软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(1):159−184 [doi: 10.13328/j.cnki.jos.006758]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

卷积神经网络的可解释性研究综述*

窦 慧[1,2], 张凌茗[1,2], 韩 峰[1,2], 申富饶[1,3], 赵 健[4]

[1](计算机软件新技术国家重点实验室 (南京大学), 江苏 南京 210023)

[2](南京大学 计算机科学与技术系, 江苏 南京 210023)

[3](南京大学 人工智能学院, 江苏 南京 210023)

[4](南京大学 电子科学与工程学院, 江苏 南京 210023)

通信作者: 申富饶, E-mail: frshen@nju.edu.cn; 赵健, E-mail: jianzhao@nju.edu.cn

摘 要: 神经网络模型性能日益强大, 被广泛应用于解决各类计算机相关任务, 并表现出非常优秀的能力, 但人类对神经网络模型的运行机制却并不完全理解. 针对神经网络可解释性的研究进行了梳理和汇总, 就模型可解释性研究的定义、必要性、分类、评估等方面进行了详细的讨论. 从解释算法的关注点出发, 提出一种神经网络可解释算法的新型分类方法, 为理解神经网络提供一个全新的视角. 根据提出的新型分类方法对当前卷积神经网络的可解释方法进行梳理, 并对不同类别解释算法的特点进行分析和比较. 同时, 介绍了常见可解释算法的评估原则和评估方法. 对可解释神经网络的研究方向与应用进行概述. 就可解释神经网络面临的挑战进行阐述, 并针对这些挑战给出可能的解决方向.

关键词: 神经网络; 可解释性; 分类; 深度学习

中图法分类号: TP18

中文引用格式: 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, 35(1): 159–184. http://www.jos.org.cn/1000-9825/6758.htm

英文引用格式: Dou H, Zhang LM, Han F, Shen FR, Zhao J. Survey on Convolutional Neural Network Interpretability. Ruan Jian Xue Bao/Journal of Software, 2024, 35(1): 159–184 (in Chinese). http://www.jos.org.cn/1000-9825/6758.htm

Survey on Convolutional Neural Network Interpretability

DOU Hui[1,2], ZHANG Ling-Ming[1,2], HAN Feng[1,2], SHEN Fu-Rao[1,3], ZHAO Jian[4]

[1](State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

[2](Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China)

[3](School of Artificial Intelligence, Nanjing University, Nanjing 210023, China)

[4](School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China)

Abstract: With the increasingly powerful performance of neural network models, they are widely used to solve various computer-related tasks and show excellent capabilities. However, a clear understanding of the operation mechanism of neural network models is lacking. Therefore, this study reviews and summarizes the current research on the interpretability of neural networks. A detailed discussion is rendered on the definition, necessity, classification, and evaluation of research on model interpretability. With the emphasis on the focus of interpretable algorithms, a new classification method for the interpretable algorithms of neural networks is proposed, which provides a novel perspective for the understanding of neural networks. According to the proposed method, this study sorts out the current interpretable methods for convolutional neural networks and comparatively analyzes the characteristics of interpretable algorithms falling within different categories. Moreover, it introduces the evaluation principles and methods of common interpretable algorithms and expounds on the research directions and applications of interpretable neural networks. Finally, the problems confronted in this regard are discussed, and possible solutions to these problems are given.

* 基金项目: 科技部科技创新 2030 重大项目 (2021ZD0201300); 国家自然科学基金 (61876076)
收稿时间: 2022-01-20; 修改时间: 2022-04-01, 2022-06-13; 采用时间: 2022-08-10; jos 在线出版时间: 2023-02-22
CNKI 网络首发时间: 2023-02-23

Key words: neural network; interpretability; taxonomy; deep learning

近年来, 人工智能 (artificial intelligence, AI) 成为最重要的科学研究领域之一, 具有巨大的社会影响力, AI 技术被广泛应用于各个领域[1,2]. 随着可扩展的高性能基础设施的发展, AI 系统在许多领域已成为不可或缺的工具, 甚至在越来越多的复杂任务上完成度超过了人类水平[3,4].

然而, AI 系统在预测、推荐和决策支持方面的出色表现通常是通过采用复杂的神经网络模型来实现的, 这些模型隐藏了内部流程的逻辑, 此类模型通常被称为黑盒模型[5–7]. 神经网络模型通过非线性、非单调和非多项式函数来近似数据集中变量之间的关系, 这使得内部运行原理高度不透明. 神经网络模型经常因为错误的原因在训练集中得到正确的预测结果, 导致模型在训练中表现出色, 但在实践中表现不佳[8–11]. 因此, 神经网络的黑盒特性使得人类难以完全相信神经网络模型的决策.

人类有进一步了解神经网络模型的意愿. 对于决策能力较人类更差的模型, 希望可以在深度了解模型后发现问题并解决问题, 从而帮助模型改善性能. 对于决策能力与人类相似的模型, 希望可以解释决策结果, 从而使人类信任模型, 应用模型. 对于决策能力较人类更好的模型, 希望可以分析其决策机制, 帮助人类更好更深入地理解需要解决的问题.

可解释人工智能 (explainable AI, XAI)[12]研究致力于以人类可理解的方式解释人工智能模型[13], 使得人类能够理解模型的内部运行逻辑和决策结果, 为模型的故障排除和广泛使用提供方便. 可视化和解释神经网络模型的研究工作引起了越来越多的关注. 2018 年, 欧洲议会在通用数据保护条例 (general data protection regulation, GDPR) 中引入关于自动化决策的条款, 规定数据主体有获得自动化决策中涉及的相关解释信息的权利. 此外, 在 2019 年, 人工智能高级专家组提出了可信赖人工智能的道德准则. 尽管法律上对这些条款存在不同意见[14,15], 但普遍认同实施这样一个原则的必要性和紧迫性. 美国国家标准与技术研究院 (National Institute of Standards and Technology, NIST) 于 2020 年 8 月发布关于 XAI 的 4 项原则[16]: 可证明性 (解释结果可以被证据证明)、可用性 (解释结果能够被模型使用用户理解并对用户有意义)、准确性 (解释结果必须准确反映模型运行机制)、限制性 (解释结果能识别出不适合其自身运行的情况).

本文主要对卷积神经网络的可解释算法进行回顾和总结. 第 1 节主要针对模型可解释的定义和研究必要性进行讨论, 归纳受到了广泛认可的可解释定义和该领域内的常见词汇定义, 并从解决伦理问题、加强模型可靠性和优化模型性能等几个方面, 介绍对模型进行解释的必要性. 第 2 节对近年来模型可解释的相关研究进行简要分析, 首先对可解释研究具有代表性的研究和发展脉络进行梳理, 随后介绍目前已有的可解释算法的分类方法, 并总结现有分类方法普遍存在的问题. 第 3 节提出一种对神经网络可解释算法进行分类的新方法, 为理解可解释算法提供新角度. 依据新的分类方法, 对当前卷积神经网络的可解释方法进行梳理, 并对不同类别解释算法的特点进行分析和比较. 第 4 节中介绍常见的可解释算法的评估原则和评估方法. 第 5 节讨论可解释神经网络的研究方向、实际应用和当前面临的挑战, 对解释算法的目标和应用任务等与可解释性研究密切相关的内容进行阐述, 并就神经网络解释模型当前面临的问题进行简述, 针对这些挑战给出可能的解决方向. 第 6 节对全文进行总结.

1 模型可解释性的定义和研究必要性

直观而言, 神经网络的可解释算法是指通过能够被人类理解的方式对网络进行解释或呈现[13,17]. 解释算法的结果取决于指定的网络模型及其需要解决的问题. 神经网络模型的非线性结构使其具有高度不透明性, 使待解决的问题和网络结构间的关系模糊不清. 许多研究将神经网络模型当作黑盒进行处理, 并对该黑盒系统进行研究[18,19], 研究内容包括: 模型可解释的意义、需要解释的模型和任务以及解释模型的方式等.

如何定义神经网络的可解释性, 这是一个得到广泛讨论的问题. 目前对神经网络可解释性的定义还没有形成明确的共识, 不同研究中对神经网络可解释性的定义往往不同甚至还偶有矛盾[20]. 文献 [21] 中对可解释性的定义为: “以可理解的术语向人类解释的能力”. 定义中用“可理解的术语”表达这一概念是自包含的, 不需要进一步的解释. 解释是人类与自动化决策之间的交互界面, 既是自动化决策的准确代理, 又是人类可以理解的内容. 文献 [22] 基于文献 [21] 的定义进行了进一步阐释. 定义中的“解释”一词, 理想情况下应该是逻辑决策规则或者可以转换为

逻辑规则. 然而, 人们通常不要求在逻辑决策规则形式中明确地进行解释, 只要求能够生成一些可用于构造解释的关键要素. 定义中的“可理解的术语”, 应该来自与任务相关的领域知识 (或常识).

“解释”一词在英文中可以有许多对应词汇, 如 interpretation, explanation 等, 一些文献就神经网络中解释相关的词汇进行了详细的区分. 文献 [23] 中定义 interpretation 是将抽象概念 (例如预测类) 映射到人类可以理解的领域 (可解释域) 中; 而 explanation 是可解释域中的特征集合, 这些特征对给定样本产生的决策 (例如分类或回归) 做出了贡献. 文献 [21,24–27] 对网络可解释性与网络其他特性间的区别和联系进行了分析, 如解释性 (interpretability) 与算法复杂度 (completeness)、解释性与算法准确度 (accuracy) 和算法保真度 (fidelity) 等, 这里不一一罗列.

越来越多的公司在其应用程序中嵌入神经网络模型, 不透明性始终被认为是神经网络模型的主要缺陷之一[28], 依赖不透明的模型可能会导致人类不能完全理解网络所采用的决策[29,30], 不能正确估计该决策带来的风险, 导致潜在的安全危机和信任危机[31,32]. 尤其在必须保证模型可靠性的高风险决策场景中, 例如医学诊断[33,34]、自动驾驶汽车[35,36]和刑事司法[37]等. 当前普遍认为可解释性是人工智能模型可以被信任的关键要素之一[38–41]. 因此, 模型可解释性在增加系统可靠性方面的重要性显而易见[42].

模型在历史数据集上进行训练, 可能会导致引入许多不易被察觉的偏见. 具有偏见的规则可能被较深地隐藏在受过训练的模型中, 这些偏见规则可能被视为一般规则. 模型的不透明性隐藏了这些潜在问题[43], 因此判断模型在涉及某些问题 (如性别、种族等问题[44]等) 时模型结果是否公平变得十分困难, . 除社会道德问题外, 模型的不透明性还会影响问责制[45]、产品安全[46]和工业职责划分等问题[47].

除可以解决系统可靠性问题和系统偏见问题外, 神经网络可解释性的研究也有助于优化网络从而提升模型性能, 甚至设计出更为有效的网络模型[48]. 对于神经网络在复杂任务上应用, 大多数研究主要关注模型性能的提升, 并未分析网络在任务中表现良好的原因[49–51]. 虽然普遍认为神经网络的性能取决于其结构, 但是目前人类对神经网络性能与其结构间的关系几乎没有系统性的理解[36]. 从这个角度看, 当前对于网络的认知仍存在一些待解决的问题: 如网络结构与其预测性能间的关系、性能良好神经网络的结构特征、神经网络的优化问题等. 如果可以解释神经网络结构中不同部分的作用及其相互间关系, 了解神经网络性能变化的原因, 有助于进一步改善模型, 进而设计出性能更优的模型[52,53].

2 相关研究概述

当前神经网络可解释性研究成果众多, 本节首先对可解释性研究具有代表性的研究成果以及发展脉络进行简要梳理, 随后对可解释性研究的相关综述进行整理, 总结其中观点.

(1) 代表性研究简介

近年来, 神经网络可解释性研究蓬勃发展. 2013 年, Simonyan 等人[54]提出两类 CNN 可视化方法: 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, 2016 年, Zhou 等人[55]提出类激活映射方法 (class activation mapping, CAM), 在 CNN 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, 基于 CAM 模型的优化 Grad-CAM 算法[56]同样可以通过热力图解释网络关注区域, 而且适用于更多网络结构和任务, 因此得到广泛应用.

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, 2017 年, Bau 等人[57]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 [58], 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.

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. 2016 年, Ribeiro 等人[59]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, 文献 [60] 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.

(2) 相关综述研究

随着可解释性研究的发展, 出现许多就神经网络可解释性研究进行讨论总结的文献. 一些文献针对可解释性研究中的关键问题展开详细的讨论. 文献 [24] 聚焦于解释网络内部数据表示这一问题. 探讨了特定输入导致特定输出的原因、网络自身包含哪些信息等问题, 并提出解释神经网络过程的两种思路: 通过创建一个代理模型来近似原始模型, 或者通过创建显著图来突出显示最相关的一小部分计算. 文献 [23] 着眼于从概念上区分可解释性的不同定义, 从确切的神经网络结构和应用领域中抽象出来, 概述了解释深度神经网络模型的技术. 文献 [61] 从解释模型在解释内容、解释受众和解释目的方面的不同, 概述不同类型的解释方法并评论它们在实践中的实用性. 文献 [42] 从不同角度出发, 讨论可解释机器学习中面临的一系列关键技术挑战, 其中包括可解释机器学习中的一系列经典问题, 例如, 为表格数据构建稀疏模型、加法模型的挑战、基于案例推理的挑战、有监督和无监督问题求解的挑战、降维问题的挑战、可解释的强化学习的挑战等, 并对不同挑战对应的可解释技术提供分析. 文献 [62] 重点讨论了解释神经网络模型在对抗性攻击中的脆弱性和鲁棒性这一问题, 利用输入中的扰动, 将损失结果可视化, 从而进行定性分析, 同时提出评估解释定义模型内在稳健性的量化指标, 进而对模型进行定量分析.

在对神经网络进行解释的诸多方法中, 网络特征可视化是探索神经网络的最直接方式[63]. 网络特征可视化为诊断网络提供了技术基础, 因此网络特征可视化是可解释性研究中的重要手段. 文献 [64] 对具有代表性的卷积神经网络 (convolutional neural network, CNN) 可视化方法进行回顾, 并讨论 CNN 可视化的实际应用, 展示网络可解释性在网络设计、优化、安全增强等领域的重要性. 文献 [65] 同样是对 CNN 的可视化特征进行分析, 但更关注对 CNN 网络中间层的可视化表示方法进行归纳, 并对不同方法进行表征信息诊断, 分析不同的目标类别对应的特征空间.

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. 文献 [66] 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, 文献 [66] 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. 文献 [67] 中将解释方法分为: 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. 文献 [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. 文献 [68] 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 4 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. 文献 [24] 同样从解释算法的目的出发, 将解释算法分为 3 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地概括. 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, 本文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. 本文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.

3 新型可解释算法的分类方法

文献 [20] 讨论了两种流行的可解释性概念: 对人类透明和事后解释. 对于对人类透明模型, 研究更关注模型

本身: 算法的收敛性、算法复杂度、模型参数的具体含义等. 事后解释是指对于一个给定的训练完成的模型, 通过易于理解的内容 (例如输入样本变量信息等) 来解释神经网络模型的预测结果 (例如类别等)[59,69]. 事后解释的算法可以在不阐明模型工作机制的情况下解释预测结果.

对于事后解释的算法, 不同解释算法的表达形式不同. 使用逻辑规则 (通常是 if-then 形式) 可以提供最清晰明确的解释, 但在实际应用中通常很难实现将解释信息完全通过逻辑规则进行表述. 相较于逻辑规则, 其他表达形式的解释没有清晰的解释文字, 也被称为“隐性解释”. 严格来说, “隐性解释”本身并不是完整的解释, 需要进一步的人工解释, 这通常是通过人们看到它们时根据经验对解释进行进一步补充完成的. 例如, 常用于事后解释的显著图, 其本身是特定输入样本上的掩码. 通过查看显著图, 可以得到的解释是模型做出当前预测结果是因为输入样本上的某些区域对模型影响较大. 如果这些区域对应于一些人类可理解的概念 (如动物器官、身体部位等), 则说明人通过对解释结果 (显著图) 的再加工 (识别概念), 完成对网络的解释. 因此, 简单通过表达形式的不同很难对解释算法进行完整地分类.

本文针对事后解释的神经网络解释算法, 提出一种新型分类方法. 在提出的分类方法中, 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. 综上所述, 不同类别的具体分类方法及概念总结见表 1. 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 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.

表 1 事后解释的神经网络解释算法新型分类方法

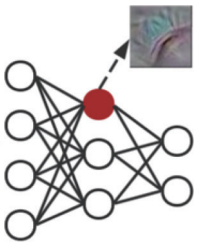
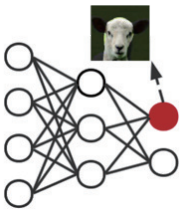
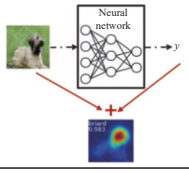
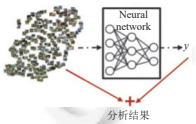
方法	概念
基于网络的解释方法	针对网络自身单元的属性进行解释, 不依赖于特定的输入输出 (1) 理想样本: 特定的网络单元自发生成最感兴趣的输入样本 (2) 真实样例: 网络单元从输入样本中寻找感兴趣样本的代表
基于输入的解释方法	针对指定输入样本, 对网络给出的输出结果进行解释 (1) 单一输入的解释: 对特定输入样本与输出结果的关系进行解释 (2) 多个输入的解释: 为一类相似的输入样本的输出结果提供统一的解释

基于网络的解释方法针对网络自身的单元 (特征图、神经元等) 的属性进行解释, 不依赖于输入输出. 这类方法主要关注神经网络本身学习到的模式, 不关注在特定输入情况下, 神经网络的表现. 深度神经网络通常不能像线性模型那样找到线性解释模型, 因此需要其他方法来解释神经网络单元. 针对这类解释方法, 一个直观的方法是可视化出指定网络单元 (例如, 隐藏神经元等) 最感兴趣的模式. 基于不同网络单元的反馈结果, 可以得到关于网络内部运作机制的启示. 此外, 还可以采取更主动的方法来使网络单元更具可解释性. 由于高层隐藏神经元通常会学习到难以解释的混合模式, 因此, 此类方法可以采用一定方式 (例如, 加过滤器等) 将不同模式进行分离, 使得指定网络单元只在固定模式下被激活. 通过不同样本激活网络中的不同部分 (例如, 过滤器等), 使得解释算法更加容易理解 (例如, 发现某过滤器只被动物的头部激活等)[70].

根据网络单元感兴趣模式的生成方式, 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. 基于理想样本的解释是指特定的网络单元根据自身习得的感兴趣模式和激活情况, 自发生成最感兴趣的输入样本, 这类样本真实并不存在. 以表 2 中理想样本的样例示意图为例, 对于给定网络, 通过解释方法生成当前选定神经元 (红色表示) 最感兴趣的样本, 样本在样本集中不存在. 基于真实样例的解释则是指从输入样本集中寻找一个或一组样本, 使得指定网络单元的激活程度最高, 则说明输入样本中明显包含该网络单元感兴趣的模式, 因此可以以这

类输入样本作为该网络单元感兴趣样本的代表. 以表 2 中真实样例的样例示意图为例, 在样本集中找到使得选定神经元 (红色表示) 激活最大的输入图像, 以此输入样本作为该网络单元感兴趣样本的代表. 两类子方法都以展示网络单元感兴趣的输入为目标, 但生成感兴趣输入的方式不同.

表 2 新型分类方法说明

分类	子类	分析	样例示意图
基于网络的解释方法	理想样本	理想样本的解释方法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. 典型方法是通过最大化某个神经元、通道或层的激活, 生成一个具有代表性的输入. 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	真实样例的解释方法是指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, 则其中明显包含该网络单元感兴趣的模式, 因此可以以这类输入样本作为该网络单元感兴趣样本的代表. 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	
基于输入的解释方法	单一输入的解释	单一输入的解释方法对特定输入进行解释, 通常利用目标输入的信息 (例如, 其特征值、梯度), 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, 对于输入网络的某一单一样本, 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	多个输入的解释方法是为一类相似的输入样本的输出结果提供统一的解释. 分解在输入样本集中具有普遍性的特征, 描述每个特征如何或在多大程度上对网络的输出结果做出贡献, 进而对模型进行解释. 示意图如图所示, 对于输入的一类样本, 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	

注: 表中的网络结构均为示意图, 不代表真实需要解释的网络结构

基于输入的解释方法针对一个或一类输入样本, 对网络给出的反馈结果进行解释. 这类方法主要关注特定的输入样本和输出结果间的关联, 而不关心神经网络的内部运行机制. 一类典型示例是利用显著图解释特定输入与输出之间的关系. 通过显著图, 可以解释输入样本会得到某个输出结果, 是因为网络专注于输入样本中的某些部分, 即输入样本中的哪些区域对输出结果的贡献最大. 最理想的结果是, 这些部分对应于一些人类可理解的概念.

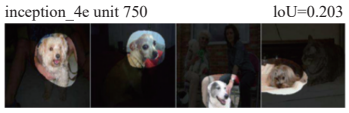
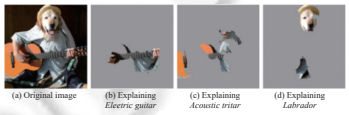
根据神经网络解释算法的输入方式, 可以将基于输入的解释方法分为单一输入的解释和多个输入的解释两个子类. 基于单一输入的解释是指对特定输入样本进行解释, 为输入样本的不同区域或像素分配重要度值或敏感度值以解释其对输出结果的影响. 基于单一输入的解释通常利用目标输入的信息 (例如, 其特征值、梯度). 以表 2 中单一输入的解释的样例示意图为例, 对于给定的输入样本, 结合网络输入结果, 给出当前结果中网络对该样本的感兴趣区域, 从而对该样本的预测结果进行解释. 基于多个输入的解释则是为一类相似的输入样本的输出结果提供统一的解释. 分析在输入样本集中具有普遍性的特征, 描述每个特征如何或在多大程度上对网络的输出结果做出贡献, 进而对模型进行解释. 多个输入的解释试图推广到尽可能广泛的输入范围 (例如, 规则学习中的顺序覆盖, 特征重要性排名的边际贡献等). 以表 2 中多个输入的解释的样例示意图为例, 对于输入的一类样本, 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. 单一输入的解释侧重于对个体预测的解释, 而多个输入的解释可以一定程度上达到对模型整体决策逻辑的理解.

目前大部分解释算法是对单一输入进行解释. 对于指定的单一输入, 输入的特征是输入图像的所有像素. 解释

算法根据图像的哪些区域 (像素) 对分类结果的贡献最大对该输入样本进行解释. 例如, 可以根据输入特征 (即所有像素) 或某些变体通过敏感性分析来计算属性. 对于多个输入的解释, 如何对样本中包含的属性进行分类是一个重要的问题. 使用输入特征的像素是一种最直接的方法. 其他方法包括定义概念并分析不同预测结果对不同概念感兴趣程度 (例如, 计算某一类预测结果对某一概念的存在的敏感度等). 文献 [58] 中将概念 (例如, 条纹) 由平面的法向量表示, 该平面将网络隐藏层空间中的有条纹和无条纹训练示例分开. 因此, 可以计算预测结果 (例如, 斑马) 对概念 (条纹) 的敏感程度, 从而获得网络对输入样本的解释.

通过新型分类方法, 本文对当前可解释算法进行分类表示, 分类结果如表 3 所示, 从算法类别、各类算法的特点、各类算法的代表论文及算法的典型效果示例等几个方面进行汇总, 有利于论文的快速索引. 本节的后续小节将对表 3 中涉及的论文进行详细说明.

表 3 新型分类方法下的 CNN 可解释算法

分类	子类	特点	文献	典型效果示例
基于网络的解释方法	理想样本	通过最大化网络单元的激活值生成指定网络单元的理想样本	[54,71–76]	 [54]
	真实样例	(1) 多模态分析: 结合多种解释方式, 如可视化解释结果和文本解释结果, 多角度为决策结果进行分析	[57,77,78]	 [57]
		(2) 特征提取: 将网络中指定单元能够学习到特征显性的进行表示	[70,79,80]	
		(3) 特征拆分: 对网络习得的特征进行拆分, 特征包括物体颜色、物体组成部分、物体尺度、物体方向等, 通过各种方法检测如何学到不同特征	[81–83]	
基于输入的解释方法	单一输入的解释	(1) 类激活映射: 通过生成类激活图来展示网络识别特定样本时感兴趣的区域	[55,56]	 [59]
		(2) 基于梯度和反向传播: 对于给定样本和结果, 将输出结果逐层反向传播至输入空间, 显示感兴趣区域和程度	[56,84–91]	
		(3) 模型未知方法: 在不知道具体模型的情况下研究模型对样本中特征的感兴趣程度	[59,60,92–94]	
		(4) 基于扰动的方法: 对样本部分信息增加扰动并观察输出结果的变化情况, 确定样本中扰动部分对结果的影响	[95–99]	
	多个输入的解释	较为普遍的做法是通过对样本类别、样本特征、样本包含属性或像素等进行一定方式的提取和结合, 寻找其中的规律	[58,100–103]	 [101]

3.1 基于网络的解释方法分类

3.1.1 基于理想样本的解释方法

基于理想样本的解释方法中最为典型方法是通过最大化某个神经元、通道或层的激活值, 找到一个具有代表性的输入样本, 这种方法被称为激活最大化方法 (activation maximization, AM)[71]. 激活最大化方法是一个最优化方法, 最初被用于非监督网络, 文献 [54] 首次将这一方法运用于卷积神经网络, 用于解决图像分类任务下的深度

网络可视化问题. 文献 [54] 中介绍了一种可视化模型指定类别的方法, 给定一个卷积神经网络和一个类别, 在输入空间生成一个选定类别最感兴趣的样本. 具体方法是, 针对卷积神经网络的全连接分类层中代表指定类别 c 的神经元, 类别 c 的激活值为 $S_c(I)$, 随机输入样本 I 激活神经元, 在反向传播的过程中, 保持网络权重不变, 迭代优化输入样本 I, 最终获得使指定该类别神经元激活值 $S_c(I)$ 最大的输入样本, 从而获得卷积神经网络中指定类别学习到的特征. 公式 (1) 中的 λ 表示正则项参数. 该可视化结果可以将网络在不同类别下学习到的特征输出:

$$\arg\max_I S_c(I) - \lambda \|I\|_2^2 \tag{1}$$

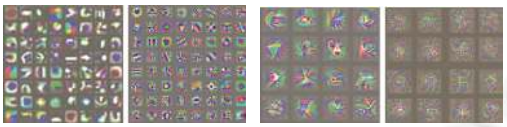
激活最大化方法可以泛化为更为一般化的方法, 从而得到网络中的任意神经元的理想样本. 针对网络中任意神经元 i, 寻找一个最优输入样本 x^*, 使神经元的激活函数 $a_i(x^*)$ 最大化, 即:

$$x^* = \arg\max_x (a_i(x) - R_\theta(x)) \tag{2}$$

其中, $R_\theta(x)$ 为正则项. 不同类别的神经元对应的最优输入样本称为该类别的类模型可视化结果. 文献 [54] 通过展示网络中不同类别的类模型可视化结果证实算法的正确性.

后续研究发现, 高频噪声是导致激活最大化方法可视化效果不佳的主要原因[72,73]. 为了避免高频噪声问题, 对输入样本 I 添加约束, 使生成样本更接近于真实样本[74]. 为了生成更加清晰且易于理解的可视化结果, 文献 [75] 分析了 4 种正则化方法: L_2 衰减、高斯模糊、小范数的被裁减和小贡献的被裁减对算法的影响, 优化算法中的不同缺陷. 比较通过不同正则化超参数生成的理想样本, 发现不同超参数优化图像的效果不同: 一些超参数有助于显示高频信息, 另一些有助于显示低频信息; 一些超参数的生成图像包含密集像素信息, 另一些超参数的生成图像只包含重要区域的稀疏轮廓信息.

文献 [74] 提供了另一种理想样本的优化思路, 即将图像模糊算子和图像去模糊算子应用于理想样本生成. 这两类算子通过使用高斯低通滤波器进行卷积和去反卷积运算来实现. 图像模糊操作主要用于滤除高频噪声, 图像去模糊操作主要用于抵消图像模糊操作后引起的模糊. 该算法可以更好地提取背景和前景中的细节信息. 文献 [74] 应用该算法对不同网络中卷积层过滤器生成感兴趣图像, 证明优化后的生成图像更具解释性. 算法结果如图 1 所示, 对 VGG 网络中不同卷积层的不同过滤器生成感兴趣图像, 可以看出不同过滤器提取到的特征不同. 为了获得更接近原始图像的生成图像, 文献 [76] 增加一个图像生成器网络用于合成图像, 通过不断优化生成器输入, 使网络中指定神经元激活值最大, 从而获得生成器的输出图像, 即指定神经元感兴趣的图像.



(a) conv2_1 (b) conv3_1 (c) conv4_1 (d) conv5_1

图 1 不同卷积层的过滤器的可视化结果

基于理想样本的解释方法共同点在于, 此类方法着重于展示模型中网络单元学习到的特征, 即生成使网络单元激活程度最高的理想样本. 对于不同的基于理想样本的解释方法, 其方法可视化效果不同. 此类研究通常致力于得到更好的可视化效果, 研究方向包括: 减少生成样本噪声和生成具有语义信息的样本等.

基于理想样本的解释方法优点在于, 基本原理简单, 可以展现模型网络单元学习到的特征, 一定程度上还原网络运行原理, 与人类对于神经网络的认知方式相似, 因此整个过程易于理解. 这类方法的缺点在于, 针对特定网络构建目标函数困难, 迭代优化过程不易, 反向传播优化过程存在信息丢失问题. 且经这类方法获得的理想样本, 语义信息往往不够明确, 与人类认知事物的方式难以匹配, 因此生成的样本解释性较弱. 如表 3 中理想样本的典型效果示例所示, 无法清晰识别生成的理想样本所代表的类别. 这类方法用于网络单元的理解和研究, 通常对结构较为简单的网络结构具有较好的解释效果.

3.1.2 基于真实样例的解释方法

基于真实样例的解释方法是指从数据集中寻找一个或一组样本, 使得指定网络单元激活程度较高, 则说明这

组样本中明显包含该网络单元感兴趣的模式, 因此可以以这组样本作为该网络单元感兴趣样本的代表. 常见的解释方法包括: 多模态分析方法、特征提取方法、特征拆分方法等.

(1) 多模态分析方法

多模态分析方法结合多种解释方式, 如可视化结果和文本说明, 多角度对决策结果进行分析. 为了阐明 CNN 模型的固有性质及训练方法对 CNN 模型的影响, 文献 [57] 中提出一种被称为网络分解的通用框架, 通过评估单个隐藏单元和一组语义概念之间的相关性对网络进行解释. 在卷积神经网络中, 利用一组通用的具有语义信息的通用数据对每个卷积层不同隐藏单元进行评分, 以高分值数据的语义信息作为隐藏单元的标签, 标签内容为物体、部件、场景、纹理、材料和颜色等信息. 该算法使用的图像数据集 Broden 具有像素级别标记的视觉概念. 算法原理如图 2 所示, 该方法通过 Broden 数据集中的视觉概念评估每个单元. 对于每个样本 x 和 Broden 数据集中的概念 c, 有与 x 尺度相同的二进制掩码 $L_c(x)$ ($L_c(x)$ 在每个像素上的值表示样本在这个像素上是否具有概念 c). $A_k(x)$ 表示样本 x 在卷积核 k 作用下的激活图, 将 $A_k(x)$ 插值放大并转换为与 x 尺度相同的二进制掩码 $M_k(x)$. 卷积核 k 与概念 c 间的对应关系通过交并比 $IoU_{k,c}$ 评估:

$$IoU_{k,c}=\frac{\sum|M_k(x)\cap L_c(x)|}{\sum|M_k(x)\cup L_c(x)|} \tag{3}$$

其中, $IoU_{k,c}$ 的值是卷积核 k 在检测概念 c 时的准确率; 如果 $IoU_{k,c}$ 超过阈值, 可以表示卷积核 k 具有检测概念 c 的能力. 文献 [77] 基于上述使用思想, 研究神经网络卷积核组合的捕获信息能力, 提出一种使用多个卷积核信息对同一概念进行解释的算法. 文献 [57] 将网络单元的可解释性与语义信息相关联, 通过实验分析单元的可解释性等同于单元的随机线性组合这一假设; 并比较不同初始化参数下网络单元的语义信息匹配能力, 从而分析不同参数对网络单元可解释性的影响.

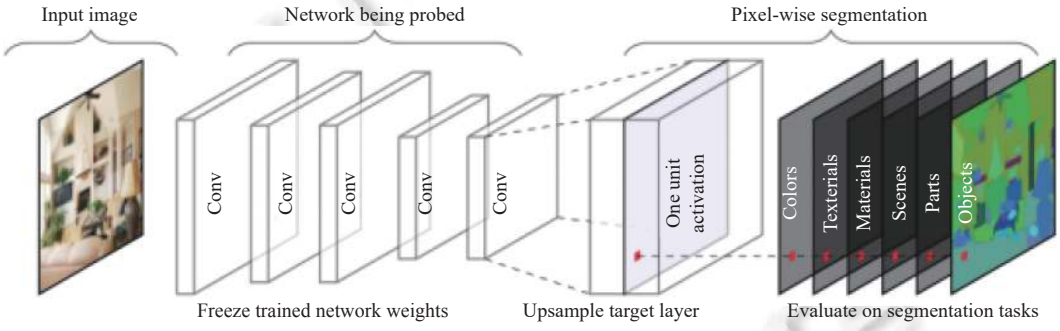


图 2 网络分解方法测量给定网络单元的语义信息[57]

文献 [78] 中提出一个被称为解释基础分解的解释方法框架, 提供热力图解释和带有标签的语义解释, 用于为分类网络提供视觉解释. 文中将输入图像的神经元激活值分解为预先训练的语义可解释元素, 类别权重被分解为一组可解释的基向量, 每个分量对应一个标签概念, 将该概念的分量投影到热力图中, 记录其在热力图中的激活程度, 量化每个概念对最终预测的贡献. 文献 [78] 使用该框架对常见的视觉识别网络提供解释, 通过 AMT (Amazon mechanical turk) 对解释结果进行评估, 以证明框架解释结果的合理性.

(2) 特征提取方法

特征提取方法主要将网络中指定单元 (对 CNN 而言通常是卷积核) 能够学习到特征显性地进行表示. 为了提高 CNN 网络单元表达信息的可解释性, 文献 [70] 提出一种将传统 CNN 修改为可解释 CNN 的方法, 阐明 CNN 高层卷积层中的知识表示. 算法用高层卷积层中的每个卷积核表示一个特定的对象部分. 在学习过程中, 算法会自动为高层卷积层中的每个卷积核分配一个对象部分, 学习到的显式知识表示可以帮助人们理解 CNN 内部的逻辑, 即 CNN 记住哪些模式进行预测. 图 3 显示了传统 CNN 与可解释 CNN 的区别: 传统 CNN 中的高层过滤器可能学习到混合模式, 网络可解释性弱. 可解释 CNN 中的过滤器只能被指定对象部分激活, 因此可以在分类任务中清晰获得网络学习到的特征. 文献 [70] 将算法应用于 4 个不同结构的 CNN 网络中, 通过对可解释卷积层中的特征图

进行可视化来说明不同过滤器关注的语义信息, 并通过对物体可解释性和位置不稳定性两个指标来定量评估过滤器的语义信息准确性. 对于给定的图像样本和对象类别, 图像分割任务可以重写为推断属于对象类别的像素, 与网络特征提取方法思路一致, 因此可以用图像分割任务结果衡量特征提取的有效性. 为了在样本中更加精准地提取属于不同类别特征的像素信息, 文献 [79] 对 CNN 模型进行修改, 模型在训练期间受到约束, 使得对图像分类重要的像素的权重值更大, 即将更有影响力的特征突出. 文献 [79] 将新算法与监督学习方法在分割任务中的结果进行比较, 证明新算法在提取物体特征中的有效性. 文献 [80] 针对知识蒸馏解释性问题, 提出一种提取在神经网络中间层编码的视觉概念, 并基于这些概念解释知识蒸馏成功的原因. 文中提出关于知识蒸馏是否有助于神经网络从原始数据中学习到更多的视觉概念的 3 种假设, 并通过在实验中量化网络中的视觉概念依次验证假设, 给出结论.

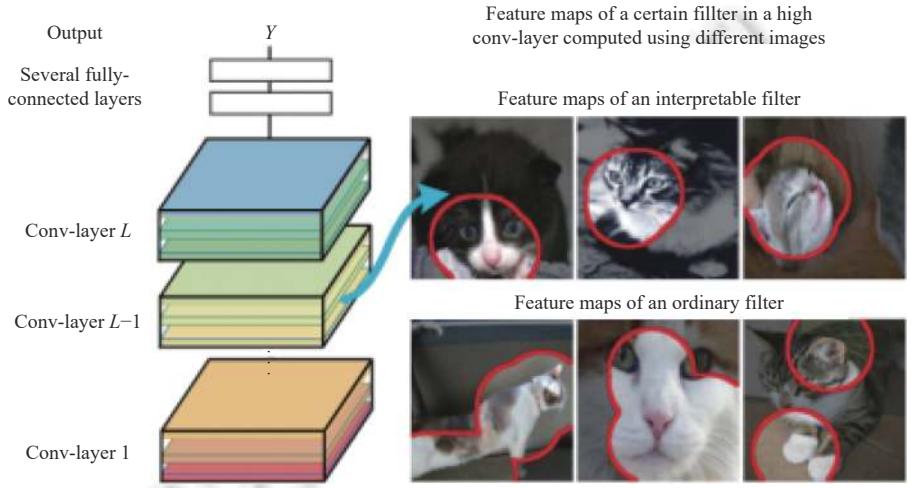


图 3 传统卷积网络与可解释卷积网络区别[70]

(3) 特征拆分方法

特征拆分方法将网络习得的特征进行拆分, 包括物体颜色、物体组成部分、物体尺度、物体方向等特征, 通过算法检测网络习得的不同特征. 文献 [81] 中提出, 神经网络的单个神经元可以检测多种类型的特征. 对输入样本特征进行分解, 通过生成不同特征来激活神经元, 从而合成不同特征下神经元感兴趣的图像. 通过对每层的选定神经元进行可视化, 发现在低层神经元不能明显区分特征的不同面. 高层神经元的可视化结果复杂, 并且可以检测特征的不同面. 在输出层, 神经元被训练为对固定类别作出反应, 学习到的特征是多样的. 为揭示 CNN 卷积层内编码的物体对象部分, 文献 [82,83] 提出一种通过提取图形模型解释图 (explanatory graph) 对卷积神经网络进行解释的方法. 考虑到每层中不同过滤器检测图像的不同特征, 算法自动从每个滤波器中提取图像的不同部分, 并构造解释图. 在解释图中, 每个节点表示一种物体模式, 不同节点间的连接表示物体模式的协同激活关系和空间关系. 文献 [82] 使用解释图对 4 种不同结构 CNN 模型进行解释, 通过解释结果可视化、评估节点表示对象的一致性和在小样本局部定位任务中测试节点可迁移性等实验, 验证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可以表示不同输入图像中的相同对象部分.

(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的共同点在于, 此类方法通过具有代表性的样例解释数据集, 并分析这些样例对模型的影响. 这些具有代表性的样例有助于识别数据中的偏差, 使模型对数据集的变化更加稳健. 基于真实样例的解释方法普遍适用于解释卷积神经网络和数据集包含语义信息的情况. 如表 3 中真实样例的典型效果示例所示, 用网络单元最感兴趣的图像区域代表该网络单元习得的特征信息.

根据解释方式的不同, 基于真实样例的解释方法分为多模态分析方法、特征提取方法、特征拆分方法等, 不同解释方法的特点归纳如表 4 所示: (1) 多模态分析方法的优点在于解释信息同时具有图像信息及对应的文本信息, 因此该方法的解释清晰且具有说服力. 缺点在于这类方法普遍存在解释模型训练较为困难的问题, 要求数据集

标注语义信息, 因此对训练数据集的筛选也具有更为严格的要求; (2) 基于特征提取的解释方法将网络基础单元与输入样本信息对应表示, 通过数据集中的数据解释网络单元学习到特征, 这类方法展示的解释信息可视化效果好, 可解释性较强. 但使用这类方法对单个网络单元提取到的特征信息进行规律总结较为困难, 可能存在提取的特征信息不具有人类可理解概念的问题, 因此很难实现对所有网络单元进行解释; (3) 基于特征拆分的解释方法对输入数据集进行拆分, 通过拆解后的特征信息对网络进行解释. 拆解的特征信息容易被人类理解, 因此解释结果也具有较强的可理解性, 且解释结果直接明了. 但这类方法普遍存在建模较为困难、计算较为复杂的问题.

表 4 基于真实样例的解释方法分析

类别	优点	缺点	适用范围
多模态分析方法	解释信息同时具有图像信息及对应的文本信息, 解释清晰且具有说服力	解释模型训练较为困难, 对网络结构和数据集要求严格	适用于解释卷积神经网络; 适用于解释数据集包含语义信息的情况
特征提取方法	解释结果可视化效果好, 可解释性较强	对单个网络单元提取到的特征信息可能不具有人类可理解概念, 很难实现对所有网络单元的解释	
特征拆分方法	拆分的特征信息易于理解, 可理解性强, 解释结果直接明了	解释模型建模较为困难, 计算较为复杂	

4 基于输入的解释方法分类

4.1 基于单一输入的解释方法

基于单一输入的解释方法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以解释其对输出结果的影响. 基于单一输入的解释方法通常利用目标输入的信息 (例如特征值、梯度等). 常见的单一输入的解释方法包括: 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、基于扰动的方法等, 其他较为典型的研究方法在其他方法中进行介绍.

(1) 类激活映射方法

类激活映射方法生成类激活图来展示卷积神经网络感兴趣的区域. 许多研究[104–106]已证明, 卷积神经网络中的卷积层, 具有显著的定位物体能力, 但使用全连接层进行分类时, 这种定位能力会丢失. 为始终保持卷积神经网络的定位能力, 文献 [55] 中提出类激活图 (class activation mapping, CAM) 方法, 通过改变网络结构保持网络定位能力. 网络主要由卷积层组成, 在最终输出层之前, 使用全局平均池化图代替原卷积神经网络中的全连接层. 通过这种结构, 可以将输出层的权重投影回卷积特征图, 形成类激活图, 从而识别图像区域的重要性. 一个特定输入样本的类激活图表示模型当前类别在识别该输入样本时的感兴趣区域和感兴趣程度. 文中通过 CAM 方法修改不同网络的结构, 并评估修改后网络的定位能力和分类能力, 实验证明 CAM 在实现定位的同时可以保持网络分类性能.

(2) 基于梯度和反向传播的方法

对于给定的输入样本, 根据网络给出的输出结果, 将输出结果逐层反向传播至输入空间, 生成与输入样本大小相同的特征图像, 描述网络对该样本的感兴趣区域和感兴趣程度[84]. 为了解释网络对单一输入图像的关注度, 文献 [54] 通过显著图为每个像素分配一个重要度分数. 文中指定输入图像 I 和输出类别 c, 通过泰勒一阶展开将输出类别得分 $S_c(I)$ 表示为输入图像 I 的线性函数:

$$S_c(I) \approx w_c I + b_c \tag{4}$$

通过梯度反向传播方法计算出该线性函数的权重 $w_c = \dfrac{\partial S_c}{\partial I}|_I$, 使用权重 w_c 生成的显著图显示了类别 c 对输入图像 I 的感兴趣区域. 文中通过展示不同输入图像的显著图来证实算法的准确性. 文献 [85] 提出一种优化显著图的 SmoothGrad 方法, 可以在基于梯度的显著图中增加平滑梯度, 减少视觉噪声. 文章在图像分类任务的网络中使用 SmoothGrad 方法, 并比较不同平滑度下的显著图效果. 实验结果表明, 使用 SmoothGrad 方法的显著图相较于原始显著图具有更好的可视化效果. 文献 [86] 认为对神经网络的显著图解释通常依赖于两个关键假设: 使用损失函数的一阶近似而忽略了损失曲率等高阶项; 孤立评估每个特性的重要性而忽略特性的相互依赖性.

文献 [86] 放宽对两个假设的要求, 在泰勒展开中保留 Hessian 项以使用损失函数的二阶近似, 并以封闭形式计算 Hessian 项, 从而得到更好的显著图解释效果. 文章将新方法应用于图像分类任务的网络解释, 实验结果表明, 该方法可以消除噪声且使得显著图解释信息与原图对象更具有一致性. 在此类基于梯度的方法中, 偏差 b 通常被忽略. 文献 [87] 中提出一种偏差反向传播算法, 它从输出层开始, 迭代地将每一层的偏差归因于其输入节点, 构建局部线性模型 $g(x)=wx+b$. 文中将该方法用于图像分类任务, 在实验证明该方法可以产生互补且可理解度高的解释信息.

反卷积方法同样是将网络中间层的特征图映射到输入样本空间, 从而展示在每层的特征图中模型学习到的特征. 反卷积方法最初在文献 [88] 中被提出, 用于构建提取低级和中级图像特征的特征检测器, 反卷积模型与卷积模型具有相同的结构, 包括卷积运算和池化运算, 但反卷积模型的结构顺序与卷积模型相反, 输入输出顺序也相反, 卷积模型将输入图映射到特征图, 反卷积模型将特征图映射到输入图. 文献 [89] 使用反卷积模型, 利用每个反卷积层的输出结果最小化与输入图像间的误差, 得到重构后的图像. 该图像可以表示当前特征图学习到的特征. 文中应用反卷积方法对 4 层分类网络进行特征提取, 实验结果表明从这些模型中提取的特征优于其他特征学习方法提取的特征. 文献 [90] 通过反卷积学到的特征, 进一步解释卷积神经网络. 通过对模型每层特征图进行可视化, 可以看到模型在低层网络中主要学习到简单边缘特征, 在中层网络中主要学习到边的连接特征, 在高层网络中主要学习到局部或全部的物体. 这与当前对卷积神经网络的认知一致.

反卷积方法需要记录池化区域最大值的位置, 文献 [91] 针对这一问题, 结合文献 [54] 中的方法, 提出一种仅由卷积层构成的新架构 Guided Backprop, 新架构具有良好的可视化解释效果. 为了量化新架构的效果, 文中在 3 个数据集上进行了分类实验, 证明新架构可以保持分类准确率. 与之相似的是文献 [56], 基于 CAM[55]的基本结构, 文献 [56] 中提出一种基于梯度的类激活映射方法 Guided Grad-CAM (gradient-weighted class activation mapping). 它可以对任何基于卷积神经网络的任务生成视觉解释, 而不需要改变结构或重新训练. 给定一个图像和一个感兴趣的类作为输入, 模型向后传播, 通过特定任务的计算获得类别的原始分数, 然后将该信号反向传播到卷积特征图以查看物体位置, 并将热力图与导向反向传播结合, 得到高分辨率具有特定概念的可视化结果. 不同方法比较结果如图 4 所示, 其中, 图 4(a) 为原始图像. 图 4(b)–图 4(f) 为根据 VGG-16 和 ResNet 的各种可视化支持猫类别: 图 4(b) Guided Backprop [91]: 突出显示所有贡献特征. 图 4(c) 和图 4(f) Grad-CAM [56]: 定位类判别区域, 图 4(d) Guided Grad-CAM [56]: 高分辨率的类判别可视化. Guided Backprop 方法可以突出图像中的细粒度细节, 但不具有类别区分性. CAM 和 Grad-CAM 方法具有类区分性但没有细节信息, Guided Grad-CAM 方法同时具有高分辨率和类别区分性. 文中将新方法应用于图像分类、图像字幕和视觉问答模型. 实验表明, 新方法在图像分类任务中: ① 有助于深入了解模型故障模式. ② 在弱监督定位任务中表现良好. ③ 忠实于底层模型. ④ 可以帮助模型泛化. 对于图像字幕和视觉问答模型, 新方法有助于实现物体定位.

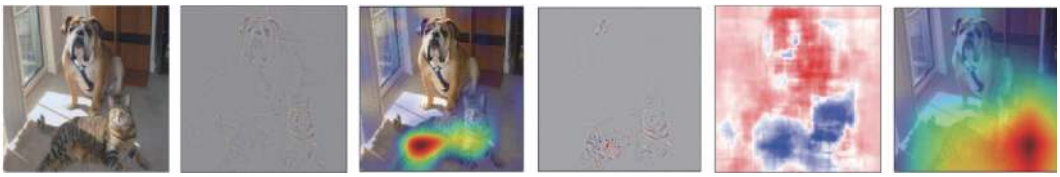
(3) 模型未知方法

如其名字所示, 模型未知方法将网络看作黑盒处理, 不需要关心解释模型的具体形式或参数. 在不知道模型如何工作的情况下研究模型对特定输入样本中特征的关注程度[107]. 鉴于待解释模型的非线性特点, 采用可解释性较好的线性模型拟合非线性模型, 以增加其可解释性, 是一种常见的思路[108,109]. 典型的模型未知方法之一是 LIME (local interpretable model-agnostic explanations)[59]. LIME 在输入样本附近采样, 构建线性解释方法, 通过线性解释方法局部拟合网络模型, 解释当前输入的分类结果. LIME 原理图如图 5 所示, 存在待解释模型 $f: R^d \to R$ 和输入样本为 x, 在其附近采样得到的样本为 $z \in \{0,1\}^{d'}$ (图像中超级像素块的存在与否), 用 $\pi_x(z)$ 衡量 z 到 x 的近似度, $\mathcal{L}(f,g,\pi_x)$ 衡量 g 在由 π_x 定义下逼近 f 时的不忠实程度. $\Omega(g)$ 是解释 $g \in G$ 的复杂性 (相对于可解释性) 的度量. 在可解释模型集合 G 中寻找一个线性模型 g, 使得:

$$\xi(x)=\arg\min_{g\in G}\mathcal{L}(f,g,\pi_x)+\Omega(g) \tag{5}$$

在 x 的附近, 可以用线性模型 g 近似解释模型 f, 并通过 x 附近的采样点解释 x. LIME 对实例进行采样, 通过对采样样本的预测结果解释实例. 图 5 中黑盒模型的复杂决策函数由蓝色/粉色表示. 粗红十字是需要解释的实例.

文中将算法应用于解释文本任务和图像分类任务的不同模型来展示算法的灵活性, 并通过模拟用户实验以评估解释结果在信任相关任务中的有效性, 并就信任解释模型问题展开一系列相关分析.



(a) Original image (b) Guided Backprop ‘Cat’ (c) Grad-CAM ‘Cat’ (d) Guided Grad-CAM ‘Cat’ (e) Occlusion map for ‘Cat’ (f) ResNet Grad-CAM ‘Cat’

(g) Original image (h) Guided Backprop ‘Dog’ (i) Grad-CAM ‘Dog’ (j) Guided Grad-CAM ‘Dog’ (k) Occlusion map for ‘Dog’ (l) ResNet Grad-CAM ‘Dog’

图 4 不同基于梯度和反向传播的方法可视化结果比较

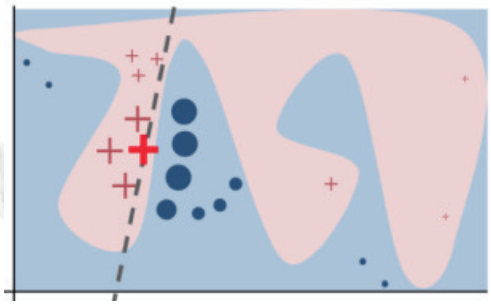


图 5 LIME 原理图[59]

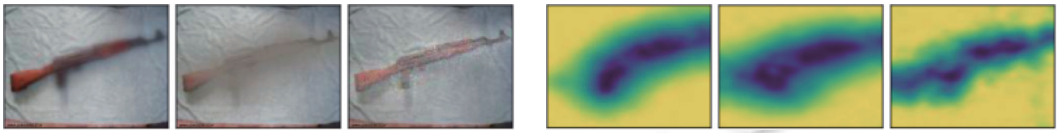
利用博弈论理论中的 Shapley 值解释模型是另一类模型未知方法. Shapley 值最初用在合作博弈中的份额分配均衡问题. 在解释模型中, 可以通过不同输入特征的 Shapley 值, 将模型预测结果与平均基线之间的差异归因于模型输入样本的不同特征. 当特征相互独立时, Shapley 值可以很好地进行解释, 但当独立性假设被违反时, 可能会导致算法解释能力变差. 文献 [92] 推导出因果 Shapley 值来解释特征对预测的总体影响, 同时考虑到它们的因果关系, 并允许通过将总因果效应分离为直接贡献和间接贡献来进一步分析特征相关性. 文中展示了在不牺牲任何所需属性的情况下为一般因果图导出这些因果 Shapley 值. 文中在实际示例中说明, 当只有部分信息可用时的基于因果链图计算因果 Shapley 值的实现方法, 以此证明算法的实用性. 对 Shapley 值的精确评估非常昂贵, 随着输入特征的数量增加, 算法复杂度呈指数级增长. 为解决算法复杂度的问题, 文献 [93] 提出了一种 Shapley 值的近似值计算方法, 对深度神经网络中 Shapley 值进行多项式时间近似, 产生有效的 Shapley 值近似值. 通过与现有的其他方法相比, 文献 [93] 证明新方法产生的 Shapley 值近似值具有明显优势. 文献 [94] 将模型未知方法用于对反事实推理实例进行解释. 算法使用编码器或通过特定类的 k-d 树获得的类原型, 在目标函数中使用类原型来将扰动快速引导到可解释的反事实. 文中定量评估生成的反事实的可解释性, 以说明提出方法在图像和表格数据集上的有效性. 文中将提出方法用于消除由于黑盒模型的数值梯度评估而出现的计算瓶颈.

(4) 基于扰动的方法

基于扰动的方法通过对输入样本部分信息增加扰动并观察输出结果的变化情况, 从而确定输入样本中扰动部分对输出结果的影响. 即对于一个网络 $f(x)$, 输入样本 x_0 的哪些区域对输出值 $f(x_0)$ 具有较大贡献. 在基于扰动的方法中, 一个重要的问题是使用什么样的方法对输入样本进行扰动, 也就是使用输入样本的哪些变体进行研究[95].

可以通过观察 $f(x)$ 的值如何随着 x 的变化而变化, 从而删除输入样本 x_0 的不同区域 R. 为了构建具有通用性的解释模型框架, 文献 [96] 中提出通过用恒定值扰动、噪声扰动和模糊扰动 3 种方式生成有意义的扰动图像的方法. 文献 [96] 中采用 3 类扰动的实验结果对比如图 6 所示. 设 $m:\Lambda\to[0,1]$ 是一个掩码, 将每个像素 $u\in\Lambda$ 与一个标量值 $m(u)$ 相关联, 那么扰动算式定义为:

$$[\Phi(x_0;m)](u)\begin{cases}m(u)x_0(u)+(1-m(u))\mu_0, \text{constant}\\ m(u)x_0(u)+(1-m(u))\eta(u), \text{noise}\\ \int g_{\sigma_0 m(u)}(v-u)x_0(v)\mathrm{d}v, \text{blur}\end{cases} \tag{6}$$



(a) 模糊扰动 (b) 恒定扰动 (c) 噪声扰动 (d) 模糊扰动掩码 (e) 恒定扰动掩码 (f) 噪声扰动掩码

图 6 不同的扰动类型[96]

第 1 类是恒定值扰动, μ_0 是颜色均值, 第 2 类是噪声扰动, $\eta(u)$ 是高斯噪声, 第 3 类是模糊扰动, σ_0 是高斯模糊核 g_σ 的标准差. 定义优化函数:

$$m^*=\mathrm{argmin}_{m\in[0,1]^\Lambda}\lambda\|1-m\|_1+f_c(\Phi(x_0;m)) \tag{7}$$

学习得到最优的掩码 m. 对于目标类别 c, 扰动后的输入样本准确度下降明显, 即 $f_c(\Phi(x_0;m))\ll f_c(x_0)$, 表示掩码遮挡了输入样本中对判别类别 c 最重要的区域. 文献 [96] 在图像分类任务中应用新算法, 通过删除游戏最小掩码以阻止网络识别正确的方式, 对网络可视化结果进行解释.

另一类基于扰动的方法是使用生成模型生成输入样本的扰动. 与上述扰动相比较, 这一类扰动更加自然、不容易被人眼识别. 但 CNN 容易受到此类人眼无法察觉的像素级扰动的对抗样本的影响[97]. 文献 [98] 结合对抗性防御技术, 对文献 [96] 的工作做进一步扩展, 在计算输入样本的扰动时, 删除所有不相关或最相关的像素, 形成一种细粒度的视觉解释图像. 文中对在不同网络中对新方法和其他方法 (BBMP[96], Gradient[54], Guided Backprop[91], Contrastive Excitation Backprop[110], Grad-CAM[56], Occlusion[90]等) 的解释结果进行定性和定量比较, 说明该算法可以提供最细粒度的解释. 文献 [99] 中针对具有像素级扰动的对抗样本对 CNN 的影响进行解释, 研究像素级扰动对图像分类的敏感性和功能, 将扰动分为 3 类: 1) 抑制型扰动; 2) 促进型扰动; 3) 平衡型扰动, 揭示了对抗性扰动的促进-抑制效应 (PSE), 并将像素级扰动的 PSE 与类激活图 (CAM) 的图像级可解释性相关联, 通过使用特定于类别的判别图像区域来解释对抗模式. 文中通过网络分解方法[57]来检查对抗效应, 说明其提供了隐藏单元的概念级可解释性.

(5) 其他方法

为了分层解释神经网络, 文献 [111] 中提出一种神经网络预测的解释方法, 对于给定的输入样本和预测结果, 通过 ACD (agglomerative contextual decomposition) 生成输入特征的分层聚类, 并计算每个聚类对最终预测的贡献. 文中对不同数据集引入 ACD 算法, 用以诊断不正确的预测、识别数据集偏差并提取不同长度的短语. 通过实验证明, ACD 使用户能够识别两个网络中更为准确的一个, 并更能够信任网络的输出结果. 文献 [112] 从信息论的角度进行模型解释, 提出一种将实例特征选择器作为解释模型的方法. 通过优化选择器来选择对输入最有用的特征子集, 选择器可以最大化所选特征和函数响应之间的互信息, 通过给定输入样本的函数响应的条件分布对模型进行解释. 文中对合成数据集和真实数据集分别进行实验, 应用于不同网络, 对运行时间、可解释性等进行定量分析, 从而证实算法的有效性.

(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的共同点在于, 此类方法提供单个预测结果的解释信息, 可视化与模型决策相关度最高的区域或像素. 算法的解释结果与预测结果可以相互验证, 从而增强模型的可信任度. 如表 3 中单一输入的典

型效果示例所示, 对于指定输入样本, 解释方法给出不同输出类别对该样本该兴趣的区域.

根据解释方式的不同, 基于单一输入的解释方法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、基于扰动的方法等. 不同解释方法的特点归纳如表 5 所示: 1) 类激活映射方法具有良好的定位能力, 能够准确定位卷积网络中选定类别对输入样本感兴趣的区域, 且可以区分显示不同区域的贡献度. 但该方法改变卷积神经网络的原始网络结构, 且生成的解释结果不具有细粒度的可视化结果. 这类方法适用于解释分类任务下卷积神经网络. 2) 基于梯度和反向传播的解释方法, 具有细粒度解释网络的能力, 且具有较强的通用性, 适用于各类任务的不同网络. 但存在计算复杂的问题, 且可能存在反向传播过程中梯度消失导致解释失败的问题. 3) 模型未知方法原理简单明了, 易于理解, 且与模型无关, 不限制网络的类型和结构, 适用于对不同类型的网络进行解释. 不足之处在于计算耗时长, 需要人工对生成的解释结果进行进一步的解释. 4) 基于扰动的方法, 通过输入经过扰动的图像, 观测网络的预测结果的变化, 从而决定扰动部分对网络的影响程度. 这类方法原理简单, 直观易于理解, 适合观测分析网络对某一类别物体的感兴趣区域. 但这类方法受网络精准度的影响较大, 在精准度较高的网络中这类方法效果明显, 在精准度较低的网络中, 预测结果的变化量并不明显, 因此解释能力下降.

表 5　基于单个输入的解释方法分析

类别	优点	缺点	适用范围
类激活映射方法	具有良好的定位能力, 可区分显示不同区域的贡献度	需要修改网络的原始结构, 生成的解释结果不具有细粒度的可视化结果	适用于解释分类任务下卷积神经网络
基于梯度和反向传播的方法	具有细粒度解释网络的能力, 且具有较强的通用性, 适用于各类任务的不同网络	计算复杂, 存在传播过程中梯度消失导致解释失败的问题	适用于解释不同任务下的卷积神经网络
模型未知方法	原理简单明了, 易于理解, 适用于对不同类型的网络进行解释	计算耗时长, 需要人工对生成的解释结果进行进一步的解释	适用于解释不同类型的网络结构
基于扰动的方法	原理简单, 直观易于理解, 适合观测分析网络对某一类别物体的感兴趣区域	受网络精准度的影响较大, 只在精准度较高的网络中这类方法效果明显	适用于解释精准度较高的网络

4.2　基于多个输入的解释方法

基于多个输入的解释方法是为一类相似的输入样本的输出结果提供统一的解释, 分解在输入样本集中具有普遍性的特征, 描述每个特征对网络的输出结果做出的贡献, 进而对模型进行解释. 一种直接的获得多个输入的解释的方法是, 对数据集进行分析, 针对单个属性的贡献度进行分析. 文献 [100] 认为, 当前许多机器学习算法存在聪明汉斯 (clever Hans) 现象, 即只是学习到样本和分类结果之间某种无意义的关联, 而不是学习到了真正的智能. 为了描述和验证模型的有效性, 评估模型是否学习到了人们所期望的特征, 文献 [100] 提出一种 SpRAy (spectral relevance analysis) 算法, 对单个属性进行聚类, 并对预测策略进行总结. 算法首先为输入样本和感兴趣的目标类别计算相关图. 进行基于特征值的聚类分析从而识别分析数据中的不同预测策略. 文中通过在图像分类任务中识别“马”这一类别举例分析, 通过实验确定, 模型具有不同的预测策略可以将输入图像分类为“马”.

另一类常见的多个输入的解释方法是将单一样本的解释方法中被解释的输入样本通过一定方式结合起来并寻找规律. MAME (model agnostic multilevel explanations)[60]是一种元方法, 基于已有的模型未知的局部解释方法, 构建多级解释树. MAME 旨在构建局部解释和全局解释间的关系, 并获得中间层级的解释. 树的底层对应于局部解释, 顶层对应于全局解释, 中间层对应于它自动聚类的数据点组的解释. 算法首先采用局部可解释性技术, 如 LIME[59]对数据进行解释, 构成解释树的底层. 并基于解释结果对不同样本进行聚合, 形成树的中间层, 直至全部样本聚合完成, 形成树的顶层. 文章在两种不同场景中应用提出的算法, 分别针对有无专家知识的情况验证算法的有效性. 文中使用两项度量标准 (广义保真度和特征重要性等级相关度) 定量衡量算法, 表明算法在两项指标中优于其他方法. 为了全局解释黑盒模型, 文献 [101] 中使用二叉树作为解释树来阐明隐含在模型中的重要决策规则. 文献 [101] 提出一种通过递归分区进行全局模型解释的方法 GIRP (global interpretation via recursive partitioning). 基于多个样本的局部解释结果, 构建各输入样本不同变量的贡献度方程, 通过最大化输入变量的贡献差异来递归地划分输入变量空间, 以此构建二叉树去近似原始模型做出各类决策的规则. 文中分别应用算法于场景理解、文本分类、重症监护病房死亡率预测等多种不同任务, 证明算法在处理不同数据不同任务时的有效性.

对比输入样本中的单个特征、像素, 利用更容易被人类理解的单元来解释机器学习模型的决策更为直观易懂, 这些单位称为概念[58,78]. 为了通过人类可理解的概念解释神经网络的内部状态, 文献 [58] 提出一种 TCAV (testing with concept activation vectors) 算法, 不研究输入样本本身具有的变量或特征, 而是通过引入概念激活向量 (concept activation vectors, CAVs), 计算用户感兴趣概念的重要程度. 根据这些概念提供对神经网络内部状态的解释. 算法首先收集有/没有目标概念的例子 (例如, 动物身上存在/不存在条纹), 这个概念可以表示为分隔这些正/负例子 (有/没有条纹的动物的图片) 的超平面上的一个分割向量. 向量随输入样本的变化程度可以衡量概念的稳定程度. 文章将算法运用于图像分类任务, 分析网络运行基本原理. 文中应用算法解释预测糖尿病视网膜病变的模型, 算法能够在模型与领域专家知识出现分歧时提供建议. 除了由人工选择之外, 这些概念也可以通过对输入片段进行聚类来自动发现[102]. 为了识别更高层次的人类可理解的适用于整个数据集的概念, 文献 [103] 提出一种基于概念的解释方法, 通过定义包含样本一类或多类基本特征、可以为人类理解的概念, 对所有数据进行解释. 算法通过分析不同输入样本, 拆分出样本中的局部特征, 并对不同样本中的局部特征进行聚类形成概念, 计算每个概念的重要程度, 从而对输入的所有样本进行解释. 文中将算法应用于物体识别模型, 通过定量人类测评和评估验证算法有效性. 实验结果表明, 该算法能够发现有对人类有意义且对网络预测结果很重要的概念.

基于多个输入的解释方法可以提供对模型行为更为一般的理解, 如识别不同的预测策略等. 这类解释对于模型学习到的策略具有更为全面的认识. 此类算法从全局出发解释分析网络, 可以理解网络运行的整体逻辑. 表 3 中多个输入的典型效果示例是其中一种方法, 通过生成具有重要语义超像素的图像解释网络分类结果, 如对于“卧室”类别 (图中第 1 列), “床”和“地板”的超像素最为重要 (图中第 4 列). 此类方法中, 不同算法的研究方式具有较大的差异性, 算法普遍较为复杂, 计算量大. 模型的不透明性和复杂性等问题也使得基于多个输入的解释方法更为困难.

4.3 可解释算法评估指标

可解释算法的评估指标用于度量各种解释方法的解释效果, 目前很难找到适用于大多数解释算法的评估标准. 文献 [24] 认为可以从以下 4 方面对可解释算法进行评估. 第 1 个方面是代理模型的相似性. 一部分解释方法采用可解释的代理模型拟合原始模型, 从而对模型进行解释. 因此可以根据代理模型与所解释的原始模型间的近似程度来直接评估代理模型, 进而评价其解释能力. 第 2 个方面是替代任务的完整性. 一些解释方法不直接解释模型的决策, 而是解释其他可以评估的属性. 例如, 可以根据模型敏感性的测量结果来评估用于揭示模型敏感性的显著性解释结果. 第 3 个方面是对有偏差的模型的检测. 可以测试对特定现象 (例如输入中存在特定模式) 的敏感性解释是否能够揭示模型偏差. 第 4 个方面是人工评价的合理性. 人类可以评估解释的合理性, 即解释与人类期望的匹配程度. 人工评价是最简单直接的评估可解释算法的方法, 但需要警惕仅使用人类对可解释算法进行评估可能引入评估偏见. 依赖人工评估可能会导致用于解决信任问题的可解释算法变得不可信任. 文献 [113] 中提出的构建可解释系统时存在的道德困境: 人为介入解释达到更好地说服用户的目的是否是道德的? 生成揭示模型的解释结果和生成符合道德标准的解释结果, 二者间该如何平衡? 不同类别的可解释算法的评估方法高度重合, 本文对常见的评估方法进行梳理.

(1) 基于视觉的评估方法

基于视觉的评估方法是可视化解释结果并对其进行评估, 通过人判断可视化结果的准确性, 这类方法适用于基于理想样本的解释方法. 基于理想样本的解释方法通过最大化网络单元的激活值生成理想样本. 对于视觉任务而言, 该类解释方法的输出结果为理想样本图像, 需要考察生成的理想样本图像与自然图像的相似程度.

如图 7 所示, 通过对解释方法生成的理想样本与自然图像的相似程度对比, 以此评估基于理想样本的解释方法算法的生成效果. 图中第 1 行图像为简单 AM 方法[71]生成的理想样本; 第 2 行图像为 AM 方法[71]加 L2 正则约束生成的理想样本; 第 3 行图像为 AM 方法[71]加样本均值约束生成的理想样本; 第 4 行图像为 DGN-AM 方法[76]生成的理想样本. 随着解释方法对先验知识的引入增多, 其生成的理想样本与自然图像的相似程度愈加提升, 其理想样本的自身的清晰度与可辨别性也随之提升, 其中 DGN-AM 方法[76]生成的理想样本最接近真实图像. DGN-AM[76]在训练过程中通过引入 GAN 网络, 通过学习训练样本的先验分布, 使得在生成理想样本的过程中, 可以直接使用来自训练样本的先验知识, 从而使其的理想样本更为接近训练样本, 而简单 AM 方法[71]则是在完全随机初

始化的输入上生成理想样本, 其输出更为抽象, 与自然图像相差更远. 可以看出, 好的解释算法生成的理想图像应该具有自然语义.

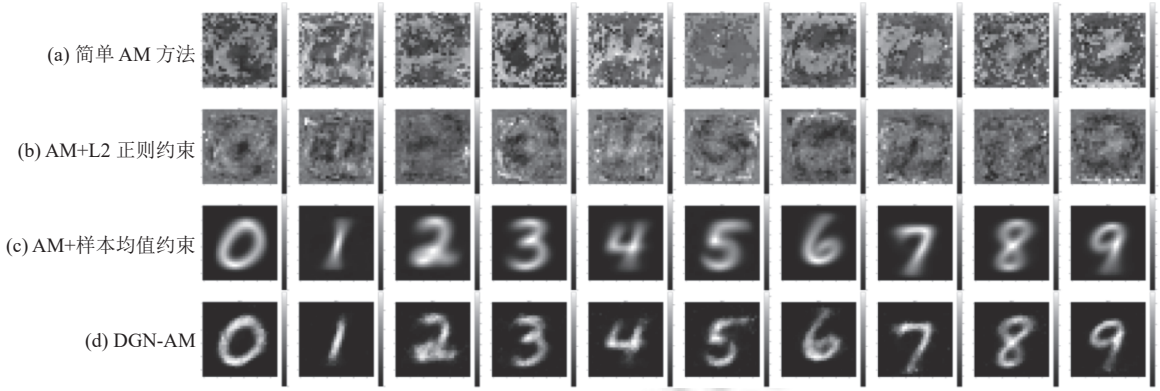


图 7　不同基于理想样本的解释方法在 MNIST 数据集生成的可视化结果[105]

此外, 基于视觉的评估方法常用对基于单一输入的解释方法的评估. 例如, 解释方法通过对输入与网络结构 (VBP[54], GBP[91], 反卷积[90]) 和输入与输出 (基于扰动的方法[90,96,114]和类激活映射[55,56,115]) 的关系, 生成相应的显著图或热力图. 为便于理解以及符合人类的视觉直觉, 显著图和热力图可以通过考察图像的连贯性和可分辨性来评估解释方法的效果好坏. 此外, 该类方法的生成结果为像素级别的显著图或热力图, 并且与输入图像之间有着紧密的局部或全局的联系.

(2) 敏感度分数

基于视觉的评估方法是一种定性评估方法, 通常需要结合定量的评估方法, 敏感度分数即是一种定量评估方法. 为考察显著图的显著程度, 文献 [96] 提出观察加入干扰对输出分数的影响, 旨在使用最小的区域干扰最大的分数变化. 解释方法的显著图, 通常会突出在人类视觉层次上对目标对象分类的关键部位, 但敏感度分数表示, 解释方法不仅可以找到图像中对算法结果决策重大的部分区域, 而且可以找到对于被解释网络而言最容易区分的部分. 该类方法简单直观但无法避免生成显著区域过程中产生的噪声影响, 导致评价结果产生误差. 如表 6 所示, 平均下降指标是指通过对得到的显著区域进行遮掩, 观察其模型对目标类的置信度下降百分比, 数值越低解释方法效果越好; 平均上升指标是通过对得到的显著区域进行正向增强, 观察其模型对目标类的置信度上升百分比, 数值越高解释方法效果越好. 其中基于扰动的 Mask 方法[96]效果最低, 基于类激活映射的 ScoreCAM 方法[115]效果最好. Grad-CAM[56]和 Grad-CAM++[116]相似, 都是基于梯度的类激活图生成方法, 故由于梯度的不稳定性, 其生成的显著图同样具有不稳定性质, 在生成过程中容易引入噪声影响, 导致评价结果较低. 而 ScoreCAM[115]通过将最高层每个通道的特征图作为一个掩码, 与原始输入图像叠加后再由卷积神经网络对分类概率进行预测, 其中使用关于特定目标类别的分类概率作为权重进行通道加权求和, 以此克服了基于梯度方法带来的不稳定性, 生成了噪声含量更少的类激活图, 故其敏感度分数指标较高.

(3) 交并比 (intersection over union, *IoU*)

仿照目标检测任务中的 *IoU*, 对生成热力图按照设定的阈值生成目标边框, 并于输入图像的真实边框进行比较, 得出交并比 *IoU*, 若 *IoU* 大于设定阈值, 表示生成的热力图正确的包含了输入图像的目标, 对其可解释性有正增强作用. 交并比适用于基于真实样例的解释方法和基于单一输入的解释方法. 表 7 表示不同解释方法在交并比度量下的解释程度. 表中用简单的两种阈值方法来拟合生成热力图的目标边框. 值域值方法通过将热力图归一化, 再用 $a \in [0:0.05:0.95]$ 对其值进行阈值, 以寻求最佳的阈值大小; 均值阈值则利用热力图的平均强度的 a 倍乘进行阈值, 其中 $a \in [0:0.5:10]$, 寻求最佳的阈值大小. 两种方法都将求得的目标边框与真实目标边框求交并比, 以阈值 0.5 计算定位错误率. 在值域值指标中, 基于扰动的 Mask 方法[96]错误率最低; 在均值阈值指标中, 基于梯度的方法 Grad[54]错误率最低. 基于扰动的 Mask[96]在通过优化的方法生成有效遮掩区域的同时, 还通过 L1 范数约束生成的掩码, 使得生成的扰动区域尽可能小, 从而在值域值指标中有较好表现. 基于梯度的 Grad 方法[54]直接通过输

入图像的梯度构成显著图, 显著图中值较高的区域即为梯度最大, 与输出相关性最大的区域, 直接构建了输入与输出的关联图像, 因而在均值阈值指标上表象较好. 但该指标更倾向于输出具有局部性的解释方法, 对更为细粒度的方法如定位相关性像素之类的解释方法并不友好, 故该类评估指标具有一定的局限性.

表 6 解释方法在敏感度分数上的评估结果 (%)[115]

解释方法	平均下降	平均上升
Mask[96]	63.5	5.29
Grad-CAM[56]	47.8	19.6
Grad-CAM++[116]	45.5	18.9
ScoreCAM[115]	31.5	30.6

表 7 解释方法在交并比 *IoU* 上的评估结果[96]

解释方法	值阈值-a	错误率 (%)	均值阈值-a	错误率 (%)
Grad[54]	0.25	46.0	5.0	41.7
Guid[91]	0.05	50.2	4.5	42.0
LRP[69]	—	—	1.0	57.8
CAM[55]	—	—	1.0	48.1
Grad-CAM[56]	0.30	48.1	1.0	47.5
Occlusion[90]	0.30	51.2	1.0	48.6
Mask[96]	0.10	44.0	0.5	43.2

(4) 指向游戏 (pointing game)

某些细粒度的解释方法会追求定位与分类最相关的元素而非类别目标整体, 这类方法得到的输出图像与输入的交并比 *IoU* 总体偏小, 故交并比 *IoU* 方法无法反映该类方法的真是解释效果. 对于特定种类, 判断热力图中最大激活值是否落入输入图像目标的真实边框中, 计算该类别的落入命中率, 并计算不同类别的平均命中率, 以此作为解释方法的评估标准. 如表 8 所示, 分别在 VOC07 测试集和 COCO 验证集上对解释算法进行指向游戏指标上的评估. 其中居中作为评估的基准线, 指最大激活值固定为图片的正中心位置. 基于类激活的解释方法取得了最好的结果. 通过指向游戏指标可发现, 基于梯度的解释方法和反卷积解释方法的性能对被解释网络的结构敏感度较高, 这两类解释方法依靠网络本身的数据流生成最大激活值, 而 CAM 类方法则更多只关注较高层特征图, 对网络结构的依赖性并不太高, 故在具有跃层连接的 ResNet50 网络, CAM 类方法能取得相较于评估基准更好的成绩. 而基于梯度和反向传播的解释方法多数需要对网络结构进行一定的解析, 对于复杂的网络, 解析的方法还需要进一步的迭代以适应更为复杂的网络结构. 指向游戏评估指标只关注于最大激活值点, 可适用于各种粗细粒度的解释方法, 但因其只关注热力图中的极小部分, 更容易被热力图生成过程中产生的噪声所影响, 进而产生评估误差.

表 8 解释方法在指向游戏上的评估结果[110]

解释方法	VOC07测试集(全集/困难)		COCO验证集(全集/困难)	
	GoogLeNet网络	ResNet50网络	GoogLeNet网络	ResNet50网络
居中	69.5/42.6	69.5/42.6	27.7/19.4	27.7/19.4
Grad[54]	79.3/61.4	75.8/50.9	42.6/36.4	30.4/24.9
Deconv[88]	74.3/49.4	73.0/53.0	35.7/27.9	38.2/31.2
LRP[69]	72.8/50.2	—	40.2/32.7	—
CAM[55]	80.8/61.9	90.6/81.8	41.6/35.0	58.4/53.5

5 可解释神经网络的研究方向、应用和面临的挑战

本节对解释算法的研究方向和应用任务等与可解释性研究密切相关的内容进行阐述, 并就神经网络解释模型当前面临的问题进行简述, 针对这些挑战给出可能的解决方向.

5.1 可解释神经网络研究方向

对神经网络进行解释的目的是更好地了解网络, 学习网络, 总结网络运行规律, 了解神经网络错误决策的原因. 同时可以为优化神经网络上提供建议, 从而为提升网络性能, 为设计更优秀的神经网络提供依据和方向.

(1) 探索神经网络原理

不同类型的解释方法对从不同的角度模型进行解释, 因而解释方法对网络的理解也不相同. 基于理想样本的

解释旨在发现网格单元的学习信息, 如深度神经网络的神经元单元. 文献 [54,76] 通过构建具有抽象概念的原型解释模型所学到的知识. 例如, 通过生成“汽车”原型图像来解释模型对“汽车”类别的理解. 构建这样的原型已被证明是研究深度神经网络单元的有效工具[61]. 基于真实样例的解释通过识别有代表性的样本来解释模型, 有助于更好地理解数据集以及其对模型的影响. 此外, 具有代表性的样例可能有助于识别数据中的偏差, 使模型对数据集的变化更加稳健. 例如, 文献 [57] 利用样例表示网络隐藏单元的习得的语义信息, 进而对具有语义的单元赋予对象、纹理、颜色等标签. 基于单一输入的解释提供有关单个预测结果的解释信息, 例如通过显著图可视化哪些像素与模型最相关以达到其决策[55,56]. 文献 [90] 从输入样本中提取特征, 进而深入了解卷积网络各中间层的功能. 这类算法的解释结果有助于验证预测结果, 从而建立对模型的信任感. 基于多个输入的解释提供对模型行为更为一般的理解, 如识别不同的预测策略等. 这类解释对于模型学习到的策略具有更为全面的认识. 文献 [100] 通过对热力图进行聚类来生成集群模型的元解释, 并生成不同集群模型学习到的预测策略. 例如, 使用分类器对“马”进行分类时识别出 4 个聚类: 马和骑手; 纵向图像中的标签; 木栅栏和骑马的其他元素; 横向图像中的标签[100].

(2) 优化神经网络结构

对网络进行解释的重要目标是对模型进行改进, 从而提升网络性能. 解释结果可以用于改进模型, 例如, 文献 [117] 结合解释算法与正则化对模型进行改进. 此外, 由于解释结果提供了有关模型的内部信息, 因此解释算法可以用于模型压缩和修剪. 文献 [90] 通过消融实验以发现不同模型的不同层对预测结果的性能贡献, 并通过可视化对模型进行诊断, 进而构建更优的模型架构. 文献 [56] 在图像分类任务下, 对模型错误预测情况提供解释, 进而深入了解模型的故障原因. 同时, 文献 [56] 通过识别数据集偏差为模型泛化提供帮助, 使模型适用于更多任务. 文献 [59] 通过解释算法在几个方面对模型进行分析: 模型是否值得信任、不同模型的选择、改进不可信模型以及给出不信任分类器的原因, 从而给出改进模型的建议.

5.2 可解释神经网络实际应用

可解释神经网络广泛应用于不同领域, 例如医疗领域、刑事司法系统和自动驾驶等. 神经网络模型应用于医疗领域的场景不断增加. 然而, 医疗临床决策要求神经网络模型不仅性能良好, 同时透明、可解释. 文献 [118] 提供一种可解释的深度学习方法, 用于断层扫描仪扫描识别感染 COVID-19 病毒的场景. 该方法可以帮助医生清楚地理解扫描过程. 这类方法可以快速判断患者是否感染 COVID-19 病毒, 同时, 这种方法可以扩展到包括更多应用场景, 例如判断患者是轻度患者还是重度患者等, 文献 [119] 提出一种具有可解释性的深度神经网络, 并应用于断层扫描仪扫描结果, 实现决策原理的可视化.

可解释神经网络的另一个应用场景是刑事司法系统. 在一些国家 (例如美国), 神经网络模型被用于预测可能犯罪地点和潜在犯罪人员等. 虽然该模型中使用的数据不包括种族信息, 但数据的其他信息可能与种族信息相关, 因此可能会导致预测中的种族偏见. 文献 [120] 中证明的可解释算法可以检测模型中偏见, 根据解释结果调整模型可以保证司法系统的公正性.

自动驾驶是一种自动化系统, 可能在完全未知的环境中使用. 因此, 相比于“黑盒”模型相比, 人类对具有透明度和解释性的系统更具有接受度. 将解释算法应用于自动驾驶中的视觉理解场景[121]和态势感知场景[122]可以提升自动驾驶中模型的可解释程度. 道路上的异常情况可能对自动驾驶汽车造成严重影响. 针对此问题, 文献 [122] 提出一种自我进化的方法, 模型可以主动从当前情况中学习知识并提供解释.

5.3 可解释神经网络面临的挑战

可解释神经网络算法近年来显著发展, 但可解释算法仍然面临一些亟需解决的问题, 本节对其进行归纳总结, 并给出解决思路.

(1) 模型性能和可解释性间存在矛盾

神经网络模型的复杂度随性能的提升而增加, 而模型自身的可解释性通常随着复杂度的提升而减弱. 在这条通向更优性能的道路上, 当性能与复杂性齐头并进时, 可解释能力的下降似乎是不可避免的[123]. 从准确性和可解释性的角度出发, 当前机器学习模型主要分为两类: 一类模型深奥但无法解释. 这类算法普遍具有较高的准确性, 但算法结构复杂, 学习到的高维特征通常是人类无法理解的. 另一类模型可解释但结构简单, 缺乏对高维特征的提

取能力, 准确性较低, 模型性能一般. 一味追求可解释性可能会牺牲准确性, 这在错误决策具有严重后果的任务中是不可取的. 因此, 如何权衡准确性和可解释性, 构建即可以满足准确性要求又能够被解释的网络成为一个可以深入研究的问题.

针对这一问题, 可以考虑平衡模型可解释性和模型性能间的关系. 通过增加可解释算法的复杂程度, 实现对算法复杂度高的模型进行解释, 并确保解释结果能够代表所研究的模型, 不会过度简化其基本特征. 根据当前解释算法的特点, 对不同类型的解释算法进行有导向型地融合, 使得新型解释算法同时具有不同类型解释算法的优点.

(2) 缺乏统一评估标准

当前研究中, 不同类型的解释方法从不同角度理解可解释的概念和解释模型. 可解释算法对于模型的解释程度没有一个统一的度量标准. 上文中提及到的当前对于解释能力的评估, 通常只适用于一个或一类解释算法. 缺乏一个或一组指标能够对不同解释算法对模型的解释程度进行统一比较, 因此很难对不同类型可解释算法的解释性进行定性的分析和比较. 目前一些研究从人类对解释结果满意程度、解释结果对人类理解模型的启发程度、解释可信度等方面对算法进行定性分析, 但是尚未出现可量化的通用评估指标或工具来对解释算法进行系统性测量.

设计出一套完备的解释算法评估系统是未来可解释研究的方向之一. 该评估系统应适用于不同解释算法之间的比较, 能够在不同的应用环境、模型和目标下对算法进行定量对比. 评估方法满足以下要求. 解释性: 用可直观理解的概念对网络进行解释. 区分度: 解释结果有区分度, 针对不同网络/类别/网络单元, 解释方法可以给出符合当前目标特点的、不同于其他目标的解释结果. 独立性: 解释算法自身具有完备的运行机制, 不依附于特定模型或任务, 也不影响模型的正常运行. 稳定性: 算法具有稳定性, 针对不同输入的解释算法的内核原理始终保持不变, 并且不随输入顺序或输入时间等无关因素的影响而变化. 为了实现统一的标准化度量, 可根据评估指标要求, 构建特定满足条件的数据集, 不同解释算法可在相同数据集下的进行比较, 从而实现统一评估.

(3) 解释算法正确性无法保证

一个成熟的解释算法应具有可信任性, 给出的解释结果可以证明自身的正确性, 从而使得人类可以信任算法给出的解释结果. 而当前的解释算法普遍并不具备这一能力. 目前大多数解释算法通常致力于给出符合人类对网络认知的解释结果, 但并不对解释结果的正确性进行分析. 正如人类无法知道网络模型给出的预测结果是否正确一样, 人类也很难知道解释算法的解释结果是否正确. 一方面, 解释算法可能会提供具有误导性或错误的特征, 造成人类对网络的错误认知, 因此对网络的信任程度变得更低, 这样就背离了解释算法的初衷. 另一方面, 为达到较好的解释效果, 解释算法可能会向原模型中添加一些额外的限制, 这可能会降低模型本身的精准性和鲁棒性, 影响模型性能. 这样会造成为了解释而解释的情况, 失去了解释模型的意义.

针对这一问题, 可以考虑充分利用网络预测结果和解释结果之间的独立性和相关性. 预测结果只来源于网络模型, 与解释算法无关. 解释结果应只来源于解释算法, 二者的产生是相互独立的 (这与第 5.2 节中提到的解释算法的独立性相同). 网络运行的原理是保持不变的, 因此两个结果产生的原理一致, 二者应具有相关性. 如果两个结果可以保持相互映证, 则可以一定程度上证明解释算法的正确性.

6 结束语

本文从模型可解释性的定义和研究必要性、模型可解释代表性研究和相关分类算法、新型神经网络可解释算法的分类方法及卷积神经网络的可解释方法梳理、可解释算法的评估、可解释神经网络的研究方向与应用以及当前面临的挑战等多个方面, 针对神经网络可解释性这一问题进行详细的讨论. 当前模型可解释这一课题的研究仍然处于较为初级的阶段, 人类了解神经网络模型的意愿仍然高涨, 期待在未来的研究中可以实现更为智能、易懂、透明的可解释算法.

References:

[1] Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S. Benchmarking and survey of explanation methods for black box models. arXiv:2102.13076, 2021.

[2] Kong XW, Tang XZ, Wang ZM. A survey of explainable artificial intelligence decision. Systems Engineering—Theory & Practice, 2021, 41(2): 524–536 (in Chinese with English abstract). [doi: 10.12011/SETP2020-1536]

[3] Goyal Y, Wu ZY, Ernst J, Batra D, Parikh D, Lee S. Counterfactual visual explanations. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 2376–2384.

[4] Wang YL, Su H, Zhang B, Hu XL. Interpret neural networks by identifying critical data routing paths. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8906–8914. [doi: 10.1109/CVPR.2018.00928]

[5] Frank Pasquale. The black box society: The secret algorithms that control money and information. Business Ethics Quarterly, 2016, 26(4): 568–571. [doi: 10.1017/beq.2016.50]

[6] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1(5): 206–215. [doi: 10.1038/s42256-019-0048-x]

[7] Su JM, Liu HF, Xiang FT, Wu JZ, Yuan XS. Survey of interpretation methods for deep neural networks. Computer Engineering, 2020, 46(9): 1–15 (in Chinese with English abstract). [doi: 10.19678/j.issn.1000-3428.0057951]

[8] T, Luigs HG, Mahlein AK, Kersting K. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2020, 2(8): 476–486. [doi: 10.1038/s42256-020-0212-3]

[9] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Medicine, 2018, 15(11): e1002683. [doi: 10.1371/journal.pmed.1002683]

[10] Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, Mcconnell MV, Percha B, Snyder TM, Dudley JT. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digital Medicine, 2019, 2(1): 31. [doi: 10.1038/s41746-019-0105-1]

[11] Hamamoto R, Suvarna K, Yamada M, Kobayashi K, Shinkai N, Miyake M, Takahashi M, Jinnai S, Shimoyama R, Sakai A, Takasawa K, Bolatkan A, Shozu K, Dozen A, Machino H, Takahashi S, Asada K, Komatsu M, Sese J, Kaneko S. Application of artificial intelligence technology in oncology: Towards the establishment of precision medicine. Cancers, 2020, 12(12): 3532. [doi: 10.3390/cancers12123532]

[12] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2019, 267: 1–38. [doi: 10.1016/j.artint.2018.07.007]

[13] Doshi-Velez F, Kim B. A roadmap for a rigorous science of interpretability. arXiv:1702.08608, 2017.

[14] Comandè G. Regulating algorithms' regulation? First ethico-legal principles, problems, and opportunities of algorithms. In: Cerquitelli T, Quercia D, Pasquale F, eds. Transparent Data Mining for Big and Small Data. Cham: Springer, 2017. 169–206. [doi: 10.1007/978-3-319-54024-5_8]

[15] Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int'l Data Privacy Law, 2017, 7(2): 76–99. [doi: 10.1093/idpl/ipx005]

[16] Phillips PJ, Hahn C, Fontana P, Yates A, Greene KK, Broniatowski DA, Przybocki MA. Four principles of explainable artificial intelligence. Technical Report NISTIR 8312, National Institute of Standards and Technology, 2021. 1–43. [doi: 10.6028/NIST.IR.8312]

[17] Pope PE, Kolouri S, Rostami M, Martin CE, Hoffmann H. Explainability methods for graph convolutional neural networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10772–10781. [doi: 10.1109/CVPR.2019.01103]

[18] Hofman JM, Sharma A, Watts DJ. Prediction and explanation in social systems. Science, 2017, 355(6324): 486–488. [doi: 10.1126/science.aal3856]

[19] Weller A. Challenges for transparency. arXiv:1708.01870, 2019.

[20] Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 2018, 16(3): 31–57. [doi: 10.1145/3236386.3241340]

[21] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608, 2017.

[22] Zhang Y, Tiňo P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Trans. on Emerging Topics in Computational Intelligence, 2021, 5(5): 726–742. [doi: 10.1109/TETCI.2021.3100641]

[23] Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 2018, 73: 1–15. [doi: 10.1016/j.dsp.2017.10.011]

[24] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In: Proc. of the 5th Int'l Conf. on Data Science and Advanced Analytics. Turin: IEEE, 2018. 80–89. [doi: 10.1109/DSAA.2018.00018]

[25] Andrews R, Diederich J, Tickle AB. Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based Systems, 1995, 8(6): 373–389. [doi: 10.1016/0950-7051(96)81920-4]

[26] Freitas AA. Comprehensible classification models: A position paper. ACM SIGKDD Explorations Newsletter, 2013, 15(1): 1–10. [doi: 10.1145/2594473.2594475]

[27] Johansson U, König R, Niklasson L. The truth is in there-rule extraction from opaque models using genetic programming. In: Proc. of the 17th Int'l Florida Artificial Intelligence Research Society Conf. Miami Beach: AAAI Press, 2004. 658–663.

[28] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82–115. [doi: 10.1016/j.inffus.2019.12.012]

[29] Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State-of-the-art in artificial neural network applications: A survey. Heliyon, 2018, 4(11): e00938. [doi: 10.1016/j.heliyon.2018.e00938]

[30] Bao XG, Zhou CL, Xiao KJ, Qin B. Survey on visual question answering. Ruan Jian Xue Bao/Journal of Software, 2021, 32(8): 2522–2544 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6215.htm [doi: 10.13328/j.cnki.jos.006215]

[31] Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 2017, 5(2): 153–163. [doi: 10.1089/big.2016.0047]

[32] Ruan L, Wen SS, Niu YM, Li SN, Xue YZ, Ruan T, Xiao LM. Deep neural network visualization based on interpretable basis decomposition and knowledge graph. Chinese Journal of Computers, 2021, 44(9): 1786–1805 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2021.01786]

[33] Zhang ZZ, Xie YP, Xing FY, McGough M, Yang L. MDNet: A semantically and visually interpretable medical image diagnosis network. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3549–3557. [doi: 10.1109/CVPR.2017.378]

[34] Zhu YH, Ma JB, Yuan CG, Zhu XF. Interpretable learning based dynamic graph convolutional networks for Alzheimer's disease analysis. Information Fusion, 2022, 77: 53–61. [doi: 10.1016/j.inffus.2021.07.013]

[35] Kim J, Canny J. Interpretable learning for self-driving cars by visualizing causal attention. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2961–2969. [doi: 10.1109/ICCV.2017.320]

[36] You J, Leskovec J, He KM, Xie SN. Graph structure of neural networks. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 10881–10891.

[37] Wang WZ, Rao Y, Wu LW, Li X. Progress of judicial judgment prediction based on artificial intelligence. Journal of Chinese Information Processing, 2021, 35(9): 1–14 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2021.09.001]

[38] Lo Piano S. Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward. Humanities and Social Sciences Communications, 2020, 7(1): 9. [doi: 10.1057/s41599-020-0501-9]

[39] Ashoori M, Weisz JD. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. arXiv:1912.02675, 2019.

[40] Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. Electronic Markets, 2021, 31(2): 447–464. [doi: 10.1007/s12525-020-00441-4]

[41] Brundage M, Avin S, Wang J, *et al*. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv:2004.07213, 2020.

[42] Rudin C, Chen CF, Chen Z, Huang HY, Semenova L, Zhong CD. Interpretable machine learning: Fundamental principles and 10 grand challenges. arXiv:2103.11251, 2021.

[43] Chen KR, Meng XF. Interpretation and understanding in machine learning. Journal of Computer Research and Development, 2020, 57(9): 1971–1986 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2020.20190456]

[44] Lakkaraju H, Kamar E, Caruana R, Horvitz E. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. Proc. of the AAAI Conf. on Artificial Intelligence, 2017, 31(1): 2124–2132.

[45] Kroll JA, Huey J, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H. Accountable algorithms. University of Pennsylvania Law Review, 2017, 165: 633–705.

[46] Danks D, London AJ. Regulating autonomous systems: Beyond standards. IEEE Intelligent Systems, 2017, 32(1): 88–91. [doi: 10.1109/MIS.2017.1]

[47] Kingston JKC. Artificial intelligence and legal liability. In: Proc. of the 2016 Int'l Conf. on Innovative Techniques and Applications of Artificial Intelligence. Springer, 2016. 269–279. [doi: 10.1007/978-3-319-47175-4_20]

[48] Zhou ZJ, Cao Y, Hu CH, Tang SW, Zhang CC, Wang J. The interpretability of rule-based modeling approach and its development. Acta Automatica Sinica, 2021, 47(6): 1201–1216 (in Chinese with English abstract). [doi: 10.16383/j.aas.c200402]

[49] Minematsu T, Shimada A, Taniguchi RI. Analytics of deep neural network in change detection. In: Proc. of the 14th IEEE Int'l Conf. on

Advanced Video and Signal Based Surveillance. Lecce: IEEE, 2017. 1–6. [doi: 10.1109/AVSS.2017.8078550]

[50] Minematsu T, Shimada A, Uchiyama H, Taniguchi RI. Analytics of deep neural network-based background subtraction. Journal of Imaging, 2018, 4(6): 78. [doi: 10.3390/jimaging4060078]

[51] 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. Harvard Data Science Review, 2019, 1(2). 1–9. [doi: 10.1162/99608f92.5a8a3a3d]

[52] Laugel T, Lesot MJ, Marsala C, Renard X, Detyniecki M. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 2801–2807.

[53] Lakkaraju H, Bastani O. "How do I fool you?" Manipulating user trust via misleading black box explanations. In: Proc. of the 2020 AAAI/ACM Conf. on AI, Ethics, and Society. New York: ACM, 2020. 79–85. [doi: 10.1145/3375627.3375833]

[54] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff, 2014. 1–8.

[55] Zhou BL, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2921–2929. [doi: 10.1109/CVPR.2016.319]

[56] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 618–626. [doi: 10.1109/ICCV.2017.74]

[57] Bau D, Zhou BL, Khosla A, Oliva A, Torralba A. Network dissection: Quantifying interpretability of deep visual representations. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3319–3327. [doi: 10.1109/CVPR.2017.354]

[58] Kim B, Wattenberg M, Gilmer J, Cai CJ, Wexler J, Viégas FB, Sayres R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 2673–2682.

[59] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 1135–1144. [doi: 10.1145/2939672.2939778]

[60] Ramamurthy KN, Vinzamuri B, Zhang YF, Dhurandhar A. Model agnostic multilevel explanations. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 501.

[61] Samek W, Müller KR. Towards explainable artificial intelligence. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, eds. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer, 2019. 5–22. [doi: 10.1007/978-3-030-28954-6_1]

[62] Yu FX, Qin ZW, Liu CC, Zhao L, Wang YZ, Chen X. Interpreting and evaluating neural network robustness. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 4199–4205.

[63] Cheng KY, Meng CY, Wang WS, Shi WX, Zhan YZ. Research advances in disentangled representation learning. Journal of Computer Applications, 2021, 41(12): 3409–3418 (in Chinese with English abstract). [doi: 10.11772/j.issn.1001-9081.2021060895]

[64] Qin ZW, Yu FX, Liu CC, Chen X. How convolutional neural networks see the world—A survey of convolutional neural network visualization methods. Mathematical Foundations of Computing, 2018, 1(2): 149–180. [doi: 10.3934/mfc.2018008]

[65] Zhang QS, Zhu SC. Visual interpretability for deep learning: A survey. Frontiers of Information Technology & Electronic Engineering, 2018, 19(1): 27–39. [doi: 10.1631/FITEE.1700808]

[66] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Computing Surveys, 2019, 51(5): 93. [doi: 10.1145/3236009]

[67] Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM. Explainable artificial intelligence: An analytical review. WIREs Data Mining and Knowledge Discovery, 2021, 11(5): e1424. [doi: 10.1002/widm.1424]

[68] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy, 2021, 23(1): 18. [doi: 10.3390/E23010018]

[69] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, 2015, 10(7): e0130140. [doi: 10.1371/journal.pone.0130140]

[70] Zhang QS, Wu YN, Zhu SC. Interpretable convolutional neural networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8827–8836. [doi: 10.1109/CVPR.2018.00920]

[71] Erhan D, Bengio Y, Courville A, Vincent P. Visualizing higher-layer features of a deep network. Technical Report 1341, Montreal: University of Montreal, 2009. 1–13.

[72] Olah C, Mordvintsev A, Schubert L. Feature visualization. Distill, 2017, 2(11): 1. [doi: 10.23915/distill.00007]

[73] Zhang QS, Wang WG, Zhu SC. Examining CNN representations with respect to dataset bias. Proc. of the 2018 AAAI Conf. on Artificial Intelligence, 2018, 32(1): 4464–4473. [doi: 10.1609/aaai.v32i1.11833]

[74] Wang F, Liu HJ, Cheng J. Visualizing deep neural network by alternately image blurring and deblurring. Neural Networks, 2018, 97: 162–172. [doi: 10.1016/j.neunet.2017.09.007]

[75] Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. arXiv:1506.06579, 2015.

[76] Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 3395–3403.

[77] Fong R, Vedaldi A. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8730–8738. [doi: 10.1109/CVPR.2018.00910]

[78] Zhou BL, Sun YY, Bau D, Torralba A. Interpretable basis decomposition for visual explanation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 122–138. [doi: 10.1007/978-3-030-01237-3_8]

[79] Pinheiro PO, Collobert R. From image-level to pixel-level labeling with convolutional networks. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1713–1721. [doi: 10.1109/CVPR.2015.7298780]

[80] Cheng X, Rao ZF, Chen YL, Zhang QS. Explaining knowledge distillation by quantifying the knowledge. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12922–12932. [doi: 10.1109/CVPR42600.2020.01294]

[81] Nguyen A, Yosinski J, Clune J. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. arXiv:1602.03616, 2016.

[82] Zhang QS, Wang X, Cao RM, Wu YN, Shi F, Zhu SC. Extraction of an explanatory graph to interpret a CNN. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(11): 3863–3877. [doi: 10.1109/TPAMI.2020.2992207]

[83] Zhang QS, Cao RM, Shi F, Wu YN, Zhu SC. Interpreting CNN knowledge via an explanatory graph. Proc. of the 2018 AAAI Conf. on Artificial Intelligence, 2018, 32(1): 4454–4463. [doi: 10.1609/aaai.v32i1.11819]

[84] Zintgraf LM, Cohen TS, Adel T, Welling M. Visualizing deep neural network decisions: Prediction difference analysis. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017. 1–11.

[85] Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: Removing noise by adding noise. arXiv:1706.03825, 2017.

[86] Singla S, Wallace E, Feng S, Feizi S. Understanding impacts of high-order loss approximations and features in deep learning interpretation. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 5848–5856.

[87] Wang SJ, Zhou TY, Bilmes JA. Bias also matters: Bias attribution for deep neural network explanation. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 6659–6667.

[88] Zeiler MD, Krishnan D, Taylor GW, Fergus R. Deconvolutional networks. In: Proc. of the 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 2528–2535. [doi: 10.1109/CVPR.2010.5539957]

[89] Zeiler MD, Taylor GW, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning. In: Proc. of the 2011 IEEE Int'l Conf. on Computer Vision. Barcelona: IEEE, 2011. 2018–2025. [doi: 10.1109/ICCV.2011.6126474]

[90] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 818–833. [doi: 10.1007/978-3-319-10590-1_53]

[91] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA. Striving for simplicity: The all convolutional net. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015. 1–11.

[92] Heskes T, Sijben E, Bucur IG, Claassen T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In: Proc. of the 34th Advances in Neural Information Processing Systems. 2020. 4778–4789.

[93] Ancona M, Öztireli C, Gross MH. Explaining deep neural networks with a polynomial time algorithm for Shapley value approximation. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 272–281.

[94] Van Looveren A, Klaise J. Interpretable counterfactual explanations guided by prototypes. In: Proc. of the 2021 European Conf. on Machine Learning and Knowledge Discovery in Databases. Bilbao: Springer, 2021. 650–665. [doi: 10.1007/978-3-030-86520-7_40]

[95] Koh PW, Liang P. Understanding black-box predictions via influence functions. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 1885–1894.

[96] Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 3449–3457. [doi: 10.1109/ICCV.2017.371]

[97] Agarwal C, Nguyen A. Explaining image classifiers by removing input features using generative models. In: Proc. of the 15th Asian

Conf. on Computer Vision. Kyoto: Springer, 2020. 101–118. [doi: 10.1007/978-3-030-69544-6_7]

[98] Wagner J, Köhler JM, Gindele T, Hetzel L, Wiedemer JT, Behnke S. Interpretable and fine-grained visual explanations for convolutional neural networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9097–9107. [doi: 10.1109/CVPR.2019.00931]

[99] Xu KD, Liu SJ, Zhang GY, Sun MS, Zhao P, Fan QF, Gan C, Lin X. Interpreting adversarial examples by activation promotion and suppression. arXiv:1904.02057, 2019.

[100] Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking clever Hans predictors and assessing what machines really learn. Nature Communications, 2019, 10(1): 1096. [doi: 10.1038/s41467-019-08987-4]

[101] Yang CL, Rangarajan A, Ranka S. Global model interpretation via recursive partitioning. In: Proc. of the 20th IEEE Int'l Conf. on High Performance Computing and Communications, the 16th IEEE Int'l Conf. on Smart City, the 4th IEEE Int'l Conf. on Data Science and Systems. Exeter: IEEE, 2018. 1563–1570. [doi: 10.1109/HPCC/SmartCity/DSS.2018.00256]

[102] Salman S, Payrovnaziri SN, Liu XW, Rengifo-Moreno P, He Z. DeepConsensus: Consensus-based interpretable deep neural networks with application to mortality prediction. In: Proc. of the 2020 Int'l Joint Conf. on Neural Networks. Glasgow: IEEE, 2020. 1–8. [doi: 10.1109/IJCNN48605.2020.9206678]

[103] Ghorbani A, Wexler J, Zou J, Kim B. Towards automatic concept-based explanations. In: Proc. of the 33rd Conf. on Neural Information Processing Systems. Vancouver, 2019. 9273–9282.

[104] Hua YY, Zhang DC, Ge SM. Research progress in the interpretability of deep learning models. Journal of Cyber Security, 2020, 5(3): 1–12 (in Chinese with English abstract). [doi: 10.19363/J.cnki.cn10-1380/tn.2020.05.01]

[105] Si NW, Zhang WL, Qu D, Luo XY, Chang HY, Niu T. Representation visualization of convolutional neural networks: A survey. Acta Automatica Sinica, 2022, 48(8): 1890–1920 (in Chinese with English abstract). [doi: 10.16383/j.aas.c200554]

[106] Shi XR, Ni L, Wang J, Guo YH. Interpretable CNN based on minimum entropy constraint. Aerospace Control, 2021, 39(5): 39–43 (in Chinese with English abstract). [doi: 10.3969/j.issn.1006-3242.2021.05.007]

[107] Adler P, Falk C, Friedler SA, Nix T, Rybeck G, Scheidegger C, Smith B, Venkatasubramanian S. Auditing black-box models for indirect influence. Knowledge and Information Systems, 2018, 54(1): 95–122. [doi: 10.1007/s10115-017-1116-3]

[108] Michelini PN, Liu HW, Lu YH, Jiang XQ. Understanding convolutional networks using linear interpreters—Extended abstract. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision Workshop. Seoul: IEEE, 2019. 4186–4189. [doi: 10.1109/ICCVW.2019.00514]

[109] Michelini PN, Liu HW, Lu YH, Jiang XQ. A tour of convolutional networks guided by linear interpreters. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4752–4761. [doi: 10.1109/ICCV.2019.00485]

[110] Zhang JM, Bargal SA, Lin Z, Brandt J, Shen XH, Sclaroff S. Top-down neural attention by excitation backprop. Int'l Journal of Computer Vision, 2018, 126(10): 1084–1102. [doi: 10.1007/s11263-017-1059-x]

[111] Singh C, Murdoch WJ, Yu B. Hierarchical interpretations for neural network predictions. In: Proc. of the 27th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019. 1–11.

[112] Chen JB, Song L, Wainwright MJ, Jordan MI. Learning to explain: An information-theoretic perspective on model interpretation. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 882–891.

[113] Herman B. The promise and peril of human evaluation for model interpretability. arXiv:1711.07414, 2019.

[114] Chang CH, Creager E, Goldenberg A, Duvenaud D. Explaining image classifiers by counterfactual generation. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019. 1–11.

[115] Wang HF, Du MN, Yang F, Zhang ZJ. Score-CAM: Improved visual explanations via score-weighted class activation mapping. arXiv:1910.01279, 2020.

[116] Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: Proc. of the 2018 IEEE Winter Conf. on Applications of Computer Vision. Lake Tahoe: IEEE, 2018. 839–847. [doi: 10.1109/WACV.2018.00097]

[117] Ross AS, Hughes MC, Doshi-Velez F. Right for the right reasons: Training differentiable models by constraining their explanations. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Melbourne: IJCAI.org, 2017. 2662–2670.

[118] Soares E, Angelov P, Biaso S, Froes MH, Abe DK. SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. medRxiv, 2020. [doi: 10.1101/2020.04.24.20078584]

[119] Tetila E, Bressem K, Astolfi G, Sant'Ana DA, Pache MC, Pistori H. System for quantitative diagnosis of COVID-19-associated pneumonia based on superpixels with deep learning and chest CT. Research Square, 2020. [doi: 10.21203/rs.3.rs-123158/v1]

[120] Soares E, Angelov P. Fair-by-design explainable models for prediction of recidivism. arXiv:1910.02043, 2019.

[121] Soares E, Angelov P, Costa B, Castro M. Actively semi-supervised deep rule-based classifier applied to adverse driving scenarios. In:

Proc. of the 2019 Int'l Joint Conf. on Neural Networks. Budapest: IEEE, 2019. 1–8. [doi: 10.1109/IJCNN.2019.8851842]

[122] Soares E, Angelov PP, Costa B, Castro MPG, Nageshrao S, Filev D. Explaining deep learning models through rule-based approximation and visualization. IEEE Trans. on Fuzzy Systems, 2021, 29(8): 2399–2407. [doi: 10.1109/TFUZZ.2020.2999776]

[123] Gao JY, Wang XT, Wang YS, Xie X. Explainable recommendation through attentive multi-view learning. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conf. and the 9th AAAI Symp. on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019. 445. [doi: 10.1609/aaai.v33i01.33013622]

附中文参考文献:

[2] 孔祥维, 唐鑫泽, 王子明. 人工智能决策可解释性的研究综述. 系统工程理论与实践, 2021, 41(2): 524–536. [doi: 10.12011/SETP2020-1536]

[7] 苏炯铭, 刘鸿福, 项凤涛, 吴建宅, 袁兴生. 深度神经网络解释方法综述. 计算机工程, 2020, 46(9): 1–15. [doi: 10.19678/j.issn.1000-3428.0057951]

[30] 包希港, 周春来, 肖克晶, 覃飙. 视觉问答研究综述. 软件学报, 2021, 32(8): 2522–2544. http://www.jos.org.cn/1000-9825/6215.htm [doi: 10.13328/j.cnki.jos.006215]

[32] 阮利, 温莎莎, 牛易明, 李绍宁, 薛云志, 阮涛, 肖利民. 基于可解释基拆解和知识图谱的深度神经网络可视化. 计算机学报, 2021, 44(9): 1786–1805. [doi: 10.11897/SP.J.1016.2021.01786]

[37] 王婉臻, 饶元, 吴连伟, 李薛. 基于人工智能的司法判决预测研究与进展. 中文信息学报, 2021, 35(9): 1–14. [doi: 10.3969/j.issn.1003-0077.2021.09.001]

[43] 陈珂锐, 孟小峰. 机器学习的可解释性. 计算机研究与发展, 2020, 57(9): 1971–1986. [doi: 10.7544/issn1000-1239.2020.20190456]

[48] 周志杰, 曹友, 胡昌华, 唐帅文, 张春潮, 王杰. 基于规则的建模方法的可解释性及其发展. 自动化学报, 2021, 47(6): 1201–1216. [doi: 10.16383/j.aas.c200402]

[63] 成科扬, 孟春运, 王文杉, 师文喜, 詹永照. 解耦表征学习研究进展. 计算机应用, 2021, 41(12): 3409–3418. [doi: 10.11772/j.issn.1001-9081.2021060895]

[104] 化盈盈, 张岱墀, 葛仕明. 深度学习模型可解释性的研究进展. 信息安全学报, 2020, 5(3): 1–12. [doi: 10.19363/J.cnki.cn10-1380/tn.2020.05.01]

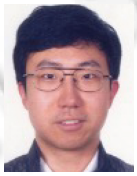
[105] 司念文, 张文林, 屈丹, 罗向阳, 常禾雨, 牛铜. 卷积神经网络表征可视化研究综述. 自动化学报, 2022, 48(8): 1890–1920. [doi: 10.16383/j.aas.c200554]

[106] 石晓荣, 倪亮, 王健, 郭宇航. 基于最小熵约束的可解释卷积神经网络. 航天控制, 2021, 39(5): 39–43. [doi: 10.3969/j.issn.1006-3242.2021.05.007]

窦慧(1992－), 女, 博士生, 主要研究领域为神经网络, 数据可视化.

申富饶(1973－), 男, 博士, 教授, 博士生导 师, CCF 杰出会员, 主要研究领域为神经计算, 机器人智能.

张凌茗(1997－), 男, 硕士生, 主要研究领域为神经网络, 计算机视觉.



赵健(1979－), 男, 博士, 副教授, 主要研究领域为通信网络, 神经计算.

韩峰(1994－), 男, 博士生, 主要研究领域为计算机视觉, 多模态学习.