何组合以分析模型的性能。

第三章提出了一种名为“神经网络扫描仪（*NNS*）”的解释方法，该方法以图像形式可视化神经元的学习结果。神经网络扫描仪可以以一种统一方式对神经网络中不同组件进行可视化，且无需改变模型架构，适用于各类人工神经元模型。然而，尽管神经网络扫描仪能够有效地可视化神经元学习到的特征，但在定量衡量这些特征方面存在不足。由于缺乏标准评估准则，神经网络扫描仪的结果往往具有主观性，通常需要额外的人工解释。针对这一问题，本章提出一种方法，通过衡量学习图像从输入样本中获取的特征，量化神经元学习到的信息。

理解神经网络模型内部工作机制一种广泛使用的技术是显著图，它突出显示了输入变量对给定样本的模型预测的贡献。显著图的概念最早由 Simonyan et al.[48] 提出，他们使用分类分数相对于输入像素的梯度生成热图。此后，诸如 Smilkov et al.[58]、Springenberg et al.[116] 和 Sundararajan et al.[117] 等研究者进一步改进了该方法，尤其是在用于分类任务的 CNN 模型中。另一个具有重要意义的方法是类激活映射，由 Zhou et al.[49]提出。CAM 在 CNN 模型中使用全局平均池化代替全连接层，使输出层的权重能够被投射回卷积特征图上，从而识别图像中的重要区域。在此基础上，进一步提出了方法，以突出显示图像中用于预测特定概念的关键区域[50-51]。梯度加权类激活映射[52] 和 Grad-CAM++[53] 通过不修改模型架构的方式解释模型。这些基于梯度的 CAM 方法使用目标类输出分数

的梯度作为显著图的加权成分。Muhammad et al.[54] 通过使用卷积层学习到的表示的主成分来创建视觉解释。类似地，Zeiler et al.[55] 使用去卷积模型揭示神经网络在低层主要学习简单的边缘特征，而在高层学习更复杂的物体特征。其他方法，如 Ren et al.[56] 和 Zhang et al.[57] 提出的工作，专注于生成高质量的视觉解释，而无需依赖非盲去卷积，或通过提取图形模型来描述特定滤波器的内容。然而，显著图和类似的基于梯度的方法有其局限性。它们通常仅反映模型在狭窄输入范围内的行为，这可能导致误导性的特征重要性估计[59]，特别是在大型视觉模型中梯度的噪声性质[58]。为了应对这一问题，诸如 Rise[60]、Sobol[61] 和 HSIC[62] 等方法通过对输入图像进行扰动，创建了更可靠的重要性图。

尽管显著性方法能够提供关于模型关注区域的视觉线索，但它们通常是主观的，需要额外的人工解释。为了增强可解释性，Kim et al.[64] 提出了基于显著图来衡量预选概念影响的方法，尽管该方法需要人工注释的数据库，成本较高。为了解决这一问题，Ghorbani et al.[65] 提出了自动化方法从数据集中提取概念，识别适用于整个数据集而非单一样本的概念。类似地，Cheng et al.[66] 开发了一种方法，可以在不需要显式标签的情况下，量化 DNN 模型中间层编码的视觉概念。确保特征测量的公平性和避免偏倚是至关重要的。Sturmfels et al.[118] 强调了无偏测量的重要性，而 Kindermans et al.[119] 提出了输入不变性的公理来评估显著性方法的可靠性。此外，Hsieh et al.[120] 基于稳健性分析建立了评估标准，通过细微的对抗性扰动来避免引入偏见和伪影。Haug et al.[121] 提出了一个新分类法，通过与基准参考比较输入特征的重要性，并评估不同数据集上的常见归因模型。这些现有的方法主要是在模型层面分析神经网络，这使得很难理解单个神经元的运作机制或模型中特定层和神经元的功能。相比之下，本章的方法将每个神经元视为一个独立的单元，为模型学习到的特征提供了更细粒度的解释。与显示模型感兴趣区域的显著性图不同，本章的方法从神经元中心的角度揭示了从输入样本中学习到的特定特征。

如图4-1所示，本章目标是通过衡量学习图像从输入样本中获取的特征，量化每个神经元编码的信息。首先使用神经网络扫描仪为每个指定神经元基于给定输入样本生成学习图像，并引入了一种新的数学度量来评估输入样本与学习图像之间的相关性，将其称为“特征量（*Feature Quantity, FQ*）”。通过特征量能够量化和比较不同神经元编码的特征，并进一步探讨这些神经元的工作机制。

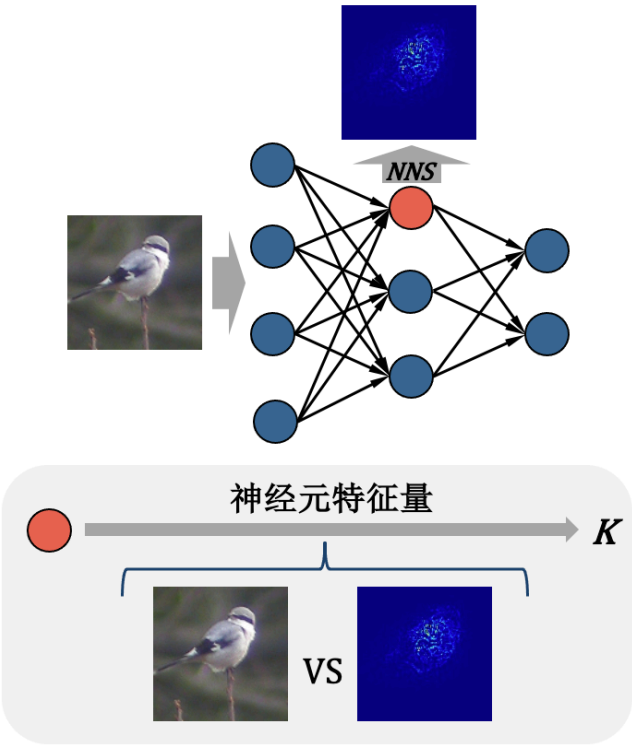


图 4-1 神经元特征量的度量

利用特征量从神经元层面分析 CNN 模型中卷积层的工作机制，并从三个方向分析模型：

- 神经元的特征量与激值的关系。神经元的激活值是其基本属性，通常被认为代表了该神经元以可量化形式学习到的信息。为了探索这种“信息”的本质，本章研究了激活值与相应特征量之间的关系。研究发现，神经元的特征量与其激活值之间存在正相关关系。激活值较大的神经元意味着学习到的特征更多。
- 不同层级的特征量的变化情况。本章分析了模型不同层之间特征量的变化。通过观察每一层神经元特征量的分布情况，发现不同层之间的特征量呈现出一种有趣的模式：随着层深的增加，神经元捕捉到的特征先变得更加多样，然后逐渐趋向统一。
- 不同滤波器的特征学习能力。为了研究不同滤波器在表示学习和模型性能中的作用，本章对每个滤波器内神经元的特征量进行了聚类，并将结果作为度量该滤波器的特征多样性的指标。聚类数较高的滤波器表现出更强的学习多样特征的能力，而聚类数较低的滤波器则专注于单一或冗余的特征

表示。实验表明，特征量多样性较高的滤波器对模型性能重要，特征量多样性较小的滤波器对模型准确度影响较小。

本章的贡献如下：

引入了神经元特征量的概念，这是一种统一的度量神经元的标准，用于量化 CNN 中神经元学习到的特征。该度量具有可解释性、普适性、有效性和可分离性。基于这一度量分析了神经元的功能，并解释了卷积层的工作机制。在不同神经网络模型上进行了实验，从三个方面分析模型，并讨论了不同模型之间的差异。

4.2　基于 NNS 的特征量生成

本节通过量化卷积层中每个神经元所学习到的特征，来研究 CNN 模型中神经元的运作。神经网络扫描仪通过生成每个神经元的学习图像，直观地表示了神经网络模型中每个神经元所学习到的特征。为了对这些神经元进行更深入的分析，建立一个统一的度量来量化学习图像中所呈现的特征是至关重要的。本节通过将每个神经元的学习图像与原始输入图像进行关联，引入了“特征量”这一概念来量化了神经元学习的特征。特征量越高，表示学习图像与原始图像越相关，也就是说该神经元从输入中提取到的特征更多。

4.2.1　神经元特征量的计算

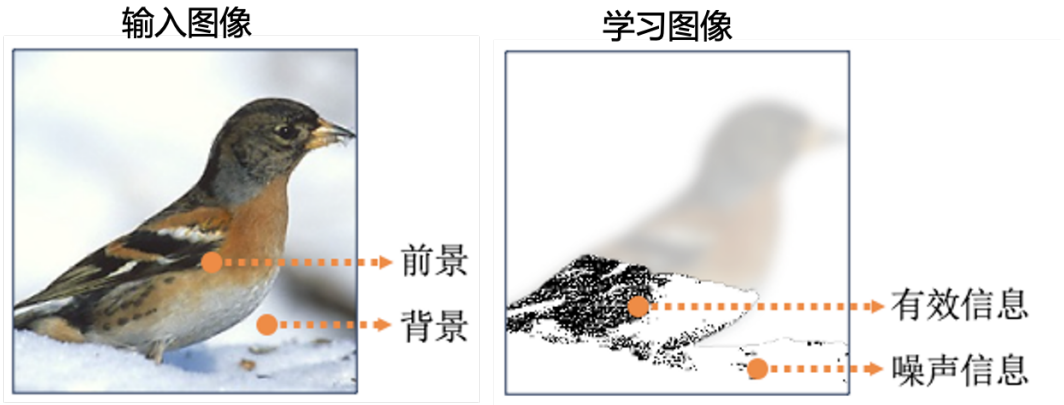


图 4-2　学习图像中信息的可视化

如图5-1所示，考虑一个输入图像为 $I \in \mathbb{R}^{H \times W}$ 的 CNN 模型，其中 I^f 和 I^b

分别表示前景图像和背景图像。对于给定的神经元 x，有效信息包括从前景图像中学习到的特征，而噪声信息则包含从背景图像中学习到的特征。神经元 x 的对应学习图像记为 I_x，其中 I_x^e 和 I_x^n 分别表示 I_x 中的有效信息和噪声信息。

$$I_x^e = I_x \cap M^f, I_x^n = I_x \cap M^b, \tag{4-1}$$

其中，M^f 和 M^b 分别是对应 I^f 和 I^b 的掩码。掩码在操作过程中被应用以确保图像尺寸的一致性。有效信息质量通过衡量 I_x^e 和相应前景图像 I_x^{fe} 之间的相似度来确定。

$$I_x^{fe} = I^f \cap M_x, \tag{4-2}$$

其中，M_x 是 I_x 的掩码。l、c 和 s 分别是 I_x^{fe} 和 I_x^e 的亮度对比、基于对比度的比较和结构对比。l、c 和 s 基于结构相似性指数（SSIM）[122]进行计算。

亮度对比为：

$$l_x = \frac{2\mu^{fe}\mu^e}{\mu^{fe2} + \mu^{e2} + C1}, \tag{4-3}$$

其中，μ^{fe} 和 μ^e 分别是 I_x^{fe} 和 I_x^e 的均值。图像的均值表示亮度信息。

对比度对比为：

$$c_x = \frac{2\sigma^{fe}\sigma^e}{\sigma^{fe2} + \sigma^{e2} + C2}, \tag{4-4}$$

其中，σ^{fe} 和 σ^e 分别是 I_x^{fe} 和 I_x^e 的标准差。图像的对比度信息通过标准差表示。

结构对比为：

$$s_x = \frac{\sigma^{fee}}{\sigma^{fe}\sigma^e + C3}, \tag{4-5}$$

其中，σ^{fee} 是 I_x^{fe} 和 I_x^e 之间的协方差。结构元素通过标准差以及相应的均值和标准差表示。$C1$、$C2$ 和 $C3$ 是为数值稳定性引入的小常数。此处设置 $C1 = 10^{-4}$，$C2 = 3 \times 10^{-4}$，$C3 = 1.5 \times 10^{-4}$。

神经元 x 的特征量通过比较 I 和 I_x 之间的差异来量化，公式如下：

$$F(I, x) = K\left(\frac{N_x^n}{N}\right)\left(\frac{N_x^e}{N^f}\right)(l_x \cdot c_x \cdot s_x), \tag{4-6}$$

其中，N 和 N_x^n 分别表示 I 和 I_x^n 中的激活像素数量，N_f 和 N_x^e 分别表示前景图

像 I^f 和 I_x^e 中的激活像素数量。K 是一个常数，此处设置 $K=10^4$。第一个因子 $\left(\frac{N_x^n}{N}\right)$ 和第二个因子 $\left(\frac{N_x^e}{N^f}\right)$ 从数量的角度量化特征，第三个因子 $(l_x \cdot c_x \cdot s_x)$ 则从质量的角度量化特征。

神经元 x 的特征量可以表示为：

$$FQ(x)=F(I^*,x), s.t. I^* := \arg \max_{i\in\{1,\cdots,M\}} F(I^i,x), \tag{4-7}$$

其中 M 为数据集样本数，这个最大值可作为衡量神经元是否具备结构保持能力的指标之一，值越大，说明存在某个样本，其局部结构在该神经元下被高度保留。特征量生成算法见算法 4.1。

算法 4.1 神经元的特征量生成

输入：输入数据集 $\mathbb{D}$，样本数为 N，神经元 x, 输入图像 I 的学习图像对神经元 x 的学习图像 I_x

参数设置：I^f 和 I^b 分别为前景图像和背景图像，I_x^e 和 I_x^n 分别为 I_x 中的有效信息和噪声信息，特征量 $FQ=0$

repeat

$I^i \in \mathbb{D}$

$l_x = \frac{2\mu^{fe}\mu^e}{\mu^{fe^2}+\mu^{e2}+C1}$

$c_x = \frac{2\sigma^{fe}\sigma^e}{\sigma^{fe^2}+\sigma^{e2}+C2}$

$s_x = \frac{\sigma^{fee}}{\sigma^{fe}\sigma^e+C3}$

$F = K\left(\frac{N_x^n}{N}\right)\left(\frac{N_x^e}{N^f}\right)(l_x \cdot c_x \cdot s_x)$

if $F > FQ$ **then**

$FQ = F$

end if

until $i = N$

输出：神经元 x 特征量 FQ

4.2.2　神经元特征量的解释能力分析

（1）可解释性：提出的算法通过特征量来衡量和解释神经元学习到的特征。通常，来自前景的特征与任务相关，而来自背景的特征则主要与任务无关。因此，一个特征量较大的神经元表示它从前景中学习到了大量的特征。特征量本身提供了强大的解释能力。学习到的特征由三个因素特征化，信息从定量和定性角度进行评估。第一个因素 $\left(\frac{N_x^n}{N}\right)$ 衡量神经元学习的特征中噪声的比例，具体来说，是从背景中学习到的特征与整体特征的比率。第二个因素 $\left(\frac{N_x^e}{N^f}\right)$ 量化了神

经元学习到的有效信息量，即神经元从原始输入前景中捕获的有效信息的比例。神经元学习到的有效信息越多，特征量越大。第三个因素 $(l_x \cdot c_x \cdot s_x)$ 评估了神经元学习到的有效信息的质量，表示神经元学习到的有效特征与原始输入图像相应部分之间的相似度。有效特征与原始图像之间的相似度越高，神经元的特征量越大。

（2）普适性：由于特征量是基于神经元设计的，而神经元是神经网络模型的基本单元，因此它适用于不同的神经网络架构。从理论上讲，这一度量允许在不同架构和层次之间进行公平比较，能够观察和分析不同模型操作机制的差异。

（3）有效性：特征量的测量结果与人类对神经网络模型的基本理解一致。如果神经元的激活值为 0，则表示该神经元未被激活，理论上没有学习到任何特征。因此，相应的学习图像将为空白。在这种情况下，神经元的特征量为 0，反映出神经元所学习特征的量也是 0。

（4）可分离性：设计的特征量由三个独立的因素组成，每个因素都可以单独测量和分析。每个因素都有明确而独特的含义。特征量区分了神经元学习到的有效信息和噪声信息，即它区分了从前景图像中学习到的特征和从背景图像中学习到的特征。每个因素都可以单独用于分析神经元。

4.2.3 特殊状态下的神经元特征量分析

为了验证特征量在度量神经元是否有效学习图像特征方面的合理性，本节从两个极端情况出发进行分析：（1）完全学习到输入样本的结构特征；（2）完全未学习任何特征。

神经元完全学习到输入样本的结构特征

给定输入图像为 $I \in \mathbb{R}^{H \times W}$ 及其对神经元 x 的学习图像 $I_x \in \mathbb{R}^{H \times W}$。若神经元完全学习到输入图像的特征，可视作学习图像为输入图像的线性缩放版本，即存在常数 $p > 0$，使得：$I_x = pI$。

设 μ_I、σ_I 分别为输入图像的均值与标准差，则有：

$$\mu_{I_x} = p\mu_I, \quad \sigma_{I_x} = p\sigma_I, \quad \sigma_{I,I_x} = p\sigma_I^2. \tag{4-8}$$

此时学习图像中激活像素点与原图中完全一致，则有：

$$\mu_{u^e} = p\mu_{u^{fe}}, \quad \sigma_{u^e} = p\sigma_{u^{fe}}, \quad \sigma^{fee} = p\sigma^2_{u^{fe}}, \quad N_x^n = N, \quad N_x^e = N^f, \tag{4-9}$$

代入公式4-6，则有：

$$F(I,x) = \frac{2p\mu^{fe^2}}{(1+p^2)\mu^{fe^2}+C1} \times \frac{2p\sigma^{fe^2}}{(1+p^2)\sigma^{fe^2}+C2} \times \frac{p^2\sigma^{fe^2}}{\sigma^{fe^2}+C3}, \tag{4-10}$$

当 $p=1$ 且 $C_1, C_2 \to 0$ 时，有：$F(I,x) \to 1$。即若学习图像与输入图像在结构上完全一致，则神经元特征量可达最大值 1，反映神经元成功学习到了图像结构。

神经元完全未学习任何结构特征

若神经元未被激活，即对图像无响应，其学习图像为全零 $I_x = 0$。此时有：

$$\mu_{u^e} = 0, \quad \sigma_{u^e} = 0, \quad \sigma^{fee} = 0, \tag{4-11}$$

代入公式4-6，则有 $F(I,x) = 0$。因此，当学习图像为空白时，与原图结构相似性为 0，意味着神经元未能学习到任何图像结构特征。

上述分析说明，特征量可用于度量神经元对输入特征的学习能力。两种情况对比如表4-1。当学习图像与输入图像相似时，特征量显著增大，反之则趋近于零。因此，特征量可作为解释神经网络特征提取效果的有效指标。

表 4-1　神经元学习状态与特征量间的关系

学习图像情况	学习图像形式	特征量	学习图像与原图相似性
完全学习特征	$I_x \propto I$	接近 1	高
完全未学习特征	$I_x = 0$	接近 0	低

4.3　基于神经元特征量的模型解释

4.3.1　神经元内在属性分析

神经元的激活值是神经网络中最基本的属性之一。普遍认为，激活值代表了神经元所学习到的信息，并且可以用数值表示。然而，这些“信息”的确切性质尚未完全明了。本节研究了神经元的激活值与其对应特征量之间的关系。使用

特征量来量化神经元学习到的信息。本质上，特征量衡量了神经元所学习到的特征与原始输入图像之间的相似度。这种方法为神经元所学习的信息提供了一个更容易理解的定义。

对于同一层中的一组神经元，通过测量皮尔逊相关系数 $r(A, K)$ 来分析激活值集 A 与特征量集 K 之间的关系：

$$r(A, K) = \frac{\sigma_{AK}}{\sigma_A \sigma_K}, \tag{4-12}$$

其中，σ_{AK} 是 A 和 K 之间的协方差，σ_A 和 σ_K 分别是 A 和 K 的标准差。

结果发现，在每一层中，A 和 K 之间的相关系数都是正相关的。这表明，神经元的激活值越大，表示它从原始图像中学习到的特征越多，也就是说，它学习到的特征与原始图像之间的相似度越高。

4.3.2 神经元学习规则讨论

在本节中探讨卷积层中神经元的操作机制。测量了每层中所有神经元的特征量，并分析了神经元学习到的特征的特性。在 CNN 模型中，不同层次的特征量呈现出一个有趣的模式：随着层深的增加，模型捕获的特征首先变得更加多样，然后逐渐趋向一致。这一观察结果可以通过检查每一层中神经元的特征量的均值和方差来定量分析。

特征量的均值反映了每一层中神经元学习到的特征与原始输入图像之间的平均相似度。较高的均值表明神经元捕获了更多输入的全局或显著特征。另一方面，特征量的方差则反映了该层中学习到的特征的多样性。较大的方差表明该层中的神经元学习了高度不同的特征，导致了丰富的表示。

结果发现，随着网络从低层到中间层的逐步推进，特征量的均值增加，这表明这些层学习到了越来越抽象但与输入相关的特征。然而，在更深的层次中，均值开始下降，这意味着网络开始更加专注于任务特定的表示，而这些表示与原始输入的原始特征的关联性较低。早期层次中观察到的方差上升表明网络正在探索广泛的特征，捕捉各种模式和细节。随着深度的增加，方差最终减少，这表明网络逐渐收敛到更统一和一致的表示上。这反映出一个逐渐聚焦的过程，其中神经元集体编码的信息更加符合特定的任务目标，从而减少了冗余。

4.3.3　神经元特征多样性分析

通过分析各个滤波器内神经元学习的特征多样性，可以深入了解不同滤波器在表示学习和模型性能中的作用。为了进行这一分析，对每个滤波器内神经元的特征量进行了聚类，并将聚类结果作为特征多样性的指标。

为了对神经元的特征量进行聚类，使用了基于密度的空间聚类算法（Density-Based Spatial Clustering of Applications with Noise, DBSCAN）[123]，该算法通过密度准则将靠近的点聚集在一起。同一聚类中的神经元具有相似的特征量。

结果发现，具有较多聚类的滤波器显示出神经元学习的特征具有显著的多样性，表明这些滤波器更能专注于多样化的特征。相反，具有较少聚类的滤波器表现出较小的特征学习变化，暗示它们更多地集中在单一或冗余的特征表示上。

为了分析特征多样性对模型损失的影响，对不同类型的滤波器进行了掩码处理。结果显示，掩码具有较多聚类的滤波器会显著增加模型的损失，证明这些滤波器对任务性能的关键贡献。这些滤波器编码了复杂且具有高度区分性的特征，这些特征对网络的整体功能至关重要。相比之下，掩码具有较少聚类的滤波器对模型损失的影响较小，表明它们所学习的特征对任务的影响较小，或者学习到冗余的特征。

4.4　实验结果与分析

4.4.1　实验设置

本节进行了多个实验以分析解释模型。为了确保全面的比较，将实验应用于 AlexNet、VGG-11、VGG-16 和 ResNet-18 模型。这些模型在 ILSVRC-2013 DET 数据集[124]、CUB-200-2011 数据集[125]和 Pascal VOC 2012 数据集[126]上进行了训练。对于 ILSVRC-2013 DET 数据集，由于计算需求较高，对多个类别进行了分类实验。定义目标的边界框为前景区域，边界框内的像素视为前景，框外的像素则视为背景。本章的重点是推导与卷积层相关的普遍原则，该层包含卷积功能和激活函数。因此，在后续的实验中强调特征量的变化趋势，而非特定的特征量值。所有实验重复三次，使用 PyTorch 实现。实验结构如下：首先，评估提出方法在不同神经网络中的有效性。然后，在不同模型上分析验证。

4.4.2 不同度量标准的比较

表 4-2 不同度量标准下的学习图像的特征

		conv1	conv2	conv3	conv4	conv5	conv6	conv7	conv8
AlexNet	ILSVRC-2013DET dataset								
	SSIM	0.95	0.52	0.24	0.76	0.42	-	-	-
	FQ	**0.0012**	**0.02**	**0.18**	**0.61**	**0.34**	-	-	-
	CUB200-2011 dataset								
	SSIM	0.92	0.48	0.35	0.74	0.41	-	-	-
	FQ	**0.0003**	**0.015**	**0.13**	**0.52**	**0.29**	-	-	-
VGG-11	ILSVRC-2013DET dataset								
	SSIM	0.96	0.92	0.84	0.67	0.71	0.74	0.68	0.53
	FQ	**0.0013**	**0.026**	**0.34**	**0.47**	**0.54**	**0.69**	**0.63**	**0.45**
	CUB200-2011 dataset								
	SSIM	0.95	0.94	0.81	0.51	0.62	0.67	0.87	0.79
	FQ	**0.0003**	**0.0013**	**0.23**	**0.37**	**0.59**	**0.64**	**0.78**	**0.54**
VGG-16	ILSVRC-2013DET dataset								
	SSIM	0.94	0.94	0.92	0.87	0.84	0.88	0.79	0.74
	FQ	**0.0011**	**0.018**	**0.23**	**0.36**	**0.48**	**0.51**	**0.57**	**0.60**
	CUB200-2011 dataset								
	SSIM	0.95	0.92	0.90	0.87	0.86	0.89	0.88	0.90
	FQ	**0.0010**	**0.013**	**0.18**	**0.27**	**0.33**	**0.39**	**0.48**	**0.51**
ResNet-18	ILSVRC-2013DET dataset								
	SSIM	0.97	0.95	0.90	0.88	0.79	0.75	0.70	0.66
	FQ	**0.0022**	**0.031**	**0.14**	**0.22**	**0.48**	**0.51**	**0.55**	**0.61**
	CUB200-2011 dataset								
	SSIM	0.96	0.91	0.89	0.83	0.77	0.74	0.71	0.59
	FQ	**0.0015**	**0.0027**	**0.18**	**0.27**	**0.32**	**0.44**	**0.47**	**0.51**

本节测量了不同层中神经元的特征量。为了方便比较，使用 SSIM 算法评估不同层中神经元的学习图像与原始样本之间的相似度。对于两种方法，使用同一层中神经元的平均值来代表该层的特征量。在 ILSVRC-2013 DET 数据集和 CUB-200-2011 数据集上对模型的前 8 层卷积层进行实验。

为简便起见，在表格中使用 FQ 来表示特征量。如表4-2所示，AlexNet 和 VGG-11 的 SSIM 值呈现出明显的规律。低层的神经元具有较高的 SSIM 值，因为这些神经元学习的特征很少，导致学习图像接近空白。因此，SSIM 值（用于衡量输入样本与几乎空白图像之间的相似度）较高。随着层数的增加，神经元学习的特征变得更加复杂，学习图像中的空白区域减少，从而导致 SSIM 值降低。在高层，神经元学习更多的特征，空白区域对 SSIM 值的影响减小，学习图像与输入样本之间的相似度随着层深增加而增加。

所提出的特征量（FQ）随着层数的增加，先下降再上升，显示出神经元响应与原始输入之间的相似度呈现明显的非单调趋势。FQ 有效地考虑了学习图像

中的空白区域的影响，为衡量神经元学习到的特征提供了更准确的度量。低层神经元学习到的是更简单的特征，而高层神经元学习到的是更复杂的特征。

4.4.3　基于特征量的模型分析

（1）神经元的特征量与激活值的关系

本节旨在分析神经元的特征量与其激活值相关关系。神经元在输入样本上的激活强度反映了其对特征提取的参与程度，而特征量作为衡量神经元编码能力的指标，应与激活水平具有一致性。采用皮尔逊相关系数（Pearson Correlation Coefficient）定量评估二者之间的相关性，在多个 CNN 模型上开展实证分析。

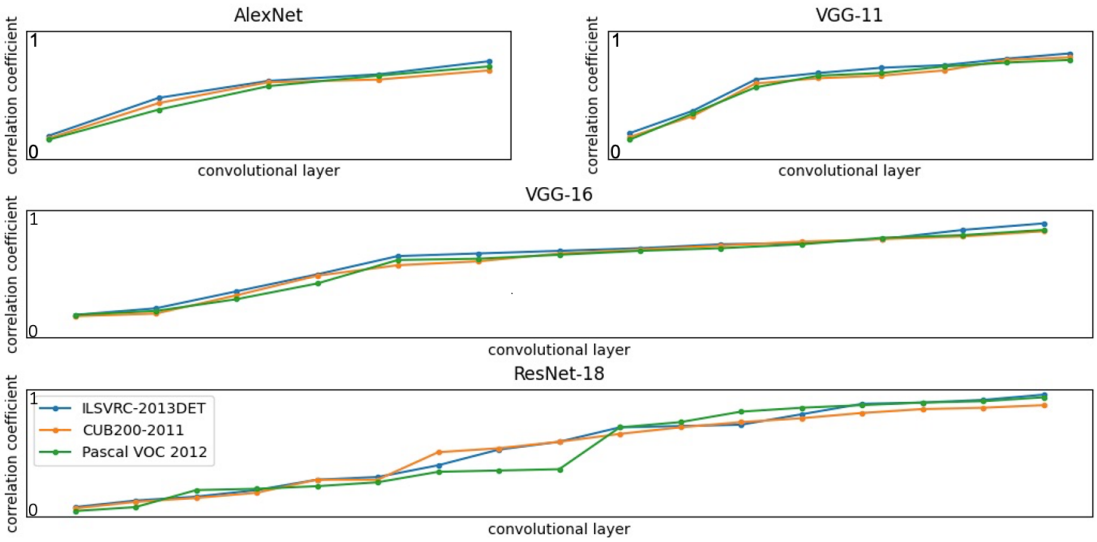


图 4-3　特征量与激活值的 Pearson 相关系数

实验对模型各个卷积层中所有神经元的特征量与其激活值进行统计与关联分析。对于每一层的每一个卷积滤波器中的神经元，计算其在多个输入样本上获得的平均激活值与平均特征量，并在样本维度上求取两者之间的 Pearson 相关系数。实验结果如图 4-3 所示，各模型中绝大多数卷积层的神经元特征量与激活值之间表现出显著的正相关性。从数值上看，在 VGG-16 模型中，中层与高层的相关系数逐步上升至 0.8 以上；在 ResNet-18 中，由于残差连接对特征表达的增强作用，该相关性整体更为明显，在多数中高层中可达到 0.85 甚至更高。值得注意的是，该相关性在网络不同层之间存在层级差异：低层神经元尽管数量众多，但由于其激活模式相对简单，特征量的分布较为集中，因此在这些层中特征量的变

化幅度较小，与激活值之间的相关性相对较弱。此外，低层主要负责低级边缘或纹理信息的捕捉，其激活受到输入图像结构的局限，表现出较强的模式一致性，也限制了特征量与激活值之间的离散度。相较之下，中层与高层卷积层的神经元所处理的特征更具语义性和抽象性，其激活值受样本内容影响更大，导致特征量的分布更广，从而增强了特征量与激活值之间的线性相关性。这种趋势表明，随着网络深度的增加，神经元对输入的选择性增强，个体间的差异更为显著，也更容易反映其在语义表征中的角色差异。

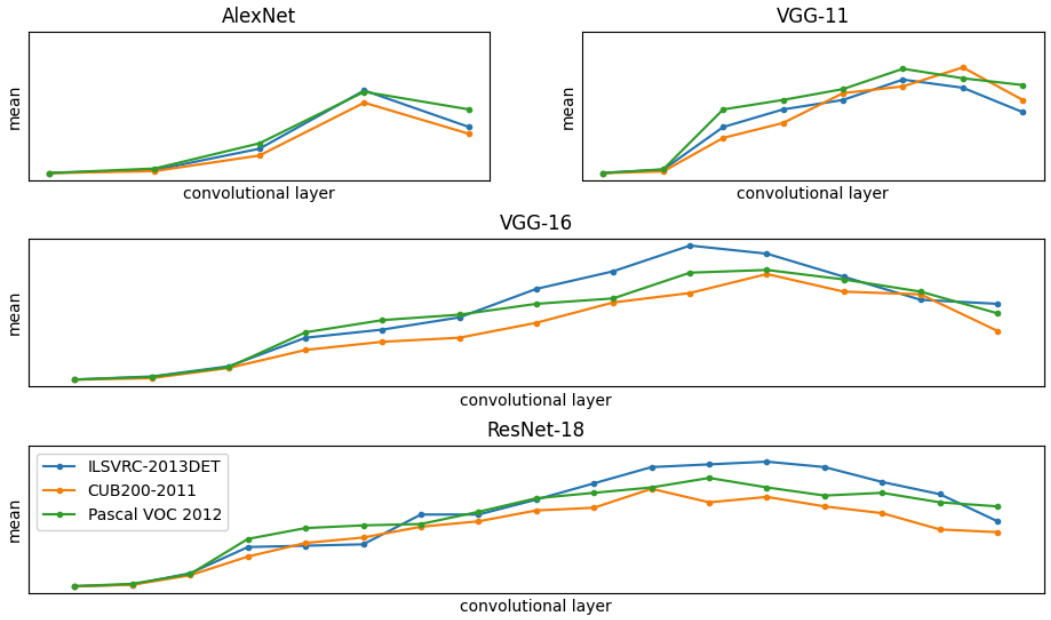
综上所述，特征量可以作为衡量神经元活跃程度及其编码重要性的有效指标，尤其在中高层中具有更强的判别力。

（2）不同层级的特征量的变化情况

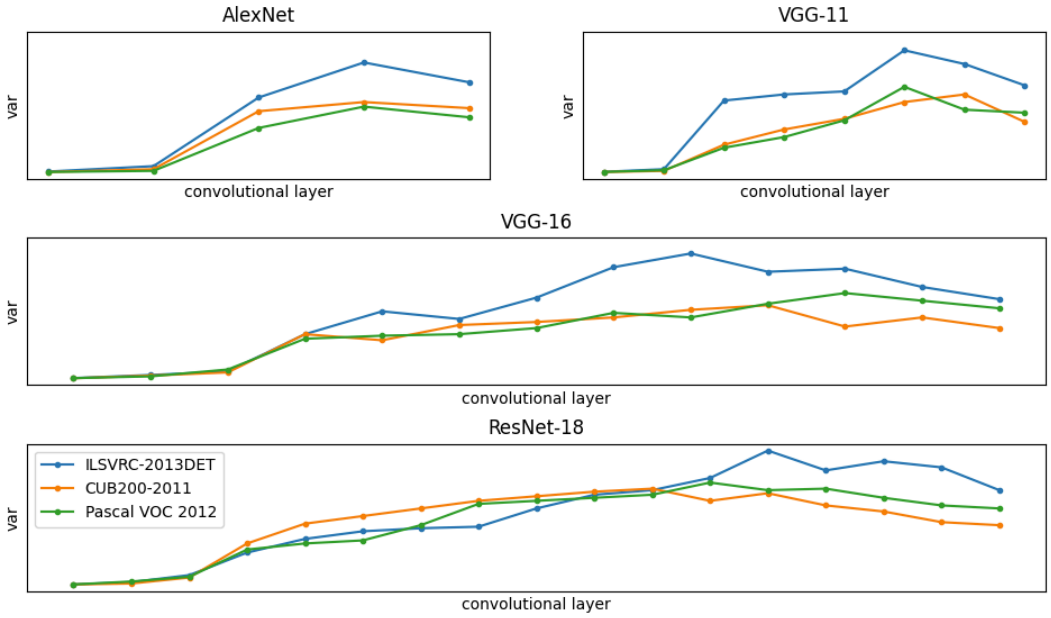
本节对神经元特征量的均值和方差进行统计分析，分别从“响应强度的集中趋势”与“表示的多样性”两个维度刻画神经元特征表示的层次性特征。具体地，选取多个神经网络模型，在标准数据集上训练后，对各个卷积层内的所有神经元特征量进行统计，计算其在样本维度上的均值和方差，并绘制其随网络深度变化的趋势图，如图 4-4 所示。

首先，特征量均值反映了神经元响应与输入特征之间的相似性。图 4-4a 展示了不同层特征量均值的变化趋势，呈现出明显的非单调型变化。在低层阶段，随着网络初步学习图像的边缘、角点、纹理等基本视觉结构，神经元的响应与输入图像具有高度的结构相关性，因此特征量均值逐渐上升。这一阶段的特征提取主要是通用的、低级的、与原始图像结构紧密相关。进入中间层后，特征量均值达到峰值，表明此阶段的神经元能够学习到更为抽象且多样的表示。这些表示一方面仍保留了原始图像中的重要结构信息，另一方面开始捕捉与任务相关的语义概念，因此响应强度较高，表征能力最强。这也说明了中间层是网络中特征表达能力最丰富的阶段。在更高层，特征量均值出现下降趋势，说明网络逐步将注意力从输入本身转移到与目标任务（如分类决策）直接相关的抽象表示。这些高层神经元往往响应于更稀疏、更具判别性的模式，特征量分布趋于集中，表征维度减少。这一变化反映了网络表示从“感知驱动”逐步转向“任务驱动”的方向性转变，体现出 CNN 从感知到决策的语义抽象路径。

其次，特征量方差的分析揭示了各层神经元编码信息的多样性变化规律。如图 4-4b 所示，特征量方差也呈现出非单调变化的趋势：在低层阶段，神经元广



(a) 均值



(b) 方差

图 4-4　不同层神经元特征量的均值和方差

泛探索输入空间中各类低级特征，学习模式种类较多，导致方差逐步上升；在中间层，方差达到最大值，神经元编码了大量不同方向、尺度、形状、语义的特征，使得特征空间达到最大复杂度。而在更深层，方差出现显著下降。这种下降表明神经元表示趋于收敛，编码的内容逐步统一为一组具有较高决策效能的特征子集。换句话说，高层神经元不再追求广泛的模式覆盖，而是聚焦于最具判别力的语义特征，以实现对目标任务的最优支持。这一特征压缩机制提高了模型在决策

阶段的效率，也减少了表示冗余。

综合均值与方差的分析结果，本节发现 CNN 模型在特征学习过程中表现出由“多样性”向“一致性”的特征表示过渡：低层更侧重于信息覆盖与结构捕捉，中间层最大程度发挥特征组合与抽象能力，而深层则通过特征聚合实现高效、目标导向的分类判别。

（3）不同滤波器的特征学习能力

本节旨在通过定量分析同一层内不同滤波器的特征多样性，并研究其对模型性能的贡献。实验发现，CNN 模型中不同滤波器的神经元在特征学习能力上存在明显不均衡，且特征量多样性较高的滤波器对于模型性能具有更为关键的作用。为此，本节采用基于密度的空间聚类算法 DBSCAN（Density-Based Spatial Clustering of Applications with Noise）来评估单层内不同滤波器神经元特征量的相似性和多样性。DBSCAN 能够有效发现神经元特征量空间中的密集区域，从而划分出具有相似特征表达的神经元簇。实验中，将 DBSCAN 的参数设置为：$eps = 0.3$（邻域半径）和 $min - samples = 6$（构成簇的最小样本数），确保聚类结果具有合理的紧密度与代表性。

本节分析聚焦于每个模型的最后一个卷积层，因为该层通常承载着较高层级的语义信息，是特征表达的重要节点。通过对该层每个滤波器内神经元的特征量进行聚类，计算并统计各滤波器的聚类数目，聚类数量即被用作该滤波器特征多样性的量化指标。聚类数越多，说明滤波器内神经元学到的特征类型越丰富，表现出更强的多样性。为了进一步探究不同滤波器特征多样性对模型性能的影响，本节基于聚类数量对滤波器进行排序，并分别对聚类数排名前 20（高多样性滤波器）和后 20（低多样性滤波器）的滤波器进行掩码实验，即将对应滤波器输出直接置零，观察模型性能变化。实验结果汇总于表 4-3 中，展示了掩码滤波器后模型损失的百分比变化情况。结果显示，当掩码聚类数较高的前 20 个滤波器时，模型损失显著增加，验证了这部分滤波器的特征多样性对于维持模型性能的重要性。这些滤波器通过编码丰富且差异显著的特征模式，有效支持了模型的判别能力与泛化性能。它们在整体特征空间中起到了关键的区分作用，缺失这些滤波器导致模型对输入信息的理解能力显著下降。相比之下，掩码聚类数较低的后 20 个滤波器对模型损失的影响较小，甚至有时仅表现为轻微的性能下降，表明这些滤波器的贡献相对有限。低聚类数对应滤波器内部神经元的特征表

达高度相似或重叠，可能存在一定的冗余性。这些滤波器学习的特征可能是对分类任务贡献较少的辅助信息，或者是模型训练过程中未充分利用的资源。

综上，本节实验证实了卷积层内部滤波器间存在显著的学习能力差异，高特征多样性的滤波器在模型性能维持中扮演着关键角色。

表 4-3　滤波器掩码后模型损失的百分比变化

模型	AlexNet		VGG-11		VGG-16		ResNet-18	
掩码滤波器	前 20	后 20	前 20	后 20	前 20	后 20	前 20	后 20
数据集 1	+108.3%	-12.8%	+93.4%	-1.9%	+95.2%	-0.01%	+78.2%	-2.6%
数据集 2	+84.9%	-8.7%	+90.5%	-10.3%	+89.6%	+0.01%	+88.3%	-0.4%
数据集 3	+112.5%	+0.03%	+92.8%	-2.7%	+88.7%	-3.1%	+82.2%	+1.3%

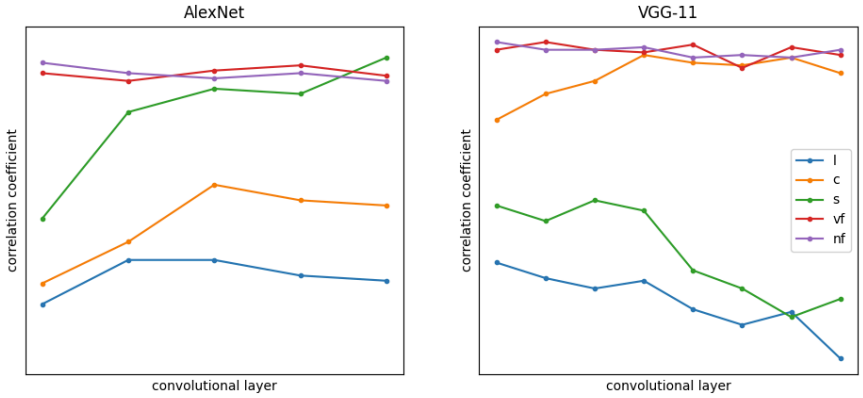
4.4.4　消融实验

为进一步揭示特征量各组成部分对模型性能的具体贡献，本节设计并实施了一系列系统性的消融实验。特征量被分解为五个主要组成元素：亮度（l）、对比度（c）、结构（s）、有效特征比例（vf）以及噪声特征比例（nf）。本节旨在分别评估每个组件在特征量构成中的独立影响，从而更深入理解它们在神经元特征表达和模型性能中的关键作用。实验以 AlexNet 和 VGG-11 两种经典 CNN 模型结构为基础，选取在 ILSVRC-2013 DET 数据集上的训练结果作为验证对象。该数据集具备多样且复杂的视觉类别，能够充分测试各特征量组成部分的表现差异。消融实验采取逐一隔离每个组成元素的策略，在保持其它元素不变的条件下，系统地调整某一组件的数值，观察其对模型性能，特别是方向一（特征量与激活值相关性）、方向二（特征量层次性变化）及方向三（滤波器特征多样性与性能关系）验证效果的具体影响。

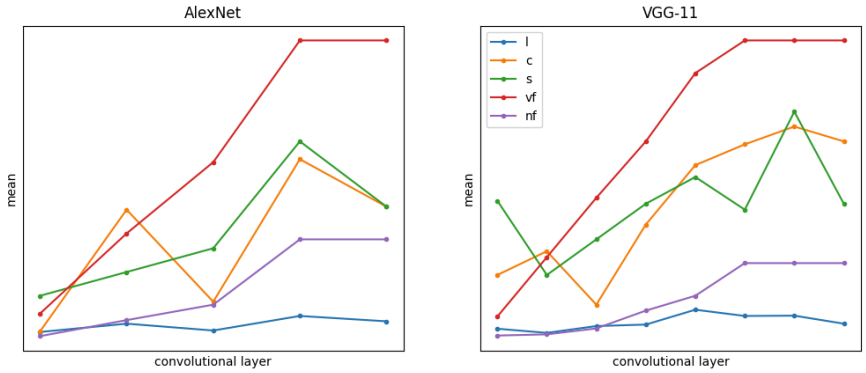
图 4-5 展示了方向一和方向二中各组件对结果的贡献变化趋势。从图中可以观察到，对比度和结构成分对相关系数的影响较为显著，尤其是在中高层卷积层，这表明输入的对比度信息和图像的空间结构是神经元激活及其特征量形成的重要因素。亮度成分的作用相对温和，但仍能在一定程度上改善特征多样性指标。有效特征比例的提升有助于增强模型对有意义信息的捕获，而噪声特征对模型性能影响较小，反映出特征纯度对学习效果的重要性。表 4-4 则总结了方向三中各组成部分对滤波器聚类数量及模型损失变化的影响。结果表明，没有单一组件能够完全独立地验证方向三的核心内容。各组件对滤波器特征多样性及整体

模型表现均产生一定影响，但其单独作用呈现出不一致性，未形成明确的性能提升或验证模式。这进一步表明，CNN 模型中神经元的特征量是多因素共同作用的结果，单一因素难以解释复杂的特征学习机制。

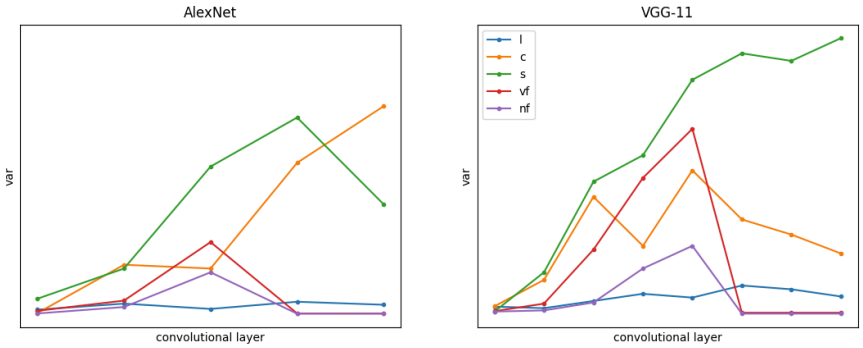
综合来看，消融实验明确证实了特征量各组成部分在整体特征表达中的关键地位。对比度、结构信息以及有效特征比例对模型性能起到主导作用。亮度作为辅助成分。该发现强调了在设计和优化神经网络特征表达时，应重视多维度特征的协调与平衡，而非单纯追求某一指标的最大化。



(a) Pearson 相关系数



(b) 均值



(c) 方差

图 4-5　特征量中不同元素对方向 1 和 2 结果的影响

表 4-4　掩码特征量中不同元素后模型损失的百分比变化

模型	AlexNet		VGG-11	
掩码滤波器	前 20	后 20	前 20	后 20
l	-32.7%	+23.4%	-0.18%	+67.3%
c	+1.4%	+3.5%	-19.6%	-8.7%
s	+34.1%	-10.4%	-4.2%	-0.12%
vf	+56.8%	+58.4%	+70.6%	+65.8%
nf	+56.8%	+58.4%	+70.6%	+65.8%

4.4.5　与其他可解释方法比较分析

为了评估不同神经网络解释方法对目标区域的定位能力，本研究采用 IOU（Intersection over Union）作为衡量显著图与原图中目标区域一致性的度量指标。IOU 是目标检测与分割任务中常用的评估标准，用于量化两个区域之间的重叠程度。

设显著图经过归一化与阈值化处理后得到的二值掩码为 M_{sal}，而对应图像的真实目标区域掩码为 M_{gt}，则其 IOU 定义为：

$$\text{IOU} = \frac{|M_{\text{sal}} \cap M_{\text{gt}}|}{|M_{\text{sal}} \cup M_{\text{gt}}|},$$

其中，$|\cdot|$ 表示区域的像素数或面积大小。IOU 值越高，表明显著图越能准确覆盖目标区域，反映出该可视化方法在空间上更具判别性。

为评估方法在定位精度方面的有效性，本节对所有神经元提取其对应特征量，并选择使输出层特征量最大的学习图像作为模型的代表性学习图像，并其与两种主流方法 CAM 和 Grad-CAM 进行了定量比较。结果如表 4-5 所示。结果显示，提出的方法在平均 IOU 上均优于 CAM 和 Grad-CAM。这表明利用特征量最大样本作为显著性基础图像，能够更集中地定位网络真实关注的目标区域。这表明提出的方法结构相关性与目标覆盖能力方面均优于传统显著图生成方式，验证了神经元特征量可作为度量模型学习能力的有效指标。

表 4-5　不同解释方法的 IOU

	数据集 1			数据集 2			数据集 3		
	CAM	Grad-CAM	**FQ**	CAM	Grad-CAM	**FQ**	CAM	Grad-CAM	**FQ**
AlexNet	0.08	0.17	**0.32**	0.09	0.23	**0.46**	0.08	0.22	**0.34**
VGG-11	0.11	0.24	**0.52**	0.14	0.19	**0.45**	0.09	0.21	**0.35**
VGG-16	0.17	0.31	**0.49**	0.16	0.36	**0.53**	0.07	0.15	**0.28**
ResNet-18	0.19	0.29	**0.58**	0.21	0.28	**0.34**	0.12	0.32	**0.44**

4.5 本章小结

本章提出了一种量化单个神经元学习到特征的方法，并通过该度量方法解释 CNN 模型中的卷积层。通过评估神经元从输入样本中学习到的特征来量化神经元的特征量。使用神经网络扫描仪为每个指定的神经元生成学习图像。并定义了一个新的数学度量，用于衡量学习图像与原始输入样本之间的相关性。以特征量作为指标从不同方向对卷积层内神经元工作机制的深入分析。该方法理论上适用于任何由人工神经元构成的神经网络架构。未来的工作计划将该方法扩展到其他架构，以进一步分析它们的机制。此外，本研究目前集中于分类任务，计划在后续研究中探索该方法在其他任务中的适用性。

第五章　基于 NNS 的滤波器层面解释

5.1　引言

当前具有量化能力的解释方法通常使用单个指标来评估性能，但这些指标本身的可解释性存疑。单一关注某个指标可能会过度简化模型中的复杂关系，从而可能导致解释不充分。因此需要更全面、更稳健的可解释性框架，整合分析的多个维度，从而更细致地了解模型如何运作和做出决策。

第三章提出的基于神经元的“神经网络扫描仪（*NNS*）”解释方法可以获得神经元学习到的特征的可视化结果。它为模型中的任何神经元提供被称为“学习图像”的解释结果，无需改变模型架构。由于学习图像与输入样本大小相同，可以在学习图像与输入样本之间进行细粒度比较，以评估学习到的特征。但该方法存在以下问题：1. 虽然可视化结果提供了直观的洞察，但它们通常需要人工解读才能得出有意义的解释结果。这种对人类判断的依赖可能引入主观性，限制了解释过程的有效性。2. 量化特征解释的方法通常使用单一度量来评估性能，而这些度量本身的可解释性可能值得质疑。单一关注某一度量可能会过于简化模型内部复杂的关系，进而导致部分解释。3. 缺乏针对神经网络模型更高阶单元——滤波器的解释，导致解释结果不够全面。第四章从神经元的定量解释出发，一定程度上解决了第一个问题。后两个问题仍需解决。针对这些问题，本章提出一种全面和稳健的可解释性框架，整合多维度的分析，从而更细致地理解模型是如何运作和做出决策的。

当前神经网络中的针对滤波器的可解释性方法通常分为两种类型：局部方法和全局方法[20]。局部方法侧重于解释特定输入，通常利用目标输入的信息（如特征值或梯度）为输入的不同区域或像素分配重要性分数，从而解释它们对模型输出的影响。例如，基于梯度的方法，如类激活映射[49]和集成梯度[117]，通过使用梯度信息生成解释结果。以此为基础，进一步实现许多通过梯度分析提供真实解释的方法[127-129]。与此同时，模型无关的方法，如 SHAP[130-131]和 LIME[132-133]，

提供独立于模型架构的可解释性策略。然而，这些局部方法通常难以提供对模型整体行为的全面解释。全局方法旨在通过识别和分析在输入数据集上具有普遍意义的特征，为整个模型提供统一的解释。这些方法描述了每个特征在多大程度上对模型输出的贡献，从而能够对模型有一个全面的理解。基于概念的方法[64,134]例如，探索人类定义的概念与模型预测之间的关系，尽管它们受限于人类主观性。数学方法将数学度量与模型分析结合，用系统化的方式评估和解释模型性能。信息瓶颈理论和傅里叶分析被用来分析模型[66,135]。局部方法提供有针对性的洞察，而全局方法则提供对模型的整体理解。现有方法缺乏普适的具有数学依据的解释结果，且缺乏能够整合局部和全局信息可解释性框架，从而更深入的分析模型的运行规律。

本章提出了一个统一的可解释性框架，称为模型可解释特性测量框架（*Test with Interpretable Properties, TIP*）。基于神经网络扫描仪方法对神经元学习特征的可视化结果，通过各种可解释特性（*Interpretable Properties, IPs*）来衡量滤波器（单元）学习到的特征，生成多个可度量的结果。如图5-1所示，TIP 通过学习图像获得滤波器的可解释特性。通过学习图像的统计特征及其与输入样本的可度量关系来获得模型的可解释特性。这些可解释特性是滤波器的固有属性，例如学习图像的内在特征或它们与原始图像的可度量关系。可解释特性具有强大的数学意义，消除了进一步解释的需要。接着，提出了三种类型的度量，旨在从不同角度定量分析和解释模型。通过利用不同的可解释特性，实现了对模型的多维度解释，从而对模型的进行更全面的理解。

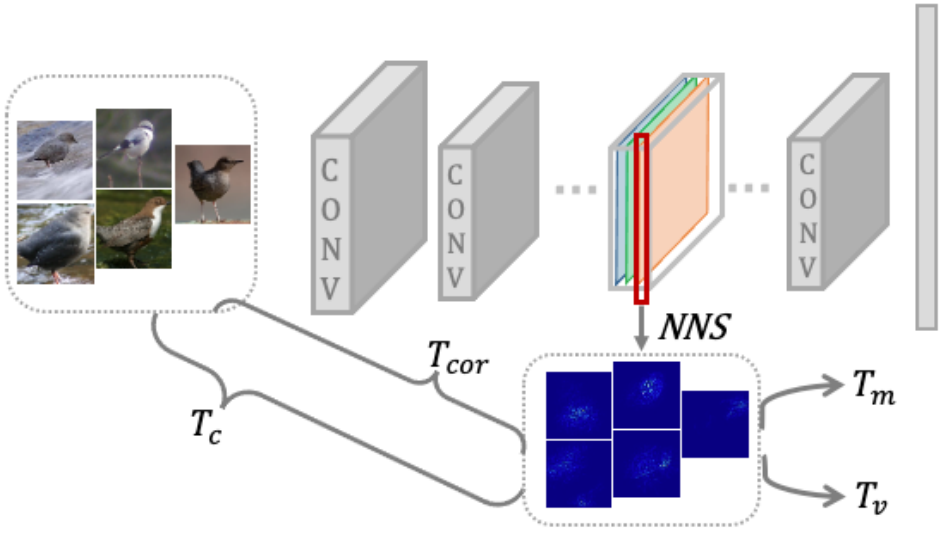


图 5-1　模型可解释特性测量框架（TIP）

本章的贡献总结如下：

- 本章提出了一个统一的可解释性框架 TIP，适用于解释各种类型的神经网络模型。该框架采用具有数学意义的可解释特性，阐明模型的运作过程。本章通过四个示例可解释特性介绍了 TIP。通过允许整合多个可解释特性，确保了该框架的可扩展性以及全面解释的潜力。
- 基于 TIP，本章提出了三种类型的度量，分别从三个方面解释模型：1. 评估不同可解释特性与模型性能的解释能力；2. 评估可解释特性阐明滤波器显著性度量的能力；3. 探讨模型的学习模式。
- 本章通过实验验证了所提出的 TIP 的有效性，并分析了三个关键问题：1. 哪些可解释特性最能解释模型性能？2. 表现良好的滤波器学习到了什么？3. 卷积层的目标是否是捕获前景信息？

5.2 基于 NNS 的模型可解释特性测量框架

本节介绍了如何生成模型可解释特性测量框架（TIP），以及如何使用可解释特性来量化模型的可解释性。并基于 TIP 提出了三种度量方法，从不同角度解释模型。

为简便起见，分析具有输入 $x \in \mathbb{R}^{N\times N}$ 的神经网络模型，并指定具有 K 个滤波器的前馈层。输入样本的数量为 S。特征图的大小为 $M \times M$。滤波器 k 的特征图为 $f^k(x) \in \mathbb{R}^{M\times M}$。

神经网络扫描仪（NNS）为每个神经元生成一个输入尺寸的样本。生成的样本被称为该神经元的“学习图像”。该学习图像以可视化的方式表示神经元从输入样本中提取的特征。神经网络扫描仪适用于模型的不同组件。对于样本 x 和指定的滤波器 k 的神经元 n，对应的学习图像 $I_n^k(x)$ 是基于输入神经元对该神经元贡献的计算结果生成的：

$$I_n^k(x) = \sum_{i,j\in[1,N]} c_{(i,j)}(x) \times I^0_{(i,j)}(x), \tag{5-1}$$

其中 $c_{(i,j)}(x)$ 是位置 (i,j) 处的输入神经元对神经元 n 的贡献值，$I^0_{(i,j)}(x)$ 是输入分辨率的图像，除 (i,j) 位置的像素外，其余像素值为零。

特征图揭示了滤波器在输入样本中学习到的特征。然而，特征图的大小通常不确定，容易丢失样本的细节信息。作为一种可以生成与输入大小一致的样本的方法，神经网络扫描仪不仅可以为神经元生成学习图像，还可以为滤波器生成学习图像。滤波器 k 的学习图像 $L^k(x)$ 是对应特征图中所有神经元学习图像的总和：

$$L^k(x) = \sum_{i,j\in[1,M]} I^k_{(i,j)}(x) \times f^k_{(i,j)}(x), \tag{5-2}$$

其中 $I^k_{(i,j)}(x)$ 是特征图 k 中位置 (i,j) 处神经元的学习图像，$f^k_{(i,j)}(x)$ 表示该位置的激活值。一个滤波器的学习图像与其对应的特征图高度相关。

本节通过神经网络扫描仪学习到的学习图像量化可解释特性。模型的可解释特性通常指的是与模型和输入样本本身固有的特征，这些特性独立于其他变量或外部因素。这些特性有助于解释和理解兴趣变量的特征或兴趣输出变量与输入样本之间的关系。换句话说，可解释特性解释了模块学习到的特征或学习到的特征与输入样本之间的关系。

给定一个预训练的 CNN、一组训练图像 X 和一个输入样本 $x \in X$，分析与滤波器 k 相关的可解释特性 P。主要从四个方面量化可解释特性。其中两个用于描述和总结滤波器学习图像的特征，另外两个是衡量学习图像与输入样本之间关系的属性。值得注意的是，所选的特性仅作为示例。通过所提出的方法，可以选择更多可解释特性以进一步分析模型。

1. 特征强度。

特征强度用于量化神经网络中滤波器的激活水平，从而深入了解其对输入样本的整体响应。学习图像的均值是所有像素值的平均值，表示整个图像的中心趋势或平均亮度。从统计学角度来看，均值给出了图像中像素值的整体水平。较高的均值表示滤波器学习到的特征具有较高的像素强度。

$$P^k_m(x) = \frac{\sum_{i,j\in[1,M]} L^k_{(i,j)}(x)}{M \times M}. \tag{5-3}$$

2. 特征多样性。

特征多样性揭示了过滤器响应的可变性，反映了其功能广度和选择性。它区分专用过滤器和广义过滤器，评估对其对输入的敏感性。用学习图像的方差描

述。方差描述了像素值的离散程度，即像素值围绕均值的波动程度。方差评估了学习图像的对比度。较高的方差表示学习图像具有较强的像素强度差异，意味着滤波器学习到的特征分布在较广的范围内。

$$P_v^k(x) = \frac{\sum_{i,j\in[1,M]}(L_{(i,j)}^k(x) - \mu_L)^2}{M \times M}, \tag{5-4}$$

其中 μ_L 是 $L^k(x)$ 的均值。在方程5-3中，$\mu_L(x)$ 和 $P_m(x)$ 都表示 L^k 的均值。将它们分别表示以保持不同可解释特性 $P_m(x)$ 和 $P_v(x)$ 的独立性。

3. 输入依赖性。

输入依赖性量化与输入样本的关联，揭示直接的特征依赖。它通过学习图像和输入样本之间的协方差解释了增强或抑制像素级的特征。协方差反映了两个变量之间联合变化的方向。学习图像和输入样本之间的协方差度量了对应位置的像素值如何共同变化。本质上，协方差评估学习图像和输入样本之间是否存在线性关系。如果协方差为正，表示在对应像素位置，当输入样本的像素值增加时，学习图像的像素值也倾向于增加，反之亦然。这表明学习图像与输入样本在焦点模式上具有相似性。协方差提供了一个指标，表示滤波器从输入样本中学习到的特征在像素强度模式上的相似性。

$$P_c^k(x) = \frac{\sum_{i,j\in[1,M]}(L_{(i,j)}^k(x) - \mu_L)(x_{(i,j)} - \mu_x)}{M \times M}, \tag{5-5}$$

其中 $x_{(i,j)}$ 是输入样本 x 在空间位置 (i,j) 处的像素值，μ_x 是 x 的均值。

4. 标准化输入相关性。

标准化输入相关性量化滤波器的输出和输入像素值之间的标准化线性关系。它便于过滤器的比较，揭示输入灵敏度和特征演变过程。与协方差类似，相关系数表示学习图像和输入样本之间线性关系的程度。不同之处在于，协方差仅关注两个变量之间线性关系的方向，而相关系数提供了方向和强度的综合信息。对于相同的输入样本 x，如果不同滤波器的 $P_c(x)$ 相同，则具有较大相关系数的滤波器意味着学习图像的像素值波动范围较小，即该滤波器学习到的特征在较窄的范围内分布。

$$P_{cor}^k(x) = \frac{\sum_{i,j\in[1,M]}(L_{(i,j)}^k(x) - \mu_L)(x_{(i,j)} - \mu_x)}{\sigma_L \times \sigma_x}, \tag{5-6}$$

其中 σ_L 是 $L^k(x)$ 的标准差，σ_x 是 x 的标准差。

通过可解释特性测量框架（TIP），使用模型的固有特性来解释整个类别输入样本上的模型行为。对于数据集 X，第 k 滤波器的可解释特性定义如下：

$$T^k = \frac{\sum_{x \in X} P^k(x)}{S}, \tag{5-7}$$

其中 S 是输入样本的数量。算法 5.1为使用单个滤波器作为示例的滤波器可解释特性的生成。

算法 5.1 滤波器可解释特性的生成

输入：输入数据集 X，样本数为 S，滤波器 k,$I^k_{(i,j)}(x)$ 是特征图 k 中位置 (i,j) 处神经元的学习图像，$f^k_{(i,j)}(x)$ 是该位置 (i,j) 的激活值，特征图大小为 $M \times M$

参数设置：$L^k(x)$ 为输入图像 x 的学习图像对滤波器 k 的学习图像

for $x \in X$ **do**

$L^k(x) = \sum_{i,j \in [1,M]} I^k_{(i,j)}(x) \times f^k_{(i,j)}(x)$

if 计算 Mean **then**

$P^k_m(x) = \frac{\sum_{i,j \in [1,M]} L^k_{(i,j)}(x)}{M \times M}$

end if

if 计算 Variance **then**

$P^k_v(x) = \frac{\sum_{i,j \in [1,M]} (L^k_{(i,j)}(x) - \mu_L)^2}{M \times M}$

end if

if 计算 Covariance **then**

$P^k_c(x) = \frac{\sum_{i,j \in [1,M]} (L^k_{(i,j)}(x) - \mu_L)(x_{(i,j)} - \mu_x)}{M \times M}$

end if

if 计算 Correlation coefficient **then**

$P^k_{cor}(x) = \frac{\sum_{i,j \in [1,M]} (L^k_{(i,j)}(x) - \mu_L)(x_{(i,j)} - \mu_x)}{\sigma_L \times \sigma_x}$

end if

end for

$T^k_m = \frac{\sum_{x \in X} P^k_m(x)}{S}$

$T^k_v = \frac{\sum_{x \in X} P^k_v(x)}{S}$

$T^k_c = \frac{\sum_{x \in X} P^k_c(x)}{S}$

$T^k_{cor} = \frac{\sum_{x \in X} P^k_{cor}(x)}{S}$

输出：滤波器 k 的可解释特性 T^k_m，T^k_v，T^k_c，T^k_{cor}

5.3 基于可解释特性的模型分析

如图 5-2所示，基于 TIP，本节提出了三种类型的度量，分别从三个方面解释模型，包括：评估不同可解释特性与模型性能的解释能力，评估可解释特性阐明滤波器显著性度量的能力，探讨模型的学习模式。

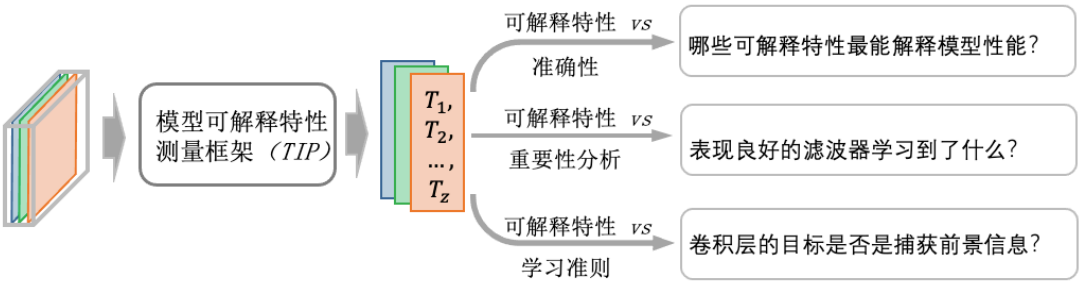


图 5-2 基于可解释特性的模型分析示意图

5.3.1 模型性能分析

本节探讨了可解释特性与模型准确性之间的关系，重点关注重要滤波器。其中一个主要的挑战是，对于某个滤波器，无法直接衡量其可解释特性与其重要性之间的关系，这使得获得精确的数值结果变得困难。通过对选择的滤波器进行掩码，并根据 TIP 对滤波器进行排序，逐步抑制每个滤波器对输入数据的响应，从而评估其对模型性能的贡献。

本节提出了一种数学度量方法来评估可解释特性：准确率可解释特性曲线（*AIPC*）。AIPC 根据指定的可解释特性对滤波器进行排序，然后按排序结果从高到低依次对滤波器进行掩码。AIPC 的 x 轴表示掩码滤波器占总滤波器的比例。y 轴表示在对指定滤波器进行掩码时，所有输入样本的准确性变化。

AIPC 下的面积用于衡量可解释特性的相对重要性。在这种意义上，本节的度量方法可以看作是接收操作特性（ROC）曲线下的面积的变种。

在相同的掩码比例下，准确率变化越大，滤波器的重要性越大。对于同一模型的不同可解释特性，AIPC 下的面积越大，准确率对该可解释特性的敏感度越高。这表明，AIPC 面积较大的可解释特性对模型更加重要。重要滤波器学习到的正是与该可解释特性相关的特征。

5.3.2　滤波器重要性分析

本节中分析了可解释特性与滤波器性能之间的关系。在神经网络模型中，一些显著性度量被用作评估滤波器性能的标准。这些度量衡量了决定滤波器如何与输入数据交互的特定行为。由于许多剪枝算法将这些度量作为模型优化的标准，因此这些度量通常被认为是重要滤波器的指示器。然而无法直接知道这些重要滤波器学习到了什么特征。本节的目标是探索这些显著性度量所指示的滤波器学习到的可解释特性。

经典的显著性度量包括滤波器的激活值和相对于滤波器的损失敏感性。激活值表示滤波器对给定输入样本的响应强度，提供了有关滤波器捕捉相关特征的有效性的信息。损失敏感性度量了滤波器对损失的影响，可以通过测量滤波器存在与否时损失的差异来计算，其他条件保持不变。使用对应滤波器特征图的平均激活值作为滤波器的激活值（AVoU）：

$$AVoU^k = \frac{\sum_{x \in X} \sum_{i,j \in [1,M]} f_{(i,j)}^k(x)}{M \times M \times S}, \tag{5-8}$$

其中 X 为输入样本集，M 为特征图的尺寸，S 为样本数量。

Lee et al.[136] 提出了一个标准，用于识别神经网络中的重要连接。它基于连接对损失函数的影响来确定每个连接的敏感度。使用类似的思路来计算不同滤波器的敏感性。

$$\triangle loss_w(x) = loss_w(\mathbf{1} \odot \omega) - loss_w((\mathbf{1} - \mathbf{e_w}) \odot \omega), \tag{5-9}$$

其中，$\mathbf{e_w}$ 是指示元素 w 是否存在的指示向量，$\mathbf{1}$ 是与 $\mathbf{e_w}$ 维度相同的向量。滤波器的敏感性（SoU）定义如下：

$$SoU^k = \frac{\sum_{x \in X} \sum_{w \in [1,W]} ||\Delta loss_w(x)||}{W \times S}, \tag{5-10}$$

其中 W 为滤波器的连接数。取损失变化的绝对值作为敏感性标准，因为如果损失变化较大，意味着该滤波器对损失有较大的影响。

检测显著性度量与可解释特性之间的关系对于解释模型的运行机制至关重

要。评估这种关系的一种有效方法是使用秩相关系数。与传统的假设线性关系并依赖于连续正态分布数据的相关度量不同，秩相关系数是非参数的，这使得它对离群值具有鲁棒性，且适用于顺序或非线性数据。

5.3.3 模型学习准则

本节的目标是研究在模型训练过程中，不同可解释特性所对应的滤波器学习的特征如何演化。具体来说，重点分析前景和背景可解释特性在不同训练时期的演化过程。滤波器根据其捕捉的前景和背景信息的量进行分类，可以将其划分为几种类型：高前景-高背景、高前景-非高背景、低前景-低背景、低前景-非低背景等。对于指定层中的可解释特性，$Tf, Tb \in \mathbb{R}^K$ 分别表示前景和背景中的 T 测量值。将这两个分布从高到低排序，并根据给定的条件识别每个分布中满足条件的元素。rf 和 rb 是排序后满足指定条件的元素索引集合：

$$\begin{aligned} rf &= \{i \in i, 2, .., K : Tf^i satisfies\ conditions\}, \\ rb &= \{i \in i, 2, .., K : Tb^i satisfies\ conditions\}. \end{aligned} \tag{5-11}$$

rfb 是 rf 和 rb 的交集。条件和相应的 rfb 被用于分类滤波器。例如，对于高前景-高背景类型的滤波器，条件是基于排序顺序，从每个分布中选出排名前 $per\%$ 的元素。

评估分为两大部分：

1. 滤波器类型的分布分析。检查在训练过程的不同阶段，各种滤波器类型的分布情况，即 $\frac{|rfb|}{K}$。这种分类可以分析每种滤波器类型随时间的数量和比例，从而揭示不同特征在训练不同阶段的突出趋势。

2. 不同滤波器类型对模型准确性的影响。通过屏蔽特定类型的滤波器并重新计算模型的准确性，评估每种滤波器类型对整体性能的贡献。准确性的变化能够评估每种类型滤波器对模型性能的贡献程度。这种双重分析将提供关于滤波器可解释性的深入理解，特别是在前景和背景特征表示如何影响模型在不同训练阶段的性能方面。

5.4　实验结果与分析

5.4.1　实验设置

为了确保全面的比较，实验应用于在 ILSVRC-2013 DET 数据集、CUB-200-2011 数据集和 Pascal VOC 2012 数据集上训练的 AlexNet、VGG-11 和 VGG-16 模型。由于实验的重点是模型性能的变化，而非特定的准确性值，因此在多个类别上进行了分类实验，并将输入尺寸设置为 32×32，以减少计算负担。将目标检测框定义为前景区域，框内的像素视为前景，框外的像素视为背景。本节的重点是推导与最后卷积层相关的通用原则，最后卷积层由卷积函数和激活函数组成。在下表中，将 ILSVRC-2013 DET 数据集、CUB-200-2011 数据集和 Pascal VOC 2012 数据集分别定义为数据集 1、数据集 2 和数据集 3。

5.4.2　基于可解释特性的可视化结果分析

为了验证提出的基于可解释特性的统一解释框架 TIP 的有效性，本节采用直观的可视化方法对模型内部滤波器学习的特征进行分析。利用不同的可解释特性对滤波器进行排序，通过展示每个属性下排名最高的滤波器所学习的图像特征，来揭示 TIP 框架在多种模型和数据集上的适用性和解释能力。

图5-3中展示了基于本研究中定义的四种主要可解释特性的滤波器排序结果，并分别列出了对应于每种属性中排名第一的滤波器所激活的学习图像。这些图像通过反向映射技术，将滤波器在输入空间中的响应特征可视化，直观反映了滤波器所捕获的图像局部或整体特征模式。通过这种方式能够从多个角度审视模型对输入信息的处理机制，并验证不同可解释特性所代表的解释维度。实验结果显示，不同的可解释特性对滤波器的关注点存在显著差异。例如，$\boldsymbol{T}_m$ 侧重于滤波器对纹理细节的捕捉，而 $\boldsymbol{T}_v$ 则更强调对整体形状或语义区域的响应。排名靠前的滤波器在其对应的属性视角下，往往能够准确聚焦输入图像中关键且有代表性的特征区域，如物体边缘、颜色分布或者特定的结构模式。这种高度一致性表明，TIP 框架所定义的可解释特性能够有效反映滤波器的功能角色，并为模型解释提供了科学的量化基础。

此外，通过对比不同属性下的滤波器学习图像，发现每种可解释特性不仅仅代表一种单一的解释维度，而是从不同层面揭示了模型的学习机制。例如，$\boldsymbol{T}_c$ 和

T_{cov} 突出滤波器的局部敏感性，反映模型对细节的捕捉能力；而 T_m 则倾向于滤波器的全局感受野特性，体现模型对整体结构的理解。通过这些多角度的解释视图，研究者可以更全面地理解模型的内部表示，避免单一视角导致的片面认知。这种基于 TIP 的可视化分析还揭示了滤波器学习特征的多样性和复杂性。不同滤波器在不同属性指导下表现出截然不同的关注焦点，说明模型通过多样化的特征提取器协同工作，以捕获输入数据中丰富的模式信息。更重要的是，选择不同的可解释特性会显著影响对模型行为的认知和解释，强调了设计合理解释指标的重要性。TIP 框架提供的多种属性指标，为理解和诊断深度模型提供了多维度的视角。

综上所述，通过基于 TIP 框架的可解释特性排序与可视化方法，不仅验证了 TIP 在模型解释方面的有效性，也展示了多视角解释策略在神经网络理解中的重要作用。

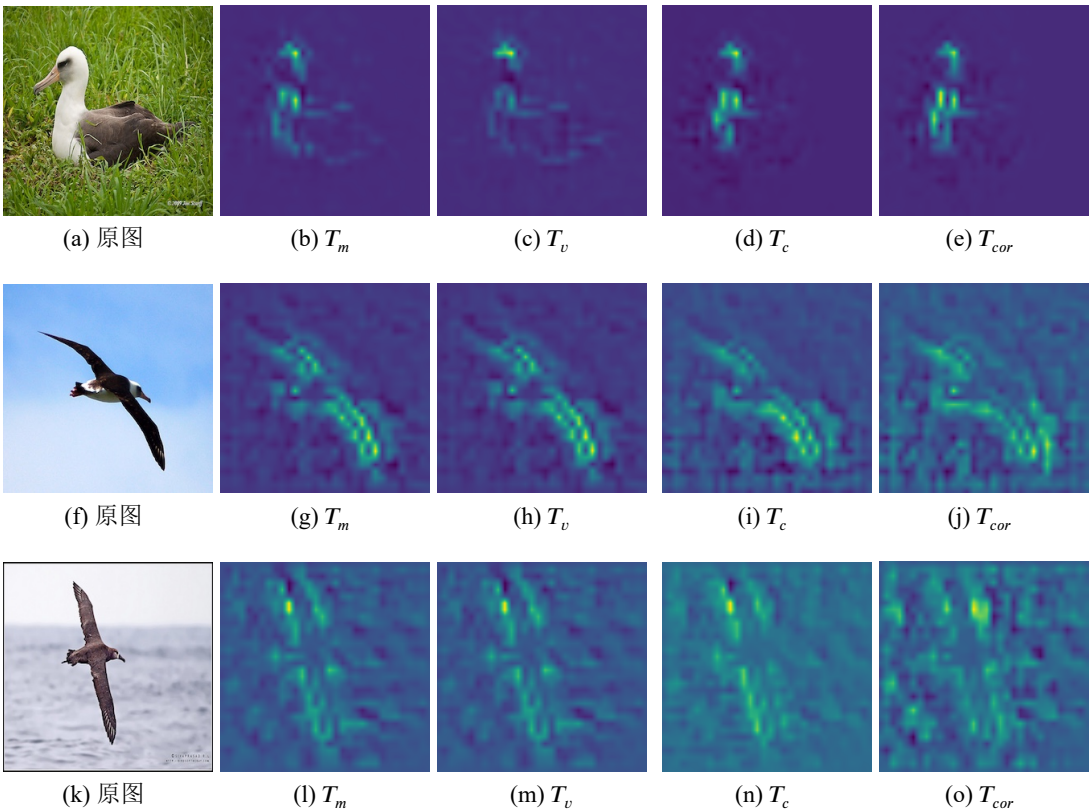


图 5-3 不同可解释特性排序下，排名最高滤波器对应的学习图像

5.4.3　基于可解释特性的模型分析

1. 哪些可解释特性最能解释模型性能？

本节旨在深入探讨可解释特性在多大程度上能够解释神经网络模型的准确性表现，特别关注那些通过 TIP 框架识别出的显著滤波器所学习到的可解释特征对整体模型性能的贡献。通过量化各 IP 与模型准确性之间的关系，揭示不同属性在模型性能解释中的有效性和重要性。为了实现这一目标，依据不同的可解释特性对模型中的滤波器进行排序，并设计了准确性可解释特性曲线（AIPC）。该曲线以掩码滤波器的比例为横轴，模型准确率作为纵轴，反映了在逐步掩盖排名靠前的滤波器时模型性能的变化。具体而言，在给定的掩码率下，模型准确性的下降幅度用以衡量被掩盖滤波器的重要性，准确性变化越显著，说明该属性所排序的滤波器对模型性能的贡献越大。进一步地，通过计算 AIPC 曲线下的面积，量化每个可解释特性与模型准确性之间的相关程度。AIPC 面积越大，意味着模型准确率对该属性排序的敏感度越高，也即该属性在解释模型性能方面越具有代表性和价值。

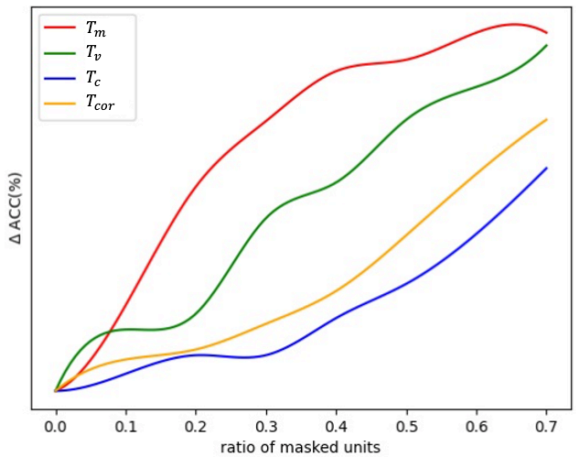


图 5-4　在 ILSVRC-2013 DET 数据集上训练的 VGG-16 模型的 APIC 值

图5-4展示了基于 TIP 框架的 AIPC 结果，以 VGG-16 模型在 ILSVRC-2013 DET 数据集上的表现为例。在图中，随着掩码滤波器比例的增加，模型准确性呈现不同程度的下降趋势。然而值得注意的是，在某些特定的掩码比例下，不同可解释特性的 AIPC 数值差异大，表明不同 IP 在不同滤波器数量级下的解释能力存在差异。这一现象反映出，当仅掩盖较少滤波器时，T_v 能够有效识别出对模型性能至关重要的核心滤波器；而当掩盖范围扩大，T_m 则可能更好地反映那

些对模型影响较小但数量众多的滤波器。为了进一步验证该结论，图5-5中绘制了多种神经网络架构上的 AIPC 曲线，以分析不同可解释特性与各模型准确性之间的相关性。结果显示，对于大多数模型，随着掩码滤波器的比例变化，可解释性属性的重要性也会发生变化。当掩码率较低时，重要的可解释性属性表明它们能够识别出少数对模型准确性有显著影响的重要滤波器。另一方面，当掩码率较高时，表现良好的可解释性属性表明它们对最重要滤波器的敏感度较低，但能够捕捉到较不重要的滤波器。这种行为突显了可解释性属性的能力，不仅能够识别关键滤波器，还能提供关于模型中各滤波器相对重要性的洞察。

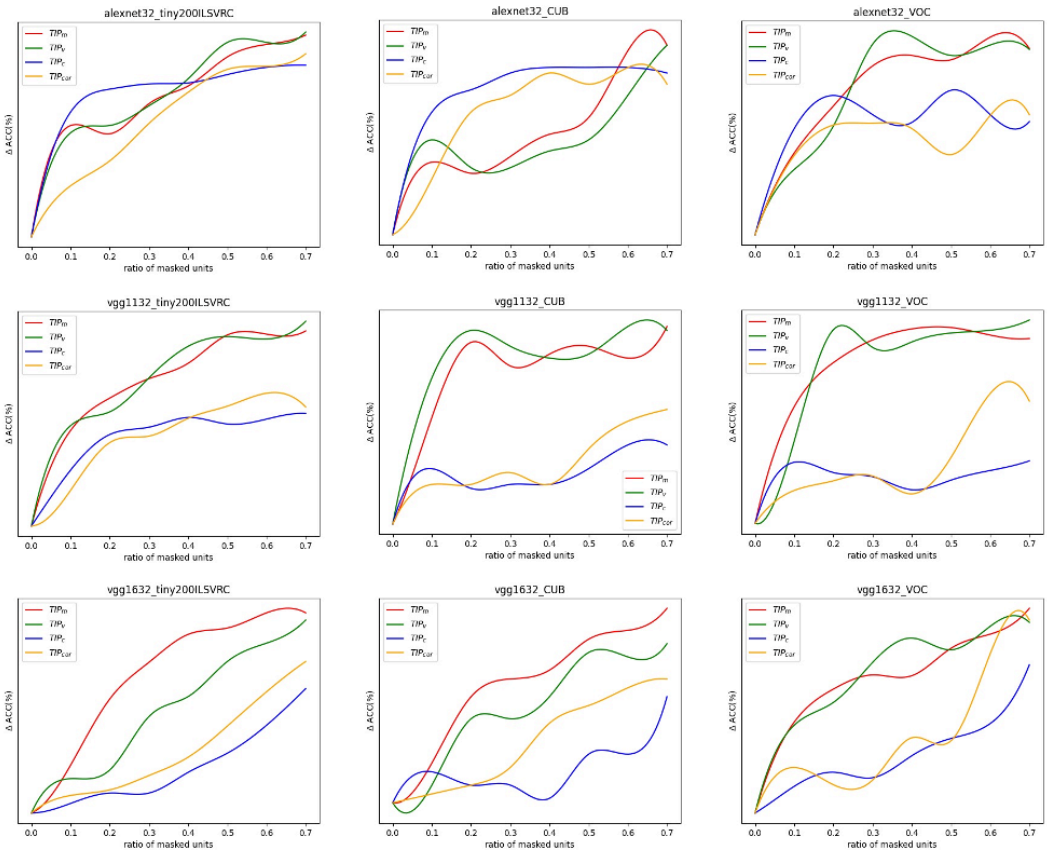


图 5-5 不同模型的 AIPC 值

针对关键滤波器的解释能力，本节进一步聚焦基于 TIP 排序中排名前 10% 的滤波器群体，统计分析了不同模型的 AIPC 面积，结果详见表5-1。面积数值较大的 IP 被视为对模型性能解释更为重要。具体而言，对于经典的 AlexNet 模型，T_c 指标表现出最高的 AIPC 面积，表明其对模型准确性最具解释力。相比之下，在 VGG 系列模型中，T_m 指标通常表现出最大的 AIPC 面积，暗示其对 VGG 模型的性能贡献更为显著。这一差异性分析进一步验证了 TIP 框架中不同可解释

特性的特定适用性。对于 AlexNet 结构，T_c 这一属性捕捉到了更具代表性的滤波器学习特征，成为理解其性能的关键指标；而对于深层的 VGG 网络，T_m 作为衡量滤波器某种特征的属性，更能揭示模型准确性的本质关联。

表 5-1　APIC 曲线下的面积

	AlexNet				VGG-11				VGG-16			
	T_m	T_v	T_c	T_{cor}	T_m	T_v	T_c	T_{cor}	T_m	T_v	T_c	T_{cor}
数据 1	41.20	41.40	**52.40**	24.40	**40.60**	19.40	7.40	10.20	**55.20**	53.00	36.40	29.40
数据 2	31.51	26.05	**36.72**	28.26	**35.67**	32.94	11.20	12.24	**52.61**	52.34	21.23	14.32
数据 3	18.46	22.31	**36.92**	24.61	**25.38**	17.70	8.46	4.62	56.15	**66.15**	17.69	15.38

综上所述，本节的 AIPC 分析不仅定量体现了 TIP 框架下各可解释特性在模型性能解释上的贡献差异，还从实证角度支持了针对不同网络架构选择合适 IP 的重要性。这为未来基于可解释特性的模型优化和诊断提供了理论与方法上的参考，也为深度神经网络的透明化与可控性研究奠定了坚实基础。

2. 表现良好的滤波器学习到了什么？

本节重点研究表现良好的滤波器所学习到的特征，尤其是基于若干显著性度量对滤波器功能特征的刻画。显著性度量通常通过衡量滤波器的激活值大小或其对模型损失函数的敏感度，反映滤波器在模型决策过程中所扮演的重要角色。通过分析这些度量与可解释特性之间的关系，旨在为基于可解释特性的滤波器功能理解提供更深入的洞见，推动两者的融合应用。

具体来说，采用公式5-8和公式5-10分别计算滤波器的平均激活值（AVoU）和敏感度（SoU），以量化滤波器的响应强度和对模型输出的影响敏感程度。AVoU 衡量滤波器在输入样本中的平均激活水平，反映滤波器对输入的整体响应能力；而 SoU 则体现滤波器对模型损失函数的梯度敏感度，揭示其在模型学习优化过程中的重要性。

为了客观衡量显著性度量与各可解释特性之间的关联程度，采用秩相关系数，其计算公式如下：

$$\rho = 1 - \frac{6\sum d_i^2}{K(K^2-1)}, \tag{5-12}$$

其中，d_i 为第 i 个滤波器在显著性度量与可解释特性排序中的秩差，K 代表滤波器的总数量。秩相关系数 ρ 的取值范围为 [-1,1]，值越接近 1 表明二者排序一致性越高，即显著性度量越能反映可解释特性所度量的滤波器功能特征。反之，较低或负相关系数则表明二者相关性较弱。

表5-2汇总了不同模型和数据集上显著性度量与多种可解释特性之间的秩相关系数分析结果。整体来看，AVoU 与可解释特性 T_m 表现出较强的正相关，表明具有高平均激活值的滤波器往往学习到 T_m 所代表的高整体均值特征。这种高度相关性说明，T_m 属性能够有效捕捉滤波器的激活强度特征，反映其在模型中的活跃程度。同时，SoU 与可解释特性 T_m 及 T_v（表示特征分布方差的属性）也呈现较强相关性。敏感度高的滤波器通常学习到具有较高均值和广泛分布的特征，说明其在模型优化过程中对损失函数的变化极为敏感，具有重要的功能意义。值得特别指出的是，T_m 与 AVoU 之间的显著相关性，不仅验证了激活值在滤波器重要性评估中的有效性，也为基于可解释特性的滤波器分析提供了坚实的实证支持。该发现表明，通过量化滤波器学习特征的统计特性，可以间接反映其激活模式和对模型输出的影响力，这为未来结合可解释特性与显著性度量的滤波器选择和模型压缩方法提供了理论依据。此外，本节分析进一步揭示，不同的可解释特性在反映滤波器功能特征时存在差异性，部分属性更侧重滤波器的激活水平，而另一些则更强调特征分布的统计特征。通过结合多种属性与显著性度量，可以更全面地理解滤波器的功能角色，从而促进更精细的模型解释。

综上所述，本节通过秩相关系数定量分析了显著性度量与可解释特性之间的关联，验证了 TIP 框架下可解释特性在捕捉滤波器功能特征方面的有效性。

表 5-2 显著性度量与可解释性属性之间的相关系数

	AlexNet				VGG-11				VGG-16			
	T_m	T_v	T_c	T_{cor}	T_m	T_v	T_c	T_{cor}	T_m	T_v	T_c	T_{cor}
AVoU												
数据 1	**0.99**	0.97	-0.46	-0.50	**0.99**	0.98	-0.40	-0.32	**0.98**	0.84	-0.54	-0.37
数据 2	**1.00**	1.00	-0.09	-0.51	**1.00**	0.98	-0.16	-0.12	**0.98**	0.88	-0.47	-0.32
数据 3	**0.99**	0.95	0.22	0.06	**1.00**	0.99	-0.54	-0.43	**0.97**	0.83	-0.42	-0.22
SoU												
数据 1	0.74	**0.77**	-0.21	-0.26	0.83	**0.84**	-0.29	-0.22	**0.57**	0.56	-0.29	-0.16
数据 2	**0.92**	**0.92**	-0.09	-0.49	0.77	**0.78**	-0.23	-0.22	0.58	**0.61**	-0.23	-0.14
数据 3	0.75	**0.79**	0.18	0.10	**0.83**	**0.83**	-0.42	-0.33	**0.57**	0.47	-0.18	-0.06

3. 卷积层的目标是否是捕获前景信息？

一个表现优异的神经网络模型，通常能够准确识别输入样本中的关键相关区域，从而实现精确的分类。这一现象引发了关于卷积层中各滤波器功能分工的核心问题：卷积层的滤波器是否主要专注于学习图像中的前景信息，而较少关注背景？另外，那些专注于前景特征的滤波器是否在模型的整体性能中扮演更关

键的角色？为了探究这些问题，本节基于 TIP 框架，分析卷积层滤波器在训练过程中的前景和背景特征学习动态，重点考察它们对模型性能的贡献差异。

本节以 T_m 作为可解释特性的代表，定量分析 VGG-16 模型在细粒度图像分类数据集 CUB200-2011 上的滤波器特征学习情况。T_m 反映了滤波器所学习特征的整体均值，是衡量滤波器激活强度和响应能力的重要指标。通过将滤波器在前景和背景区域学习到的特征分别量化，并基于这两个维度对滤波器进行排序，划分出“高信息滤波器”和“低信息滤波器”两个群体，分别对应排序中前 30% 和后 30% 的滤波器。

进一步地，为深入理解滤波器的功能侧重，将滤波器分为六种类型：高前景-高背景滤波器（同时对前景和背景特征学习强烈）；高前景-非高背景滤波器（专注于前景特征）；非高前景-高背景滤波器（专注于背景特征）；低前景-低背景滤波器（对两者响应均弱）；低前景-非低背景滤波器；非低前景-低背景滤波器。模型训练过程被划分为三个阶段：阶段 1 代表训练刚开始，阶段 2 为训练中期，阶段 3 为训练完成。通过比较不同阶段中各类型滤波器所占比例，动态观察滤波器对前景与背景特征的关注度变化。

首先，针对第一个问题，即卷积层滤波器是否更侧重于前景特征，统计高信息滤波器群体中不同类型滤波器的比例变化。表5-3显示，在所有训练阶段，高前景-高背景滤波器的比例始终位居前列，且明显高于仅关注前景或仅关注背景的滤波器。低信息滤波器中也呈现类似趋势。这一现象表明，绝大多数滤波器并未严格区分前景与背景信息，而是同时对两者具有较强的响应能力，表明滤波器在学习过程中倾向于捕捉多样化的图像特征，而非单纯聚焦于前景。进一步分析不同训练阶段的滤波器分布，发现随着训练的深入，专注于高前景和高背景特征的滤波器比例并未显著改变。这表明在训练初期，滤波器即已形成对图像多区域信息的共同关注，这种多元化的特征学习策略有助于模型构建更为丰富的表征空间，提升对复杂图像内容的解析能力。

对于第二个问题，分析专注于前景特征的滤波器是否对模型性能更具影响力。本节通过掩码不同类型的滤波器，量化它们对模型准确性的影响，从而评估其重要性。以 T_m 为基础，对 VGG-16 模型在 CUB200-2011 数据集上的训练过程进行分析，并关注前述六类滤波器在不同训练阶段对性能的贡献。

表5-3展示了在模型训练阶段 1、阶段 2 和阶段 3 分别掩码特定类型滤波器

表 5-3 滤波器的分布和重要性

	Proportion of units			Change of accuracy		
	Phase1	Phase2	Phase3	Phase1	Phase2	Phase3
H-H	**0.28**	**0.28**	**0.28**	**17.4**	**22.00**	**26.00**
H-noH	0.03	0.03	0.03	0.40	0.40	0.80
noH-H	0.03	0.03	0.03	1.60	1.40	1.80
L-L	**0.29**	**0.29**	**0.29**	**0.00**	**0.40**	**0.00**
L-noL	0.02	0.02	0.02	0.20	0.20	0.80
noL-L	0.02	0.02	0.02	0.00	0.20	0.00

注释：此表按顺序展示了具有高前景-高背景、高前景-非高背景、非高前景-高背景、低前景-低背景、低前景-非低背景和非低前景-低背景的滤波器。

后的模型准确性变化情况。结果表明，掩码高前景-高背景滤波器所导致的性能下降显著高于掩码低前景-低背景滤波器。这种趋势在三个训练阶段中都非常稳定，尤其是在训练后期（阶段 3），该差异最为明显。相比之下，掩码低前景-低背景滤波器对模型性能的影响则较小，甚至在部分情况下近似于无效掩码，说明这类滤波器对模型分类结果的贡献相对有限。此外，随着训练轮次的增加，掩码高前景-高背景滤波器所引起的准确性下降逐渐增大。这一趋势表明，这类滤波器的学习能力在训练过程中不断增强，所提取的特征对模型最终决策的重要性持续上升。它们不仅在训练初期就具备一定的判别能力，而且在后续迭代中进一步强化了这种能力，成为支持模型性能的关键构件。相对而言，高前景-非高背景滤波器并未表现出类似的行为。尽管这类滤波器在前景区域具有较强响应，但其对整体准确性的影响相较于高前景-高背景滤波器而言要小得多，且这种影响在训练过程中变化不大。类似地，非高前景-高背景滤波器同样未对模型性能产生显著的影响。这些结果说明，滤波器的有效性不仅取决于其是否关注前景信息，还与其是否能在前景和背景之间建立起有效的特征联系密切相关。

综合分析表明，单纯关注前景信息的滤波器并不一定就是最重要的；相反，那些能够在前景和背景之间进行联合建模的高前景-高背景滤波器才是真正影响模型性能的关键因素。这些滤波器具有更强的语义建模能力，能够捕捉到图像中多个区域之间的结构性联系，因此其掩码会显著破坏模型的判别能力。

为进一步深入讨论进行扩展实验，在不同可解释性属性下分析不同模型结果。表5-4、表5-5、表5-6中的实验结果显示，对于不同的模型和不同的可解释性属性，大多数滤波器倾向于同时学习前景和背景信息。此外，具有高前景和高背景信息的滤波器对模型性能的影响大于其他类型的滤波器。这一发现突显了捕

表 5-4　滤波器滤波器的分布和重要性（AlexNet 在 ILSVRC-2013 DET 数据集训练）

		Proportion of units			Change of accuracy		
		Phase1	Phase2	Phase3	Phase1	Phase2	Phase3
H-H	T_m	**0.28**	**0.28**	**0.28**	**15.8**	**20.6**	**22.6**
H-noH		0.02	0.02	0.02	0.00	0.60	0.60
noH-H		0.02	0.02	0.02	0.40	2.80	1.40
L-L		**0.29**	**0.29**	**0.29**	**-0.20**	**-0.20**	**-0.60**
L-noL		0.01	0.02	0.01	0.00	0.00	-0.40
noL-L		0.01	0.02	0.01	0.00	-0.40	0.40
H-H	T_v	**0.28**	**0.28**	**0.28**	**18.20**	**20.40**	**22.40**
H-noH		0.02	0.02	0.02	-0.40	0.80	-0.20
noH-H		0.02	0.02	0.02	0.60	1.80	1.60
L-L		**0.29**	**0.29**	**0.29**	**0.00**	**-0.20**	**0.20**
L-noL		0.02	0.02	0.01	0.00	0.20	0.20
noL-L		0.02	0.02	0.01	0.20	-0.60	0.20

注释：此表按顺序展示了具有高前景-高背景、高前景-非高背景、非高前景-高背景、低前景-低背景、低前景-非低背景和非低前景-低背景的滤波器。

表 5-5　滤波器滤波器的分布和重要性（VGG-11 在 CUB-200-2011 数据集训练）

		Proportion of units			Change of accuracy		
		Phase1	Phase2	Phase3	Phase1	Phase2	Phase3
H-H	T_m	**0.21**	**0.21**	**0.23**	**15.38**	**19.23**	**22.31**
H-noH		0.09	0.10	0.08	3.85	2.31	1.54
noH-H		0.09	0.10	0.08	4.62	6.92	6.92
L-L		**0.24**	**0.23**	**0.24**	**-0.77**	**1.54**	**-1.54**
L-noL		0.07	0.07	0.06	0.77	0.77	1.54
noL-L		0.07	0.07	0.06	-1.54	0.77	0.00
H-H	T_v	**0.26**	**0.23**	**0.23**	**14.62**	**22.31**	**23.80**
H-noH		0.05	0.07	0.07	6.15	6.15	7.69
noH-H		0.05	0.07	0.07	-2.31	1.54	6.15
L-L		**0.26**	**0.26**	**0.27**	**-2.31**	**-0.77**	**-0.77**
L-noL		0.05	0.05	0.04	0.77	3.08	0.77
noL-L		0.05	0.05	0.04	0.00	-0.77	0.00

注释：此表按顺序展示了具有高前景-高背景、高前景-非高背景、非高前景-高背景、低前景-低背景、低前景-非低背景和非低前景-低背景的滤波器。

捉前景和背景特征的滤波器的重要性，表明这些滤波器在模型整体决策过程中更具影响力。

5.4.4　与其他可解释方法比较分析

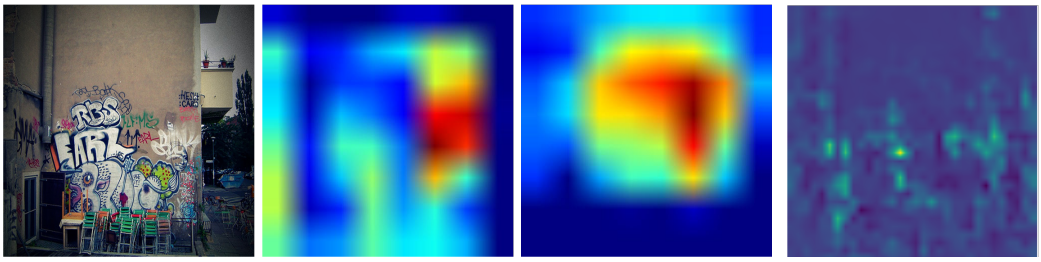
本节对当前解释方法与 CAM 和 Grad-CAM 这两类广泛应用的解释方法进行可视化比较，CAM 和 Grad-CAM 通过突出输入图像中对分类结果贡献最大的区域，帮助理解模型的决策机制。然而，这些方法通常只提供粗粒度的热图信息，较难揭示模型对细节特征的关注情况。相比之下，TIP 方法在细粒度解释方面展现出显著优势。本节采用相关系数可解释特性的可视化结果，实验结果如

表 5-6 滤波器滤波器的分布和重要性（VGG-16 在 Pascal VOC 2012 数据集训练)

		Proportion of units			Change of accuracy		
		Phase1	Phase2	Phase3	Phase1	Phase2	Phase3
H-H	T_m	**0.27**	**0.27**	**0.28**	**-0.65**	**5.34**	**13.54**
H-noH		0.04	0.04	0.03	0.13	0.13	2.60
noH-H		0.04	0.04	0.03	-1.17	0.52	0.52
L-L		**0.25**	**0.25**	**0.25**	**6.64**	**4.95**	**6.64**
L-noL		0.05	0.05	0.05	2.34	0.26	2.73
noL-L		0.05	0.05	0.05	0.52	-1.04	1.04
H-H	T_v	**0.27**	**0.27**	**0.27**	**4.56**	**8.33**	**18.49**
H-noH		0.04	0.03	0.03	-0.65	2.86	0.00
noH-H		0.04	0.03	0.03	-2.60	-5.47	2.34
L-L		**0.25**	**0.25**	**0.26**	**3.91**	**1.43**	**2.73**
L-noL		0.05	0.05	0.05	0.65	2.08	4.17
noL-L		0.05	0.05	0.05	0.52	-0.78	1.69

注释：此表按顺序展示了具有高前景-高背景、高前景-非高背景、非高前景-高背景、低前景-低背景、低前景-非低背景和非低前景-低背景的滤波器。

图5-6所示，与 CAM 和 Grad-CAM 相比，TIP 能够更清晰地定位图像中语义上关键但区域上较小的细节特征，例如小物体、边缘轮廓以及与文本描述精确对齐的区域。这一能力使 TIP 提供了更具针对性的可视化解释。



(a) 原图 (b) CAM (c) Grad-CAM (d) TIP

图 5-6 不同解释方法的可视化比较

5.5 本章小结

本章提出了一个通用且可扩展的可解释性框架——TIP，旨在通过模型固有的、可度量的和可解释的可解释特性来解释神经网络模型。首先，使用神经网络扫描仪生成了滤波器的学习图像。随后，从不同的角度评估了模型的各种可解释特性。基于 TIP 提出了三个度量标准，以多维度解释模型的工作机制。通过图像分类任务验证的实验结果表明了方法的有效性。TIP 提升了对模型行为和性能的理解，有助于提高神经网络的可解释性。

TIP 在理论上适用于任何由人工神经元组成的神经网络架构。在未来的工作

中计划将该方法扩展到其他架构，以进一步解释它们的工作机制。可解释特性的计算是并行进行的，这使得 TIP 能够扩展以包含更多的可解释特性。该方法为模型剪枝提供了新的视角，利用 TIP 度量评估滤波器的重要性，从而基于解释结果进行有针对性的剪枝。

第六章　基于 NNS 的模式层面解释

6.1　引言

现有的可解释性方法在局部或特征层面提供了一定程度的模型解释，但这类解释缺乏全局结构性，难以揭示神经网络模型在更高层次上从输入样本所学习到的模式。

第三章提出的基于神经元的“神经网络扫描仪（*NNS*）”解释方法可以获得神经元学习到的特征的可视化结果。它为模型中的任何神经元提供称为“学习图像”的解释结果，无需改变模型架构。作为一种可以生成与输入大小一致的样本的方法，神经网络扫描仪不仅可以为神经元生成学习图像，还可以为滤波器生成学习图像。基于神经网络扫描仪的结果，第四章和第五章提出度量方法和框架，分别从神经元的定量解释和滤波器的定量解释入手解释模型的运行机制。现有的可解释性方法在局部或特征层面提供了一定程度的模型解释，但这类解释缺乏全局结构性，难以揭示神经网络模型在更高层次上在输入数据中所学习到的模式。本章同样基于神经网络扫描仪的结果，从一个更大的解释目标入手探究模型。通过对滤波器的学习图像进行聚类，从而实现模式级的模型解释，并探索聚类结果与模型过拟合现象的关系。

CNN 中的滤波器通常是被认为是特征提取器[82,137]。这些提取的特征不能直接为人类所理解，需要进一步处理，诸如可视化和谱分析等技术可以帮助人类更好地理解这些特征[138-139]。通常，提取的特征是多方面且错综复杂的，随着模型深度的增加，其复杂性也会增加[82-83]。此外，有研究指出，低层的滤波器往往提取一些基本特征，如形状、颜色或纹理，这些特征在不同类别之间是共有的。而深层的滤波器则趋向于学习抽象的概念，包括特定物体（如眼睛或身体部位），这些概念通常更具类别特异性，并在 CNN 决策结果中起着重要作用[140]。

此外，众多研究[141][142][143]已证明 CNN 中存在滤波器冗余。这表明，可以评估每个滤波器的贡献，并去除那些不重要甚至影响模型性能的滤波器。一类

方法通过幅值来评估滤波器的作用，如 L1 范数[141]、L2 范数[144] 或批量归一化因子[145]。这些方法直接明了，但缺乏细致的分析。另一类研究则聚焦于每个滤波器的特征图。研究[146] 分析了特征图的统计数据，评估了学习模式的多样性与相似性。还有一类研究[147][148] 倾向于根据滤波器的激活模式对滤波器进行聚类，从而为每个聚类选择代表滤波器，进而压缩整个网络。与此不同，本章的方法在每个滤波器内部进行聚类，分析其在不同输入样本上的激活分布。这种滤波器内的聚类分析提供了更细粒度的视角，能够识别出在传统滤波器层级分析中可能被忽视的冗余或过拟合模式。它还分析了特征如何在内部进行表示和共享，提供了一个理解滤波器的新视角。

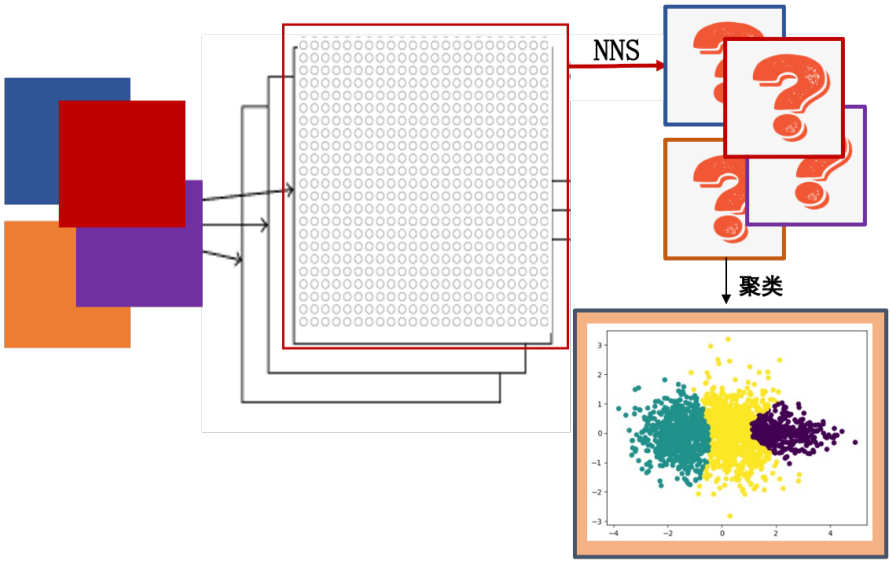


图 6-1　通过 GMM 对单个滤波器的所有学习图像聚类

本章通过探索滤波器的学习图像的聚类结果与模型过拟合之间的关系来解释 CNN。给定一个预训练的模型，本章对与特定滤波器对应的所有学习图像进行聚类并可视化聚类结果，如图6-1所示，其中每个数据点表示一个由输入图像生成的学习图像。如图6-2所示，在正常情况下，学习图像的聚类结果良好，这与传统观点一致，即滤波器作为聚类函数发挥作用。在少数情况下，聚类结果中可能会出现离群的异常点。为了方便起见，聚类结果中的异常点对应的样本被称为异常样本（*Outlier Samples*），将这种具有异常点的聚类结果对应的滤波器称为异常滤波器（*Anomaly Filters*）。结果发现，异常滤波器的存在与 CNN 的过拟合现象存在关系。针对这一问题，本章设计了一系列实验从三个角度进行详细分析：

- 异常滤波器数量与过拟合模型关系。

观察模型在训练过程中不同时期的异常滤波器数量，发现当模型发生过拟合时，异常滤波器的数量显著增加。

- 与异常滤波器相关的异常样本与模型过拟合的关系。

  在训练过程中，异常样本的梯度通常比正常样本大几倍。这意味着模型在异常样本上过度学习细节，从而导致模型更可能出现过拟合现象。

- 异常滤波器与模型性能的关系。

  通过分析在过拟合模型中移除异常滤波器前后的训练准确性和验证准确性发现，移除异常滤波器通常会导致模型准确性上升。这意味着在移除异常滤波器后缓解了模型的过拟合现象。

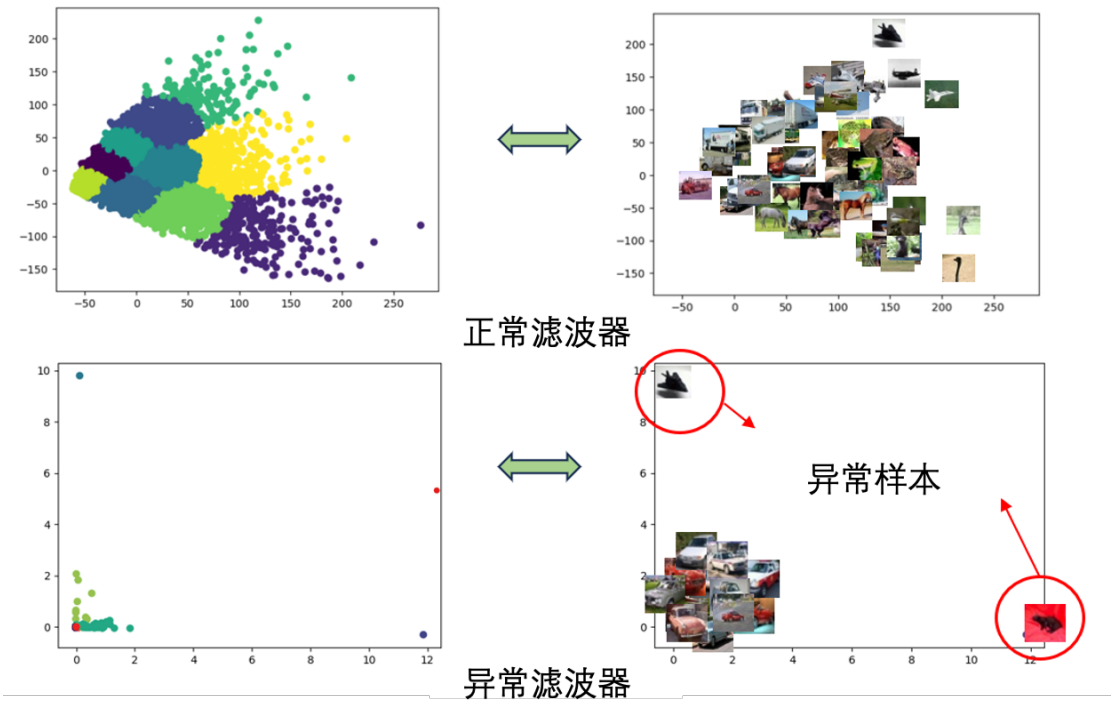


图 6-2 聚类结果可视化

本章主要分析了模型过拟合与 CNN 中异常滤波器之间的关系。主要贡献如下：

- 本章提供了一种细粒度的解释方法，通过无监督聚类方法，探索 CNN 的本质特征。本章的方法不依赖于先验的人工知识，并且无需修改模型，确保解释结果与模型的内在特性密切一致。
- 本章提出了定量指标来评估异常滤波器。为了实现在滤波器层级上评估 CNN，本章提出一种客观指标来检测与模型过拟合相关的异常滤波器。在

对每个滤波器对应的学习图像进行聚类后，本章使用指标来评估聚类结果的质量，并将其进一步用于筛选异常滤波器。

- 本章揭示了 CNN 的行为模式。本章发现一种指示潜在过拟合与异常滤波器关系的独特模式，并从三个方面进行分析。本章使用多个模型和数据集来验证了异常滤波器与过拟合之间的联系。

6.2　基于 NNS 的异常滤波器识别

本节首先使用高斯混合模型（Gaussian Mixture Model, GMM）对每个滤波器的学习图像进行聚类。聚类的数量 K 是动态确定的。算法流程如图6-3所示。随后，本节定义了识别异常滤波器的标准，异常滤波器是指滤波器的聚类结果中存在离群点，并讨论异常滤波器与模型过拟合的关系。

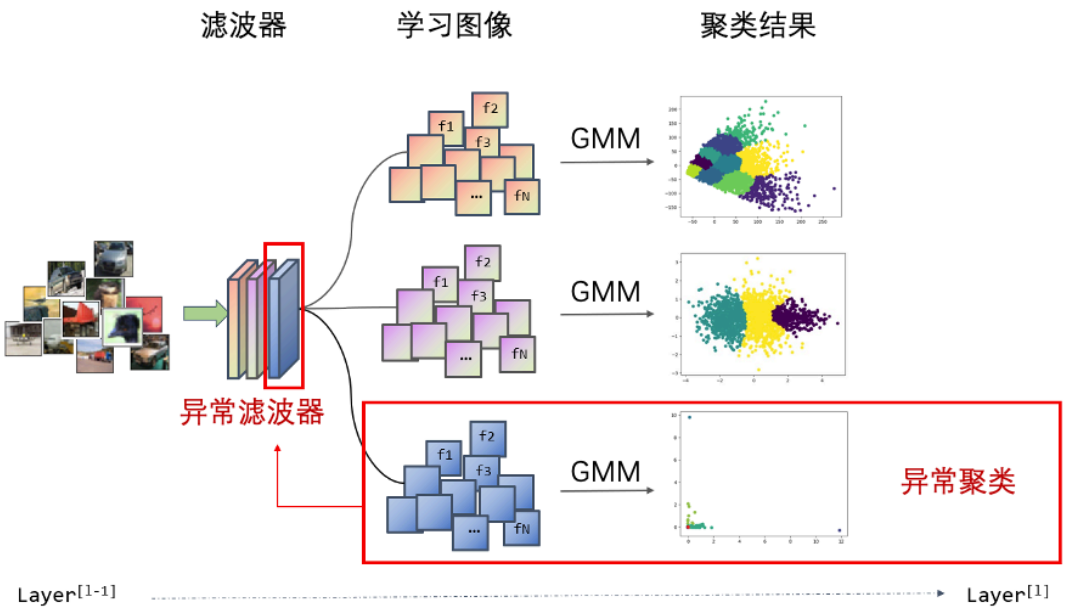


图 6-3　异常滤波器检测流程

为简便起见，本节考虑具有输入 $x \in \mathbb{R}^{N\times N}$ 的神经网络模型，并指定具有 K 个滤波器的前馈层。输入样本的数量为 S。特征图的大小为 $M \times M$。滤波器 k 的特征图为 $f^k(x) \in \mathbb{R}^{M\times M}$。

神经网络扫描仪（NNS）为每个神经元生成一个输入尺寸的样本。生成的样本被称为该神经元的“学习图像”。该学习图像以可视化的方式表示神经元从输入样本中提取的特征。神经网络扫描仪适用于模型的不同组件。对于样本 x 和

指定的滤波器 k 的神经元 n，对应的学习图像 $I_n^k(x)$ 是基于输入神经元对该神经元贡献的计算结果生成的：

$$I_n^k(x) = \sum_{i,j\in[1,N]} c_{(i,j)}(x) \times I_{(i,j)}^0(x), \tag{6-1}$$

其中 $c_{(i,j)}(x)$ 是位置 (i,j) 处的输入神经元对神经元 n 的贡献值，$I_{(i,j)}^0(x)$ 是输入分辨率的图像，除 (i,j) 位置的像素外，其余像素值为零。滤波器 k 的学习图像 $L^k(x)$ 是对应特征图中所有神经元学习图像的总和：

$$L^k(x) = \sum_{i,j\in[1,M]} I_{(i,j)}^k(x) \times f_{(i,j)}^k(x), \tag{6-2}$$

其中 $I_{(i,j)}^k(x)$ 是特征图 k 中位置 (i,j) 处神经元的学习图像，$f_{(i,j)}^k(x)$ 表示该位置的激活值。

6.2.1 过拟合的定义

过拟合是指模型的泛化能力随着训练的进行而退化，特别是在训练周期延长后，测试误差出现上升的现象。过拟合模型通常会记住数据集中的噪声，使其对训练数据以外的偏差非常敏感。本节从经验和数学两个角度定义了过拟合。

在经验上，过拟合通常表现为训练误差最小化后，训练曲线与验证曲线之间的差异变大，通常表现为验证误差上升。这一现象通常通过偏差-方差权衡来解释。然而，Nakkiran et al.[149]挑战了这一观点，揭示了双重下降现象，即随着训练时间的增加，模型的性能先提升，然后下降，最后再次提升的现象。此行为与模型大小、数据规模、训练轮次，甚至优化器的选择等因素有关。

在实验中刻意避免了可能出现双重下降的区域。相反，专注于比较泛化良好的模型和过拟合模型。具体而言，将训练得较好的模型定义为获得峰值测试精度的模型，并将测试性能下降的后期模型视为过拟合模型。如果观察到双重下降现象将在进入过参数化阶段之前截断训练。

此外，本节还从数学角度提供了过拟合的定量度量。过拟合的模型通常会收敛到损失曲面上的尖锐且不规则的极小值（sharp minima），这些极小值与较差的泛化能力相关联。这一几何特性可以通过 Hessian 矩阵捕捉，Hessian 矩阵是

损失函数相对于模型参数的二阶导数。形式上，给定损失函数 $\mathcal{L}(\theta)$，Hessian 矩阵定义为：

$$H = \nabla_{\theta}^{2}\mathcal{L}(\theta) \in \mathbb{R}^{N\times N} \tag{6-3}$$

Hessian 矩阵的谱特性，例如主特征值、迹和特征值分布，反映了损失曲面曲率。用 $\lambda_1, \cdots, \lambda_d$ 代表 Hessian 矩阵特征值，则过拟合可以被定义为：

$$\lambda_{\max}(H(\theta^{*})) \gg \delta, \tag{6-4}$$

其中 δ 是阈值。较大的特征值通常表明较尖锐的极小值，表示模型可能对扰动更加敏感，从而导致更差的泛化能力。相反，较平坦的极小值（对应较小的 Hessian 特征值）通常与较好的泛化能力相关。因此，基于 Hessian 的度量提供了一种有原则的方式，用以定量分析过拟合。

6.2.2　滤波器层级的 GMM 聚类

聚类方法通常非常适合分析 CNN 滤波器的激活，因为它们能够以无监督的方式揭示潜在的模式，展示多样的激活行为，并为模型性能提供见解。本节对每个滤波器进行层级聚类。

首先，本节从特定层 l 提取学习图像 L，其中 $L \in \mathbb{R}^{B\times C\times N\times N}$ 表示一个小批量大小为 B 时，C 个滤波器的学习图像，$L = [L^1, L^2, \cdots, L^C]$。通过并将其按所有位置拼接，重塑学习图像，得到 $S \in \mathbb{R}^{B\times C\times (N\cdot N)}$。

接下来，应用主成分分析（Principal Component Analysis, PCA）将每个学习图像的空间维度减少到 2，从而得到压缩表示 $D \in \mathbb{R}^{B\times C\times 2}$，该表示有助于可视化和聚类。该压缩表示 D 作为后续聚类步骤的输入。

本节采用基于概率的聚类算法 GMM。选择 GMM 进行聚类因为它能够建模复杂的重叠聚类，不像 K-means 那样假设聚类为球形并对离群点敏感。GMM 还提供了一个概率框架，能够提供关于聚类成员的更多信息，这对于分析 CNN 滤波器的学习图像非常有用。与 DBSCAN 和层次聚类（Hierarchical Clustering）相比，后者对参数的选择敏感，并且在面对不同密度的聚类时可能表现较差，GMM 提供了一种更灵活、更合理的方式来识别不同形状和大小的聚类。

对于 D，本节在滤波器层级上进行 GMM 聚类，得到每个滤波器的聚类图。需要注意的是，GMM 要求预先指定类的数量 K，即高斯分布的总数。K 的选择将在第6.2.4节中进一步讨论。

6.2.3 滤波器层级聚类评估

在 CNN 中，滤波器作为特征提取器，具有相似特征的输入往往会产生相似的激活。这一过程与聚类非常相似，其中滤波器将具有特定特征的输入分组为一个聚类。由于滤波器本质上执行了聚类操作，因此通过评估学习图像上的聚类结果为探索模型提供了一个独特的视角。显然，产生高质量聚类结果的滤波器被认为对模型更为重要，因为它们的学习图像可能捕捉了共同且有用的信息。另一方面，产生差聚类结果的滤波器则被认为对模型的影响较小，甚至可能是有害的。因此，建立合理的评估聚类质量的度量标准至关重要。本节使用 CH 指数作为评估指标。

$$CH = \frac{SSB/(K-1)}{SSW/(N-K)}, \tag{6-5}$$

其中

$$SSB = \sum_{i=1}^{K} Z_i \cdot ||\boldsymbol{m_i} - \boldsymbol{m}||_2. \tag{6-6}$$

以及

$$SSW = \sum_{i=1}^{K} \sum_{j=1}^{Z_i} ||\boldsymbol{x_{ij}} - \boldsymbol{m_i}||_2, \tag{6-7}$$

在公式(6-5)中，SSB 和 SSW 分别表示类间和类内的散度矩阵，Z_i 表示第 i 个聚类中的样本数量。SSB 通过计算每个聚类中心 $\boldsymbol{m_i}$ 与数据中心 $\boldsymbol{m}$ 之间的加权欧几里得距离来得到，而 SSW 则计算每个数据点 $\boldsymbol{x_{ij}}$ 与相应聚类中心 $\boldsymbol{m_i}$ 之间的欧几里得距离之和。CH 指数通过直观的方式衡量聚类质量，其中高质量的聚类应在同一类内具有较高的相似性，并且不同类之间具有良好的可区分性。

CH 指数作为一种简单但高效的无监督聚类评估度量，在评估聚类质量时具有很好的适用性。本节对 D 所形成的聚类结果进行滤波器层级评估，并对同一层内的滤波器进行比较。

6.2.4　聚类数目 K 的动态分配

在上述的聚类过程中强调了预先设定的聚类数目 K 的重要性。尽管已经证明了每个滤波器对应的聚类反映了滤波器所学习到的某些模式，但这些模式的确切数量本质上是未知的。

为了解决这个问题，本节提出了一种动态策略，为每个滤波器单独分配聚类数目。本节不再预先固定 K，而是定义一个合理的候选范围，并选择在该范围内最大化 CH 指数的 K 值。

这一选择基于以下原则：更高的 CH 指数表明更好的聚类结构，从而更准确地捕捉到滤波器所学习的潜在激活模式。因此，所选的 K 可以被视为滤波器所编码的不同模式的近似数量。

6.2.5　聚类结果评估

CH 指数提供了一个量化评估单个滤波器的度量标准。通过进一步的可视化，观察到某些聚类结果中出现了一个显著且独特的模式。如图 6-2所示，典型的聚类特征是数据点分布相对均衡，而一个对比模式偶尔会出现——大多数数据点紧密聚集在原点（或某个特定点）周围，而少数离群点则形成遥远、孤立的聚类，将相关滤波器标记为异常滤波器。

识别出异常滤波器的聚类结果的三个关键特征包括：

1. 类别分布不平衡——大多数数据点落入单一聚类，只有少数离群点远离聚集，形成孤立的聚类。
2. 异常高的 CH 指数——根据公式 (6-5)，异常聚类会导致 SSW 非常低，而 SSB 很高，从而导致 CH 指数膨胀。
3. 高激活值——这一点有助于区分具有重要影响的滤波器和冗余滤波器，确保关注那些显著影响模型行为的滤波器。

上述分析展示了识别异常滤波器的步骤，具体步骤见算法 6.1。

算法 6.1 异常滤波器检测

输入：学习图像 D, 最大聚类数 K
参数设置：最大 CH 指数
for $c = 1$ to C **do**
  初始化最大 CH 指数 $\leftarrow 0$
  初始化 $k \leftarrow 0$
  for $k = 2$ to K **do**
    用 k 初始化 GMM 参数
    用 GMM 对 F_c^l 聚类并计算 CH 指数
    if CH 指数 $\geqslant$ 最大 CH 指数 **then**
      最大 CH 指数 $\leftarrow$ CH 指数
      最终 $k \leftarrow k$
    end if
  end for
  用最终 k 初始化 GMM 参数
  用 GMM 对 D 聚类并获得聚类结果
  if 聚类结果满足三个关键特征 **then**
    设定对应滤波器为异常滤波器
  end if
end for
输出：聚类结果和异常滤波器

6.3 实验结果与分析

6.3.1 实验设置

在前面的章节中介绍了一种基于聚类的 CNN 滤波器级别解释和评估的新方法，并强调了异常滤波器的重要性。通过对异常滤波器特征的细致分析，认为它是模型过拟合的潜在指示器，为 CNN 的解释提供了新的视角。本节设计了三个实验来验证异常滤波器与模型过拟合之间的关系。实验在五个 CNN 模型和四个数据集上进行。

表 6-1 实验超参数设计

	λ	α	β	θ
实验一（1）	50	0.2	1	0.1
实验一（2）	10	0.2	1	0.1
实验二	5	-	-	-
实验三（1）	20	0.2	1	0.5
实验三（2）	10	0.2	1	0.1

注释：（1）LeNet-5 和 simple CNN，（2）AlexNet, ResNet-18 和 VGG-16。

模型：为了展示方法的广泛适用性，选择了五种具有不同架构的模型。实验中包括了经典的 CNN 模型，如 AlexNet 和 LeNet-5，以及一个三层的简单 CNN，

其滤波器配置为 32-64-64。此外，还对更复杂的 CNN 模型进行了实验，如 ResNet-18 和 VGG-16。这些模型在深度、卷积核大小、滤波器数量和机制上有所不同，从而验证了方法的通用性。

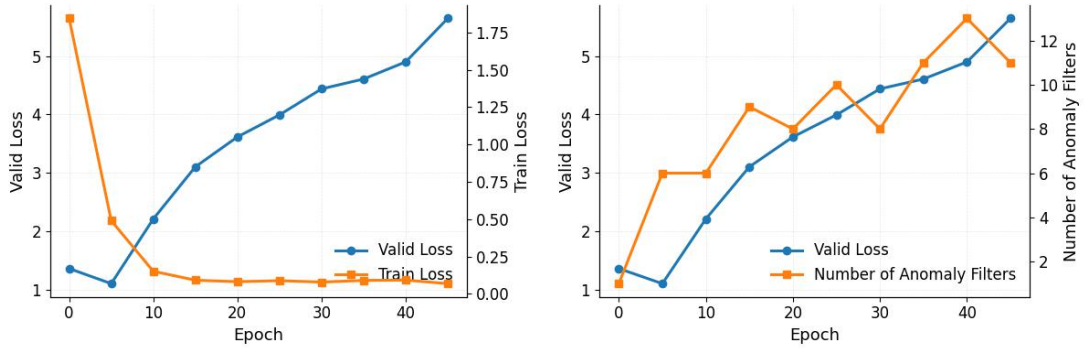
数据集：在数据集方面，使用了广泛应用的 CIFAR-10 和 CIFAR-100，这些数据集在图像分类任务中具有代表性。此外还采用了更为复杂的数据集，如 ILSVRC-2013 DET 数据集。为了方便实验，从原始的 ILSVRC-2013 DET 数据集中随机选择了 10 个类别。需要注意的是，当将小型数据集应用于 AlexNet、ResNet-18 和 VGG-16 时，需要进行额外的插值操作。在本实验中采用了双三次插值方法，这是一种高质量的插值方法，通过 16 个邻近像素的加权平均来估算像素值，从而确保与特定的输入要求一致，最终得到 224x224x3 的图像尺寸。

超参数：如第 6.2.5节所述，本节为识别异常滤波器制定了三个标准：1. 不平衡的类分布，2. 异常高的 CH 指数，3. 足够大的激活值。本节为这些标准设置了一些超参数。对于标准 1，将不超过 λ 个点的聚类视为异常聚类。如果异常聚类的数量超过总聚类数量的 α 倍，则认为相应的滤波器表现出不平衡的类分布。对于异常高的 CH 指数，计算给定层的平均 CH 指数，将 CH 指数超过平均值 β 倍的滤波器视为满足该标准。对于标准 3，计算给定层的平均激活值，并设置阈值 θ，将那些激活值小于平均值 θ 倍的滤波器排除。表6-1提供了三个实验中使用的超参数概览。通过网格搜索来设定这些超参数，选择可以通过手动检查聚类质量及其相应激活模式来验证的超参数。通过调整超参数，可以很好地区分异常滤波器、正常滤波器以及那些提供少量信息的冗余滤波器。

需要注意的是，对于不同规模的模型使用了不同的超参数。较大规模的模型如 VGG-16 相比于异常滤波器，更容易出现未激活的滤波器。因此放宽了阈值 θ。此外，小规模模型在实验中部署了更严格的超参数，因为小规模模型很少有未激活的滤波器，掩码过多的滤波器会不可避免地导致性能的大幅下降。

6.3.2　异常滤波器分布分析

为了深入探究异常滤波器与模型过拟合之间的潜在关联性，本节从异常滤波器数量的变化角度入手，分析其在神经网络训练不同阶段中的变化趋势。具体而言，标准训练框架下，对模型在多个训练轮次中的表现进行了跟踪与记录，从而为理解网络内部特征提取器的异常行为与模型泛化能力之间的关系提供实



(a) (b)

图 6-4 CNN 模型在 CIFAR-10 数据集:（a）模型损失；（b）异常滤波器数量

证支持。在深度神经网络训练过程中，过拟合现象往往表现为模型在训练集上取得极高准确率的同时，在验证集上的性能却未能同步提升，甚至出现下降。这一现象反映出模型对训练数据的过度记忆，丧失了对新样本的泛化能力。为了评估模型是否过拟合，本节采用训练准确率与验证准确率的对比方法：当两者间的差距随着训练轮次增加而扩大，并且验证准确率在达到峰值后开始下降，即可认为模型进入过拟合阶段。为了系统地分析这一过程，本节引入“良好训练模型（Well-Trained Model, WT）”和“过拟合模型（Overfitted Model, OF）”两个概念。WT 模型指的是训练过程中在验证集上取得最佳性能，同时训练损失亦处于较低水平的模型，体现出良好的泛化能力。相对应的，OF 模型则是训练轮次结束时的最终模型，其虽然在训练集上表现优异，但在验证集上的准确率通常已经出现下滑。为了确保对比的公平性，统一选取训练准确率最高的轮次模型作为 WT 模型，并选择最后一个训练轮次的模型作为 OF 模型，二者具备一致的训练环境和模型架构，仅在训练程度上存在差异。为确保该分类的合理性，验证训练过程中未出现双重下降现象，从而排除了在训练后期仍能选出良好泛化模型的可能性。

图6-4a是一个简单 CNN 模型在 CIFAR-10 数据集上的损失变化曲线。相应地，图6-4b则展示了异常滤波器数量随训练轮次的变化趋势。结果表明，在模型训练后期，尤其是在发生过拟合的阶段，异常滤波器数量显著增加，并与验证损失的上升趋势高度一致。这表明，异常滤波器的数量与过拟合密切相关。

为了进一步进行分析，分别检测 WT 模型和 OF 模型应用异常滤波器，以评估在不同训练状态下模型内部滤波器表现出的异常行为。各模型中异常滤波器的数量结果如表 6-2 所示。考虑到数据集复杂度与模型容量之间的匹配程度会

显著影响过拟合的产生，在实验设计中刻意避免在部分不匹配配置下进行测试。例如，对于规模较小的数据集，若采用过于复杂的模型结构，极易产生过拟合，从而影响异常滤波器统计的普适性与代表性。因此，在实验中进行了合理的筛选与控制变量设计，确保数据集复杂度、模型容量以及训练轮次之间保持相对平衡，从而提升实验结果的可信度。实验结果显示，在大多数对比配置中，过拟合模型（OF）中的异常滤波器数量显著高于良好训练模型（WT）。这表明异常滤波器的形成与模型的过拟合状态密切相关，随着模型在训练数据上的过度拟合，部分滤波器可能不再对新的输入模式产生有效响应，成为异常滤波器。这一现象从侧面支持了异常滤波器检测作为评估模型过拟合程度的可行性与有效性。进一步分析显示，同一模型在不同数据集上的异常滤波器数量也存在明显差异。以 ILSVRC-2013 DET 数据集为例，在该数据集上训练得到的过拟合模型中异常滤波器的数量普遍低于在 CIFAR-10 等相对简单数据集上得到的模型。这一发现说明数据集的多样性与复杂度可以有效抑制模型的过拟合倾向，从而减少异常滤波器的产生。复杂的数据分布促使模型需提取更具判别性和泛化能力的特征，提升了各个滤波器的参与度。综上所述，本节通过比较良好训练模型与过拟合模型中的异常滤波器数量，揭示了两者之间存在显著的关联性。

表 6-2　不同模型中的异常滤波器数量

数据集 / 模型	CIFAR-10		CIFAR-100		ILSVRC-2013 DET	
	WT	OF	WT	OF	WT	OF
LeNet-5	2	**12**	8	**13**	-	-
Simple CNN	5	**12**	3	**20**	-	-
AlexNet	10	**53**	30	**37**	6	4
ResNet-18	123	**374**	104	**419**	207	**301**
VGG-16	11	10	16	**29**	2	3

注释：WT 表示训练良好的模型，OF 表示训练过拟合的模型。

6.3.3　异常样本贡献评估

本节旨在进一步探讨异常样本与模型过拟合之间的关系，进而揭示训练数据内部结构对模型泛化能力的潜在影响机制。已有研究表明，过拟合往往与损失函数在参数空间中的尖锐极小点有关，如 Keskar et al.[150]指出，训练过程中的尖锐最小化倾向导致模型在面对未见数据时表现不稳健，从而显著降低了其泛化能力。这类尖锐极小点通常对应于模型对训练数据的过度拟合，其中个别样本的

扰动能够显著改变模型的输出。本节通过分析训练过程中各样本所引发的梯度大小，尝试量化不同样本对过拟合行为的诱发程度。

在神经网络的反向传播过程中，梯度反映了模型对误差的敏感程度，即模型需要多大幅度地调整参数以正确分类当前样本。因此，若某些样本在训练过程中持续引发极高的梯度，则可能表明这些样本难以被模型轻易拟合，模型为此进行了大量的参数调整，从而导致学习过程的不稳定，甚至偏离真实的分类边界。这类样本往往是引发过拟合的重要因素，模型倾向于习得这些样本的特征，而非学习具有普遍性的判别规律。为了进一步分析这一现象，首先需要将输入样本划分为异常样本与正常样本。在异常滤波器激活空间中，不同输入样本之间形成若干簇状分布，少数样本偏离主流聚类，形成孤立的聚类结构。若某一簇中包含的数据点不超过设定目标数，则该簇内样本判定为异常样本。其余样本则被归为正常样本。在样本分类完成后，分别在良好训练模型（WT）和过拟合模型（OF）上，计算每个样本对应的梯度幅值，以评估样本对模型学习动态的影响。采用标准反向传播机制，对每个输入样本计算其在各层引发的梯度，然后取其绝对值并在所有网络层中求平均。所有计算过程在相同超参数设置下完成，确保可比性。最终得到的平均梯度幅值结果见表 6-3。

实验结果表明，在多数数据集与模型组合中，OF 模型在异常样本处的平均梯度幅值显著高于 WT 模型。例如，在 CIFAR-10 数据集上使用 LeNet-5 网络进行训练时，异常样本的平均梯度幅值约为正常样本的 8.8 倍，且在 OF 模型中该差距进一步拉大。这一现象说明，异常样本在训练过程中引发了更加剧烈的参数更新，从而加剧了模型对这些样本的记忆倾向，形成了不平滑的、复杂的决策边界。这种行为正是过拟合的典型特征之一。进一步观察还发现，即使在良好训练模型中，异常样本的梯度值也普遍高于正常样本。这表明，异常样本本身在特征空间中就具有较大的不确定性，可能包含噪声、非典型分布或难以判别的特征组合。模型在处理这些样本时，即使尚未过拟合，也已经表现出学习难度加大的迹象。而在进入过拟合状态后，模型会进一步调整参数以极力拟合这些高梯度样本，最终导致模型在验证集上的性能下降。换句话说，异常样本在一定程度上驱动了模型向尖锐最小化的方向演化。

综上所述，本节从梯度的角度揭示了异常样本与模型过拟合之间的强相关性。异常样本引发的高梯度更新表明，模型在这些样本上形成了高度复杂的决策

超平面，其带来的训练不稳定性最终导致了泛化能力的下降。

表 6-3　不同样本的平均梯度值

数据集 / 模型	CIFAR-10		CIFAR-100		ILSVRC-2013 DET	
	Outlier	Normal	Outlier	Normal	Outlier	Normal
LeNet-5 (WT)	**0.461**	0.052	**0.163**	0.032	-	-
LeNet-5 (OF)	**2.811**	0.201	**0.264**	0.162	-	-
Simple CNN (WT)	**0.079**	0.024	**0.083**	0.015	-	-
Simple CNN (OF)	**0.281**	0.066	**0.312**	0.178	-	-
AlexNet (WT)	**0.005**	0.001	**0.017**	0.002	**0.012**	0.016
AlexNet (OF)	**0.014**	0.005	**0.063**	0.005	**0.045**	0.014
ResNet-18 (WT)	0.003	0.003	**0.003**	0.002	0.006	**0.007**
ResNet-18 (OF)	0.006	**0.008**	**0.009**	0.008	0.010	0.010
VGG-16 (WT)	**0.006**	0.004	0.004	0.004	0.018	**0.020**
VGG-16 (OF)	0.003	**0.004**	**0.017**	0.010	**0.070**	0.032

6.3.4　异常滤波器重要性评估

在前述实验中已初步揭示了过拟合现象与异常滤波器之间存在显著的相关性。为了进一步验证这种关联并探究异常滤波器对模型性能的具体影响，本节设计并实施了一组掩码实验。该实验通过掩码异常的滤波器，观察其对训练与验证性能的影响，从而为异常滤波器在模型过拟合过程中的作用提供更具实证意义的支持。具体而言，采用一种直接而有效的方式对异常滤波器进行掩码，即将异常滤波器的输出值置为零，从而实质上阻断其在前向传播过程中的贡献。为评估掩码策略的实际效果，在模型训练结束后分别在训练集与验证集上测试了模型的分类准确率，并分析其变化趋势。如果掩码的滤波器导致过拟合，那么其对训练准确率的影响应为负面，即训练准确率下降；而验证准确率应有所提升或保持稳定，从而反映出模型对未见数据的泛化能力得以提升。首先进行单滤波器掩码实验，即在每轮实验中仅掩码一个由异常检测算法识别出的异常滤波器。实验结果汇总于表 6-4 中。可以观察到，在大多数情形下，掩码后的模型在训练集上的准确率略有下降，而验证集准确率则出现上升。这种趋势表明，这些被识别并掩码的滤波器可能确实引导了模型对训练集的过度学习，掩码操作有效缓解了过拟合的影响，提升了模型的泛化表现。然而，仍然存在一些例外情况。有部分滤波器的掩码并未提升验证准确率，甚至在个别情况下略有下降。这说明，在当前的异常滤波器判定标准下，仍存在一定程度的误判，即部分对最终任务具有实际贡献的滤波器可能被误归类为异常。

表 6-4 掩码后发生准确率变化的滤波器数量

数据集 / 模型	CIFAR-10		CIFAR-100		ILSVRC-2013 DET	
	训练集 ↓	验证集 ↑	训练集 ↓	验证集 ↑	训练集 ↓	验证集 ↑
LeNet-5	3/3	1/3	2/2	0/2	-	-
Simple CNN	3/3	1/3	8/8	6/8	-	-
AlexNet	6/6	5/6	13/15	13/15	4/4	2/4
ResNet-18	218/234	102/234	316/356	247/356	3/3	2/3
VGG-16	7/8	5/8	14/16	12/16	209/237	114/237

为了验证该方法相较于随机滤波器掩码的有效性，进一步设计了随机单滤波器掩码对比实验。在该实验中，每轮随机选取一个滤波器进行掩码操作，并重复若干轮以取平均结果，表 6-5实验结果显示，相比随机掩码方式，基于异常识别的掩码策略在绝大多数测试场景中对验证准确率的影响更小，甚至在部分情形中准确率出现提升。这进一步验证了异常滤波器与过拟合强相关。即，相较于随机滤波器，异常滤波器更可能是造成模型过拟合的关键因素。

表 6-5 掩码不同滤波器后验证准确率变化

数据集 / 模型	CIFAR-10		CIFAR-100		ILSVRC-2013 DET	
	Anomaly	Random	Anomaly	Random	Anomaly	Random
LeNet-5	**-1.21%**	-3.94%	**-1.27%**	-6.92%	-	-
Simple CNN	**+0.10%**	-1.51%	**-0.18%**	-0.74%	-	-
AlexNet	**+0.06%**	-0.15%	**+0.01%**	-0.43%	**+0.02%**	-0.10%
ResNet-18	**+0.15%**	-0.04%	**+0.02%**	-0.30%	**+0.01%**	-0.09%
VGG-16	**+0.01%**	-0.48%	**+0.00%**	-0.61%	**+0.02%**	-0.04%

本节通过掩码实验进一步验证了异常滤波器与模型过拟合之间的关系。掩码异常滤波器能够减少过拟合，提升验证准确率，改善模型的能力，帮助 CNN 模型避免过拟合并提高其在任务中的表现。

6.4 本章小结

本章提出了一种新颖的方法，通过使用高斯混合模型聚类来研究异常滤波器与模型过拟合之间的关系。通过对单个滤波器生成的学习图像进行聚类，识别出与模型过拟合密切相关的异常滤波器。本章设计了一系列实验，涵盖了多种 CNN 架构和数据集。通过引入包括经典网络（如 LeNet-5 和 AlexNet）以及更复杂的架构（如 ResNet-18 和 VGG-16）在内的多种 CNN 模型，本章的方法展示了其广泛的适用性。实验结果支持异常滤波器与模型过拟合之间的强关联。本章从三个方面进行分析并发现：1. 异常滤波器在过拟合模型中更为常见。2. 与异

常滤波器相关的异常样本促使了模型的过拟合。3. 移除异常滤波器能缓解模型的过拟合。

未来计划优化异常滤波器检测的标准，并将方法扩展到更多的神经网络架构和应用中。此外，未来的工作将探索如何进一步通过优化异常滤波器实现模型的剪枝。

第七章　总结与展望

7.1　本文总结

在本研究中，本文探讨了基于神经元的卷积神经网络可解释性方法。传统的 CNN 通常被视为“黑箱”模型，缺乏对其决策过程的透明度和可解释性。随着深度学习模型在各个领域的广泛应用，如何理解和解释 CNN 的内部机制已成为一个重要的研究方向。当前的神经网络解释方法通常与固定的神经网络模型架构紧密相关，不同的网络结构在结构设计和功能模块方面存在显著差异，因此无法有效地对比不同模块的工作机制。这种局限性使得对复杂模型的全面理解变得困难，也影响了跨模型的系统性分析。针对这一问题，本文主要关注采用基于神经元的可解释性方法分析模型，这类方法具备了跨模块的可解释性能力，提供一致的解释框架，从而为神经网络提供统一的理解和评估工具。根据解释目标的颗粒度不同，本文的解释方法分为神经元级定性解释、神经元级定量解释、滤波器级定量解释和模式级定量解释。

在第三章中，本文提出了一种神经元级定性解释方法，称为神经网络扫描仪，通过可视化神经元学习过程来解释神经网络的全新方法。该方法能够针对指定神经网络提取每个神经元所学习到的特征，并将其以人类可理解的形式展示出来。通过整合不同神经元所学习到的特征，可以对各种神经网络模型的工作机制进行详细分析。该方法适用于多种神经网络结构，且无需对模型进行任何结构上的修改。本文将神经网络扫描仪应用于不同的模块，并在图像分类任务的模型中进行了实验，验证了该方法的有效性。通过这些实验，本文深入分析了神经网络的工作原理，并从多个角度评估了 CNN 模型的可解释性。

第四章主要探索了神经元级别的定量解释方法。针对神经网络扫描仪的结果具有主观性但缺乏标准评估准则的问题。本文引入了“特征量”这一概念，用来量化并评估每个神经元所编码的特征信息。通过这一方法，本文提出了关于卷积层的三项假设：神经元的特征量与其激活值呈正相关；随着网络深度的增加，

特征量的分布由多样性逐渐过渡到均一性；不同通道在特征学习能力上存在显著差异。本文通过实验在多种模型架构上验证了这些假设，并对卷积层进行了深入分析。

第五章聚焦于滤波器级的定量解释方法。针对神经网络扫描仪缺乏对滤波器的分析及从单一维度进行解释的问题，本文将神经网络扫描仪的神经元级解释方法，进一步扩展其应用到滤波器的分析上，以揭示滤波器的行为与特性。本文引入了“可解释属性”的概念，从多个维度对滤波器学习到的特征进行有效量化，进而提取出多种可解释属性。本文提出了一个统一且具有良好扩展性的可解释性框架——测量可解释特性框架，旨在通过不同的可解释属性对模型进行多角度解释。通过该框架，本文设计了三种具有代表性的评估指标，以提供从不同视角出发的模型解释。通过实验，本文分析了如何利用这一框架进行模型的多角度解释。

第六章探索了模式级的定量解释方法。针对神经网络扫描仪缺乏对模型特殊模式和现象的解释的问题，本文通过神经网络扫描仪对滤波器进行扫描，并对学习图像进行聚类，进而探究聚类结果与模型过拟合现象之间的关系。为此，本文设计了量化指标来评估滤波器的特性，并识别异常滤波器，揭示其在模型过拟合中的作用机制。本文提出了以下三个假设：异常滤波器在过拟合模型中更为常见；与异常滤波器相关的异常样本促进了模型的过拟合；移除异常滤波器能够缓解模型的过拟合现象。本文设计并开展了多个实验研究，从不同角度验证了这些假设。

7.2 未来展望

尽管神经网络的可解释性方法取得了显著进展，但仍存在挑战和改进空间。未来的研究可以从以下几个方面进一步拓展和优化：

- **进一步拓展分析模型的表现能力和学习能力。**尽管现有的可解释性方法提供了深入理解神经网络内部机制的窗口，但如何通过可解释性分析来评估和提升模型的表现能力和学习能力仍然是一个挑战。基于现有研究成果，未来的工作可以进一步探索如何将现有可解释性分析方法泛化到各种神经网络结构中，从而揭示不同架构下神经元与层级之间的协同模式。这将有

助于深入理解不同模型在处理多样化任务时的特征提取路径和学习行为，为构建更加通用且高效的深度学习系统提供理论支持。

- **基于可解释性的模型优化。**当前的模型优化策略通常专注于模型的性能提升，常常通过自动化的算法调节模型参数，但这些优化方法通常缺乏对模型内部机制的理解，难以为模型的可解释性提供支持。未来的研究应致力于基于可解释性的模型优化策略，这不仅是提升模型性能的途径，也是确保优化过程透明和可控的重要手段。基于可解释性的优化方法可以理解模型在训练过程中的行为，识别影响模型性能的关键特征和神经元，从而有的放矢地优化网络结构、选择合适的正则化方法，并提高模型在各种应用场景中的适应性。
- **基于可解释性的模型安全。**随着 AI 技术在各个领域的广泛应用，模型安全问题变得尤为重要。当前的模型安全算法大多专注于提高模型的鲁棒性和防御能力，但这些方法通常忽视了模型的可解释性，导致防御机制缺乏透明度和可理解性。未来的研究可以基于可解释性进行模型安全研究，通过可解释性分析提升 AI 模型的防御能力。通过揭示模型对不同输入数据的反应模式和决策边界，可以更好地识别潜在的漏洞和薄弱环节，进而设计出更加稳健的防御策略。这不仅能够提升模型的安全性，还能够增强用户对模型的信任。

参考文献

[1] BODRIA F, GIANNOTTI F, GUIDOTTI R, et al. Benchmarking and survey of explanation methods for black box models[J]. Data Mining and Knowledge Discovery, 2023, 37: 1719-1778.

[2] KONG X, TANG X, WANG Z. A survey of explainable artificial intelligence decision[J]. Systems Engineering - Theory & Practice, 2021, 41: 524-536.

[3] GOYAL Y, WU Z, ERNST J, et al. Counterfactual visual explanations[C]//Proceedings of the International Conference on Machine Learning. 2019: 2376-2384.

[4] WANG Y, SU H, ZHANG B, et al. Interpret neural networks by identifying critical data routing paths[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8906-8914.

[5] RUBEL A. The black box society: the secret algorithms that control money and information[J]. Business Ethics Quarterly, 2016, 26: 568-571.

[6] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. Nat. Mach. Intell, 2019, 1: 206-215.

[7] SU J, LIU H, XIANG F, et al. Survey of interpretation methods for deep neural networks[J]. Computer Engineering, 2020, 46: 1-15.

[8] SCHRAMOWSKI P, STAMMER W, TESO S, et al. Making deep neural networks right for the right scientific reasons by interacting with their explanations[J]. Nat. Mach. Intell, 2020, 2: 476-486.

[9] ZECH B, JR, MA L, M C, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study[J]. PLoS Medicine, 2018, 15: 1002683.

[10] BADGELEY M, ZECH O R, JR, L G, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables[J]. Npj Digital Medicine, 2019, 2: 31.

[11] HAMAMOTO R, SUVARNA K, YAMADA M, et al. Application of artificial intelligence technology in oncology: towards the establishment of precision medicine[J]. Cancers, 2020, 12: 3532.

[12] PHILLIPS P, HAHN C, FONTANA P, et al. Four principles of explainable artificial intelligence[J]. Natl. Inst. Stand. Technol. Interag. Intern. Rep, 2021, 8312: 1-43.

[13] MILLER T. Explanation in artificial intelligence: insights from the social sciences[J]. Artif. Intell, 2019, 267: 1-38.

[14] DOSHI-VELEZ F, KIM B. A roadmap for a rigorous science of interpretability[J]. ArXiv preprint arXiv:1702.08608, 2017.

[15] POPE P, KOLOURI S, ROSTAMI M, et al. Explainability methods for graph convolutional neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019: 10772-10781.

[16] HOFMAN J, SHARMA A, WATTS D. Prediction and explanation in social systems[J]. Science, 2017, 355: 486-488.

[17] WELLER A. Transparency: motivations and challenges[G]//Explainable AI: interpreting, explaining and visualizing deep learning. Springer, 2019: 23-40.

[18] LIPTON Z. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery[J]. Queue, 2018, 16: 31-57.

[19] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning[J]. ArXiv preprint arXiv:1702.08608, 2017.

[20] ZHANG Y, TIŇO P, LEONARDIS A, et al. A survey on neural network interpretability[J]. IEEE transactions on emerging topics in computational intelligence, 2021, 5: 726-742.

[21] MONTAVON G, SAMEK W, MÜLLER K R. Methods for interpreting and understanding deep neural networks[J]. Digital Signal Processing, 2018, 73: 1-15.

[22] GILPIN L, BAU D, YUAN B, et al. Explaining explanations: an overview of interpretability of machine learning[J]. IEEE International Conference on Data Science and Advanced Analytics, 2018: 80-89.

[23] FREITAS A. Comprehensible classification models: a position paper[J]. SIGKDD Explor, 2013, 15: 1-10.

[24] JOHANSSON U, KÖNIG R, NIKLASSON L. The truth is in there - rule extraction from opaque models using genetic programming[C]//Proceedings of the International Florida Artificial Intelligence Research Society Conference. 2004: 658-663.

[25] ARRIETA A, DÍAZ-RODRÍGUEZ N, DEL SER J, et al. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai[J]. Inf. Fusion, 2020, 58: 82-115.

[26] ABIODUN O, JANTAN A, OMOLARA A, et al. State-of-the-art in artificial neural network applications: a survey[J]. Heliyon, 2018, 4: 00938.

[27] CHOULDECHOVA A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments[J]. Big Data, 2017, 5: 153-163.

[28] RUAN L, WEN S, NIU Y, et al. Deep neural network visualization based on interpretable basis decomposition and knowledge graph[J]. Chinese Journal of Computers, 2021, 44: 1786-1805.

[29] ZHANG Z, XIE Y, XING F, et al. MDNet: a semantically and visually interpretable medical image diagnosis network[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3549-3557.

[30] ZHU Y, MA J, YUAN C, et al. Interpretable learning based dynamic graph convolutional networks for alzheimer’s disease analysis[J]. Information Fusion, 2022, 77: 53-61.

[31] KIM J, CANNY J. Interpretable learning for self-driving cars by visualizing causal attention[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2942-2950.

[32] YOU J, LESKOVEC J, HE K, et al. Graph structure of neural networks[C]// International Conference on Machine Learning. 2020: 10881-10891.

[33] WANG W, RAO Y, WU L, et al. Progress of judicial judgment prediction based on artificial intelligence[J]. Journal of Chinese Information Processing, 2021, 35: 1-14.

[34] PIANO S L. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward[J]. Humanities and Social Sciences Communications, 2020, 7: 1-7.

[35] THIEBES S, LINS S, SUNYAEV A. Trustworthy artificial intelligence[J]. Electronic Markets, 2021, 31: 447-464.

[36] BRUNDAGE M, AVIN S, WANG J. Toward trustworthy ai development: mechanisms for supporting verifiable claims[J]. ArXiv preprint arXiv:2004.07213, 2020.

[37] RUDIN C, CHEN C, CHEN Z, et al. Interpretable machine learning: Fundamental principles and 10 grand challenges[J]. ArXiv preprint arXiv:2103.11251, 2021.

[38] CHEN K, MENG X. Interpretation and understanding in machine learning[J]. Journal of Computer Research and Development, 2020, 57: 1971-1986.

[39] LAKKARAJU H, KAMAR E, CARUANA R, et al. Identifying unknown unknowns in the open world: representations and policies for guided exploration[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2017: 2124-2132.

[40] KROLL J, HUEY J, BAROCAS S, et al. Accountable algorithms[J]. University of Pennsylvania Law Review, 2017, 165: 633-705.

[41] DANKS D, LONDON A. Regulating autonomous systems: beyond standards[J]. IEEE Intell. Syst, 2017, 32: 88-91.

[42] KINGSTON J. Artificial intelligence and legal liability[C]//Proceedings of the SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. 2016: 269-279.

[43] ZHOU Z, CAO Y, HU C, et al. The interpretability of rule-based modeling approach and its development[J]. Acta Automatica Sinica, 2021, 47: 1201 1216.

[44] MINEMATSU T, SHIMADA A, UCHIYAMA H, et al. Analytics of deep neural network-based background subtraction[J]. J. Imaging, 2018, 4: 78.

[45] RUDIN C, RADIN J. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition[J]. Harvard Data Science Review, 2019, 1.

[46] LAUGEL T, LESOT M J, MARSALA C, et al. The dangers of post-hoc interpretability: unjustified counterfactual explanations[C]//Proceedings of the International Joint Conference on Artificial Intelligence. 2019: 2801-2807.

[47] LAKKARAJU H, BASTANI O. How do i fool you?: manipulating user trust via misleading black box explanations[J]. AAAI/ACM Conference on AI, Ethics, and Society, 2020: 79-85.

[48] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps[C]//International Conference on Learning Representations. 2013: 1-8.

[49] 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.

[50] WANG H, WANG Z, DU M, et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and pattern recognition workshops. 2020: 24-25.

[51] LEE J R, KIM S, PARK I, et al. Relevance-cam: Your model already knows where to look[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14944-14953.

[52] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.

[53] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks[C]//2018 IEEE Winter conference on Applications of Computer Vision. 2018: 839-847.

[54] MUHAMMAD M B, YEASIN M. Eigen-cam: Class activation map using principal components[C]//2020 International Joint Conference on Neural Networks. 2020: 1-7.

[55] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision. 2014: 818-833.

[56] REN D, ZHANG K, WANG Q, et al. Neural blind deconvolution using deep priors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3341-3350.

[57] ZHANG Q, WANG X, CAO R, et al. Extraction of an Explanatory Graph to Interpret a CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2021, 43: 3863-3877.

[58] SMILKOV D, THORAT N, KIM B, et al. Smoothgrad: removing noise by adding noise[J]. ArXiv preprint arXiv:1706.03825, 2017.

[59] GHALEBIKESABI S, TER-MINASSIAN L, DIAZORDAZ K, et al. On locality of local explanation models[J]. Advances in Neural Information Processing Systems, 2021, 34: 18395-18407.

[60] PETSIUK V, DAS A, SAENKO K. Rise: Randomized input sampling for explanation of black-box models[J]. ArXiv preprint arXiv:1806.07421, 2018.

[61] FEL T, CADÈNE R, CHALVIDAL M, et al. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis[J]. Advances in Neural Information Processing Systems, 2021, 34: 26005-26014.

[62] NOVELLO P, FEL T, VIGOUROUX D. Making sense of dependence: Efficient black-box explanations using dependence measure[J]. Advances in Neural Information Processing Systems, 2022, 35: 4344-4357.

[63] BAU D, ZHOU B, KHOSLA A, et al. Network dissection: quantifying interpretability of deep visual representations[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3319-3327.

[64] KIM B, WATTENBERG M, GILMER J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)[C]// International Conference on Machine Learning. 2018: 2668-2677.

[65] GHORBANI A, WEXLER J, ZOU J Y, et al. Towards automatic concept-based explanations[J]. Advances in Neural Information Processing Systems, 2019, 32.

[66] CHENG X, RAO Z, CHEN Y, et al. Explaining knowledge distillation by quantifying the knowledge[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and pattern recognition. 2020: 12925-12935.

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

[68] RAMAMURTHY K, VINZAMURI B, ZHANG Y, et al. Model agnostic multilevel explanations[J]. Advances in Neural Information Processing Systems, 2020: 5968-5979.

[69] GUIDOTTI R, MONREALE A, RUGGIERI S, et al. A survey of methods for explaining black box models[J]. ACM Comput. Surv, 2019, 51: 1-93 42.

[70] ANGELOV P, SOARES E, JIANG R, et al. Explainable artificial intelligence: an analytical review[J]. Wiley Interdiscip. Rev. Data Min. Knowl. Discov, 2021, 11.

[71] LINARDATOS P, PAPASTEFANOPOULOS V, KOTSIANTIS S. Explainable ai: a review of machine learning interpretability methods[J]. Entropy, 2021, 23: 18.

[72] ZHANG Q, WU Y N, ZHU S C. Interpretable convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and pattern recognition. 2018: 8827-8836.

[73] ERHAN D, BENGIO Y, COURVILLE A, et al. Visualizing higher-layer features of a deep network[J]. Univeristé de Montréal, 2009, 1341: 1-13.

[74] OLAH C, MORDVINTSEV A, SCHUBERT L. Feature visualization[J]. Distill, 2017, 11: 1.

[75] ZHANG Q, WANG W, ZHU S C. Examining cnn representations with respect to dataset bias[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018: 4464-4473.

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

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

[78] NGUYEN A, DOSOVITSKIY A, YOSINSKI J, et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks[J]. Advances in Neural Information Processing Systems, 2016, 29: 3387-3395.

[79] FONG R, VEDALDI A. Net2Vec: quantifying and explaining how concepts are encoded by filters in deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8730-8738.

[80] ZHOU B, SUN Y, BAU D, et al. Interpretable basis decomposition for visual explanation[C]//Proceedings of the European Conference on Computer Vision. 2018: 119-134.

[81] PINHEIRO P O, COLLOBERT R. From image-level to pixel-level labeling with convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and pattern recognition. 2015: 1713-1721.

[82] NGUYEN A, YOSINSKI J, CLUNE J. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks[J]. ArXiv preprint arXiv:1602.03616, 2016.

[83] ZHANG Q, WANG X, CAO R, et al. Extraction of an explanatory graph to interpret a cnn[J]. IEEE transactions on Pattern Analysis and Machine Intelligence, 2020, 43: 3863-3877.

[84] ZHANG Q, CAO R, SHI F, et al. Interpreting CNN knowledge via an explanatory graph[C]//Proceedings of the AAAI conference on artificial intelligence: vol. 32. 2018.

[85] HUA Y, ZHANG D, GE S. Research progress in the interpretability of deep learning models[J]. Journal of Cyber Security, 2020, 5: 1-12.

[86] SI N, ZHANG W, QU D, et al. Representation visualization of convolutional neural networks: a research survey[J]. Acta Automatica Sinica, 2021: 1-31.

[87] SHI X, NI L, WANG J, et al. Interpretable cnn based on minimum entropy constraint[J]. Aerospace Control, 2021, 39: 39-43.

[88] ZINTGRAF L, COHEN T, ADEL T, et al. Visualizing deep neural network decisions: prediction difference analysis[C]//International Conference on Learning Representations. 2017: 1-11.

[89] SINGLA S, WALLACE E, FENG S, et al. Understanding impacts of high-order loss approximations and features in deep learning interpretation[C]//Proceedings of the International Conference on Machine Learning. 2019: 5848-5856.

[90] WANG S, ZHOU T, BILMES J. Bias also matters: bias attribution for deep neural network explanation[C]//Proceedings of the International Conference on Machine Learning. 2019: 6659-6667.

[91] ZEILER M D, KRISHNAN D, TAYLOR G W, et al. Deconvolutional networks[C]//IEEE Computer Society Conference on computer vision and pattern recognition. 2010: 2528-2535.

[92] ZEILER M D, TAYLOR G W, FERGUS R. Adaptive deconvolutional networks for mid and high level feature learning[C]//International Conference on Computer Vision. 2011: 2018-2025.

[93] SPRINGENBERG J, DOSOVITSKIY A, BROX T, et al. Striving for simplicity: the all convolutional net[C]//International Conference on Learning Representations. 2015: 1-11.

[94] SELVARAJU R, COGSWELL M, DAS A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization[C]//IEEE International Conference on Computer Vision. 2017: 618-626.

[95] ADLER P, FALK C, FRIEDLER S, et al. Auditing black-box models for indirect influence[J]. Knowl. Inf. Syst, 2018, 54: 95-122.

[96] HESKES T, SIJBEN E, BUCUR I, et al. Causal shapley values: exploiting causal knowledge to explain individual predictions of complex models[J]. Advances in Neural Information Processing Systems, 2020: 4778-4789.

[97] ANCONA M, ÖZTIRELI C, GROSS M. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation[C]//Proceedings of the International Conference on Machine Learning. 2019: 272-281.

[98] VAN L A, J K. Interpretable counterfactual explanations guided by prototypes[C]//European Conference on Machine Learning and Knowledge Discovery in Databases. 2021: 650-665.

[99] KOH P W, LIANG P. Understanding black-box predictions via influence functions[C]//International Conference on Machine Learning. 2017: 1885-1894.

[100] FONG R C, VEDALDI A. Interpretable explanations of black boxes by meaningful perturbation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3429-3437.

[101] AGARWAL C, NGUYEN A. Explaining image classifiers by removing input features using generative models[C]//Asian Conference on Computer Vision. 2020: 101-118.

[102] WAGNER J, KÖHLER J, GINDELE T, et al. Interpretable and fine-grained visual explanations for convolutional neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9097-9107.

[103] ZHANG J, BARGAL S, LIN Z, et al. Top-down neural attention by excitation backprop[J]. Int. J. Comput. Vis, 2018, 126: 1084-1102.

[104] XU K, LIU S, ZHANG G, et al. Interpreting adversarial examples by activation promotion and suppression[J]. ArXiv preprint arXiv:1904.02057, 2019.

[105] SINGH C, MURDOCH W, YU B. Hierarchical interpretations for neural network predictions[C]//International Conference on Learning Representations. 2019: 1-11.

[106] CHEN J, SONG L, WAINWRIGHT M, et al. Learning to explain: an information-theoretic perspective on model interpretation[C]//Proceedings of the International Conference on Machine Learning. 2018: 882-891.

[107] LAPUSCHKIN S, WÄLDCHEN S, BINDER A, et al. Unmasking clever hans predictors and assessing what machines really learn[J]. Nature Communications, 2019, 10: 1-8.

[108] YANG C, RANGARAJAN A, RANKA S. Global model interpretation via recursive partitioning[C]//IEEE International Conference on High Performance Computing and Communications; IEEE International Conference on Smart City; IEEE International Conference on Data Science and Systems. 2018: 1563-1570.

[109] SALMAN S, PAYROVNAZIRI S, LIU X, et al. DeepConsensus: consensus-based interpretable deep neural networks with application to mortality prediction[C]//International Joint Conference on Neural Networks. 2020: 1-8.

[110] PENG Z, HUANG W, GU S, et al. Conformer: Local features coupling global representations for visual recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 367-376.

[111] WANG Z J, TURKO R, SHAIKH O, et al. CNN explainer: Learning convolutional neural networks with interactive visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 27: 1396-1406.

[112] SAMEK W, MONTAVON G, LAPUSCHKIN S, et al. Explaining deep neural networks and beyond: A review of methods and applications[J]. Proceedings of the IEEE, 2021, 109: 247-278.

[113] YANG Z, KAFLE K, DERNONCOURT F, et al. Improving visual grounding by encouraging consistent gradient-based explanations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 19165-19174.

[114] YANG R, WANG B, BILGIC M. Idgi: A framework to eliminate explanation noise from integrated gradients[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 23725-23734.

[115] AGARWAL C, KRISHNA S, SAXENA E, et al. Openxai: Towards a transparent evaluation of model explanations[J]. Advances in Neural Information Processing Systems, 2022, 35: 15784-15799.

[116] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for simplicity: The all convolutional net[J]. ArXiv preprint arXiv:1412.6806, 2014.

[117] SUNDARARAJAN M, TALY A, YAN Q. Axiomatic attribution for deep networks[C]//International Conference on Machine Learning. 2017: 3319-3328.

[118] STURMFELS P, LUNDBERG S, LEE S I. Visualizing the impact of feature attribution baselines[J]. Distill, 2020, 5: e22.

[119] KINDERMANS P J, HOOKER S, ADEBAYO J, et al. The (un) reliability of saliency methods[J]. Explainable AI: Interpreting, explaining and visualizing deep learning, 2019: 267-280.

[120] HSIEH C Y, YEH C K, LIU X, et al. Evaluations and methods for explanation through robustness analysis[J]. ArXiv preprint arXiv:2006.00442, 2020.

[121] HAUG J, ZÜRN S, EL-JIZ P, et al. On baselines for local feature attributions[J]. ArXiv preprint arXiv:2101.00905, 2021.

[122] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13: 600-612.

[123] ESTER M, KRIEGEL H P, SANDER J, et al. Density-based spatial clustering of applications with noise[C]//Int. Conf. knowledge discovery and data mining: vol. 240. 1996.

[124] SERMANET P, EIGEN D, ZHANG X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. ArXiv preprint arXiv:1312.6229, 2013.

[125] WAH C, BRANSON S, WELINDER P, et al. The caltech-ucsd birds-200-2011 dataset[J]., 2011.

[126] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88: 303-338.

[127] LEEM S, SEO H. Attention Guided CAM: Visual Explanations of Vision Transformer Guided by Self-Attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 38. 2024: 2956-2964.

[128] WALKER C, JHA S, CHEN K, et al. Integrated decision gradients: Compute your attributions where the model makes its decision[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 38. 2024: 5289-5297.

[129] LUNDSTROM D D, HUANG T, RAZAVIYAYN M. A rigorous study of integrated gradients method and extensions to internal neuron attributions[C]// International Conference on Machine Learning. 2022: 14485-14508.

[130] LUNDBERG S. A unified approach to interpreting model predictions[J]. ArXiv preprint arXiv:1705.07874, 2017.

[131] SINGH R, SHUKLA A, TURAGA P. Improving Shape Awareness and Interpretability in Deep Networks Using Geometric Moments[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 4159-4168.

[132] 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.

[133] TAN Z, TIAN Y, LI J. GLIME: general, stable and local LIME explanation[J]. Advances in Neural Information Processing Systems, 2024, 36.

[134] REN J, ZHOU Z, CHEN Q, et al. Can we faithfully represent masked states to compute shapley values on a dnn?[J]. ArXiv preprint arXiv:2105.10719, 2021.

[135] XU Z J. Understanding training and generalization in deep learning by fourier analysis[J]. ArXiv preprint arXiv:1808.04295, 2018.

[136] LEE N, AJANTHAN T, TORR P H. Snip: Single-shot network pruning based on connection sensitivity[J]. ArXiv preprint arXiv:1810.02340, 2018.

[137] ATHIWARATKUN B, KANG K. Feature representation in convolutional neural networks[J]. ArXiv preprint arXiv:1507.02313, 2015.

[138] DOSOVITSKIY A, BROX T. Inverting visual representations with convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and pattern recognition. 2016: 4829-4837.

[139] ZHANG X, XU J, YANG J, et al. Understanding the learning mechanism of convolutional neural networks in spectral analysis[J]. Analytica Chimica Acta, 2020, 1119: 41-51.

[140] GIRISH D, SINGH V, RALESCU A L. Unsupervised clustering based understanding of CNN[C]//CVPR Workshops. 2019: 9-11.

[141] FILTERS' IMPORTANCE D. Pruning filters for efficient convnets[J]. ArXiv preprint arXiv:1608.08710, 2016.

[142] MONDAL M, DAS B, ROY S D, et al. Adaptive CNN filter pruning using global importance metric[J]. Computer Vision and Image Understanding, 2022, 222: 103511.

[143] WANG W, YU Z, FU C, et al. COP: customized correlation-based Filter level pruning method for deep CNN compression[J]. Neurocomputing, 2021, 464: 533-545.

[144] HE Y, KANG G, DONG X, et al. Soft filter pruning for accelerating deep convolutional neural networks[J]. ArXiv preprint arXiv:1808.06866, 2018.

[145] LIU Z, LI J, SHEN Z, et al. Learning efficient convolutional networks through network slimming[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2736-2744.

[146] LI H, MA C, XU W, et al. Feature statistics guided efficient filter pruning[J]. ArXiv preprint arXiv:2005.12193, 2020.

[147] CHANG J, LU Y, XUE P, et al. Automatic channel pruning via clustering and swarm intelligence optimization for CNN[J]. Applied Intelligence, 2022, 52: 17751-17771.

[148] LIU Y, WU D, ZHOU W, et al. EACP: An effective automatic channel pruning for neural networks[J]. Neurocomputing, 2023, 526: 131-142.

[149] NAKKIRAN P, KAPLUN G, BANSAL Y, et al. Deep double descent: Where bigger models and more data hurt[J]. Journal of Statistical Mechanics: Theory and Experiment, 2021, 2021: 124003.

[150] KESKAR N S, MUDIGERE D, NOCEDAL J, et al. On large-batch training for deep learning: Generalization gap and sharp minima[J]. ArXiv preprint arXiv:1609.04836, 2016.

致　　谢

博士求学之路漫长而充满挑战，至此完成学业，回首过往，心中充满感激之情。在此，我谨向在这段旅程中给予我关心、指导与支持的所有人致以最诚挚的谢意。

首先，我要向我的导师，申富饶教授，致以最深切的谢意。在我博士生涯的每一个关键时刻，申老师都给予我最坚定的支持与最细致的指导。是每周一次的个人讨论之路铺成了我的科研之路。从最初的科研启蒙，到独立思考与科研推进，申老师始终以严谨治学的态度和宽广的学术视野引导我。申老师用实际行动诠释了一位真正学者和师者的风范。

我还要特别感谢我的家人。感谢我的父母，始终如一地支持我、鼓励我，是你们的理解与无私奉献让我能够心无旁骛地追求学术理想。感谢我的丈夫，在我最疲惫和焦虑的时候，是你默默承担起生活的责任，为我撑起一片安心的天地。

感谢 426 和 415 的各位同学和朋友们，与你们在一起的日子让我收获了友谊、启发与成长。感谢郭苏涵、易梦军，你们的同行与鼓励使这段旅程不再孤单，也增添了无数温暖和欢笑。

博士论文的完成不仅是我个人努力的成果，更凝聚了所有在我生命中给予我温暖和力量的人。感谢你们的陪伴，使我有勇气走到今天，未来的路上，我会铭记这份恩情，继续努力，不负所学，不负所托。

学术成果

攻读博士学位期间完成的学术成果

已发表论文：

1. **Hui Dou**, Furao Shen, Jian Zhao, Xinyu Mu. Understanding neural network through neuron level visualization. Neural Networks, 2023.（*CCF-B* 类期刊）
2. **Hui Dou**, Baile Xu, Furao Shen, Jian Zhao. V-SOINN: A topology preserving visualization method for multidimensional data. Neurocomputing, 2021.（*CCF-C* 类期刊）
3. **窦慧**, 张凌茗, 韩峰, 申富饶, 赵健. 卷积神经网络的可解释性研究综述. 软件学报, 2024.（中文 *CCF-A* 类期刊）

在投论文：

1. **Hui Dou**, Xinyu Mu, Furao Shen, and Jian Zhao. Explaining Model Overfitting in CNNs via GMM Clustering.
2. **Hui Dou**, Furao Shen, and Jian Zhao. A Unified Approach to Explaining CNNs through Interpretable Properties.
3. **Hui Dou**, Xinyu Mu, Furao Shen, and Jian Zhao. Explaining the Convolutional Layer from the Neuron-Level Perspective.

攻读博士学位期间参与的科研课题

1. 面向增量式无监督学习的新型神经网络研究, 国家自然基金面上项目, 2023.1-至今.

2. 基于神经可塑性的脉冲网络高效学习机制与类脑智能系统, 科技部科技创新 2030 重大项目, 2022.1-至今.

3. 基于深度感知增量式联想记忆神经网络的信息融合系统研究, 国家自然基金面上项目, 2019.1-2022.12.

已授权发明专利

1. **窦慧**, 徐百乐, 申富饶. 一种支持拓扑结构保持的高维数据可视化方法, ZL201911179884.0, 2023.

攻读博士学位期间完成的教材

1. 申富饶. 简明神经网络 [M]. 电子工业出版社, 2025. (该教材为“十四五”规划省级重点教材). 参与核心内容编写和全书校对.