◎热点与综述◎

深度学习模型可解释性研究综述

曾春艳[1],严 康[1],王志锋[2],余 琰[1],纪纯妹[3]
1.湖北工业大学 太阳能高效利用及储能运行控制湖北省重点实验室,武汉 430068
2.华中师范大学 数字媒体技术系,武汉 430079
3.中国移动通信集团广东有限公司 汕头分公司,广东 汕头 515041

摘 要:深度学习技术以数据驱动学习的特点,在自然语言处理、图像处理、语音识别等领域取得了巨大成就。但由于深度学习模型网络过深、参数多、复杂度高等特性,该模型做出的决策及中间过程让人类难以理解,因此探究深度学习的可解释性成为当前人工智能领域研究的新课题。以深度学习模型可解释性为研究对象,对其研究进展进行总结阐述。从自解释模型、特定模型解释、不可知模型解释、因果可解释性四个方面对主要可解释性方法进行总结分析。列举出可解释性相关技术的应用,讨论当前可解释性研究存在的问题并进行展望,以推动深度学习可解释性研究框架的进一步发展。

关键词:深度学习;可解释性;人工智能;因果可解释;自解释
文献标志码:A **中图分类号:**TN912 doi:10.3778/j.issn.1002-8331.2012-0357

Survey of Interpretability Research on Deep Learning Models

ZENG Chunyan[1], YAN Kang[1], WANG Zhifeng[2], YU Yan[1], JI Chunmei[3]
1.Hubei Key Laboratory for High-efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan 430068, China
2.Department of Digital Media Technology, Central China Normal University, Wuhan 430079, China
3.Shantou Branch, China Mobile Group Guangdong Co., Ltd., Shantou, Guangdong 515041, China

Abstract: With the characteristics of data-driven learning, deep learning technology has made great achievements in the fields of natural language processing, image processing, and speech recognition. However, due to the deep learning model featured by deep networks, many parameters, high complexity and other characteristics, the decisions and intermediate processes made by the model are difficult for humans to understand. Therefore, exploring the interpretability of deep learning has become a new topic in the current artificial intelligence field. This review takes the interpretability of deep learning models as the research object and summarizes its progress. Firstly, the main interpretability methods are summarized and analyzed from four aspects: self-explanatory model, model-specific explanation, model-agnostic explanation, and causal interpretability. At the same time, it enumerates the application of interpretability related technologies, and finally discusses the existing problems of current interpretability research to promote the further development of the deep learning interpretability research framework.

Key words: deep learning; interpretability; artificial intelligence; causal interpretability; self-explanatory

在深度神经网络(Deep Neural Networks,DNN)的推动下,深度学习在自然语言处理[1]、图像处理[2]、语音识别[3]等相关领域取得了重大突破。深度神经网络成功的一个关键因素是它的网络足够深,大量非线性网络层的复杂组合能对原始数据在各种抽象层面上提取特征。然而,由于大多数深度学习模型复杂度高、参数多、

基金项目:国家自然科学基金(61901165,61501199);湖北省自然科学基金(2017CFB683);华中师范大学中央高校基本科研业务费项目(CCNU20ZT010)。

作者简介:曾春艳(1986—),女,博士,副教授,CCF会员,研究领域为信号处理、可解释机器学习;王志锋(1985—),通信作者,男,博士,副教授,CCF会员,研究领域为信号处理、可解释机器学习,E-mail:zfwang@mail.ccnu.edu.cn。

收稿日期:2020-12-21 **修回日期:**2021-01-21 **文章编号:**1002-8331(2021)08-0001-09

透明性低,如黑盒一般,人们无法理解这种“端到端”模型做出决策的机理,无法判断决策是否可靠。

深度学习模型在实际应用中产生了各种问题,可解释性的缺失使这些问题难以解决。例如在医疗领域,看似非常精确的用于肺炎患者预后的系统,非常依赖数据集中的虚假相关性。该系统预测有哮喘病史的病人死于肺炎的风险较低,但这是因为哮喘病人得到了更快、更好的关注,才导致他们的死亡率更低。实际上,这类病人死于肺炎的风险会更高[4]。在司法领域,借助COMPAS软件对罪犯再犯风险的评估,法官能更合理地对保释金额、判刑等做出决策。但由于训练和测试模型时所用数据库的代表样本不足或是无关统计相关等原因,模型存在潜在的人种偏见、性别歧视,或是其他各种主观偏见[5]。在图像处理领域,高度精确的神经网络面对对抗攻击却很脆弱。只需对图像的像素进行不可察觉的改变,图像处理神经网络就可以将其任何预测改变为任何其他可能的预测[6]。

对于安全性要求苛刻的应用,例如健康诊断、信用额度、刑事审判,人们依旧选择线性回归、决策树等不那么精确,但人类可理解的模型。深度学习模型可解释性的缺失,严重阻碍了其在医学诊断[7]、金融[8]、自动驾驶[9]、军事[10]等高风险决策领域的应用。因此研究可解释性就显得格外重要,可解释性有利于系统的使用者更好地理解系统的强项和不足,并明确系统的知识边界,了解系统在何种情况下有效,从而恰当地信任和使用系统来进行预测。对系统的设计者而言,可解释性有利于优化和改进系统,避免模型中的歧视和偏见,并加强对系统的管理和监控。

最近5年(2016—2020)来,深度学习模型可解释性研究引起了学术界和企业界的高度关注,人们相继提出各类解释方法来试图解决模型“黑盒”问题,发表在各大顶级期刊和会议上关于可解释性的论文呈上升趋势。中国于2019年7月在中央全面深化改革委员会第九次会议上审议通过了《国家科技伦理委员会组建方案》,全面启动了包含“知识可解释”在内的科技伦理建设工作,以确保人工智能安全、可靠、可控。欧盟于2019年4月,出台了正式版的《人工智能道德准则》,提出了实现可信赖人工智能全生命周期框架,包含可解释、安全、隐私和透明等方面。美国国防高级研究计划局(Defense Advanced Research Projects Agency,DARPA)在2016年8月提出了可解释的人工智能(Explainable Artificial Intelligence,XAI)计划[10],目的是建设一套全新、可解释的深度学习模型。该计划的四大原则草案由美国国家标准局于2020年8月公开,分别是提供解释、解释有意义、解释的精确度、系统的知识边界。

由于不同研究者看待问题的角度不同,他们赋予可解释性的定义也不同,目前可解释性没有一个统一的定义。本质上说,在人工智能领域,深度学习可解释性理解为模型决策结果以可理解的方式向人类呈现,它有助于人们理解复杂模型的内部工作机制以及它们如何做出特定决策等重要问题。关于可解释性理论和方法的梳理总结,最早由Miller[11]从哲学、心理学、认知科学和人机交互领域,对人们如何定义、选择、评估和呈现解释进行综述。随后Du等人[12]根据获得可解释性的时间不同,将其分为内在、事后可解释性,并进一步根据解释范围不同将事后可解释性分为全局、局部可解释性。全局可解释性从整体上理解模型如何进行预测,而局部可解释性为模型的单个预测提供局部解释。最近苏炯铭等人[13]对卷积神经网络(Convolutional Neural Networks,CNN)、循环神经网络(Recurrent Neural Networks,RNN)、生成对抗网络(Generative Adversarial Networks,GAN)等典型网络的解释方法进行科学的总结。化盈盈等人[14]基于可解释性研究的原理,从模型结构、特征分析、可解释性迁移三个角度对可解释性方法进行系统性的介绍。

然而,目前深度学习模型可解释性研究综述缺乏从因果关系的角度对该模型做出决策的原因进行分析总结,而因果关系比可视化等解释方法更能深入可解释性问题本质。基于此,本文加入因果可解释性研究的最新进展,并整体梳理了目前可解释性框架。本文结构如下:首先从自解释模型、特定模型解释、不可知模型解释、因果可解释性四个方面对可解释方法进行分类。自解释模型主要指传统机器学习算法,其本身具备可解释性,同时也是其他可解释方法建立的基础。其次特定模型解释专门为特定类型模型设计,并通过研究模型内部来获得解释。而不可知模型解释不关心模型的中间过程,只分析模型的输入输出。最后因果可解释性从因果关系的角度对模型做出决策的原因进行分析总结,可解释方法分类及其各方法特点分析如表1所示。同时列举出可解释性相关技术的应用,最后讨论可解释性研究当前存在的问题,展望其未来研究方向。

1 自解释模型

实现可解释性的最简单方法是仅使用创建自解释模型的算法子集[15]。自解释模型本身内嵌可解释性,通常结构简单、易于实现,人们很容易理解其决策过程,线性回归、决策树是这类模型典型代表,常见自解释模型及其特点总结如表2所示。

一般规律下,自解释模型的内嵌可解释性与准确性之间存在一个平衡。如果自解释模型结构简单,可解释性好,那么模型的拟合能力必会受到限制,导致其预测精度不高,会限制这些算法的应用场景。为解决此问题,研究者们将复杂的黑盒模型迁移到自解释模型中,从而实现对黑盒模型决策结果的解释。但规则复杂的

表1 方法分类及特点分析

类别	方法	特点分析
自解释模型	线性回归	形式简单,只能表示线性关系
	决策树	训练数据少,容易过拟合
特定模型解释	激活最大化	理解和呈现DNN的内部表征,可能产生噪声图像
	基于梯度解释方法	计算效率高,无法量化特征对决策结果的重要程度
	基于类激活映射方法	出色的定位物体的能力,提供粗粒度的解释结果
不可知模型解释	LIME	实现简单,无法解释模型的整体决策行为
	知识蒸馏	易于理解,对原始模型的全局近似
因果可解释性	基于模型的解释	估计模型组件对输出的因果影响
	反事实解释	回答"为什么"的问题
	决策公平性	辨别歧视问题
	认知和因果推理	结合认知和因果推理构建模型

表2 自解释模型总结

模型	优点	缺点	适用场景
线性回归	形式简单,易于建模	只能表示线性关系,难以表达复杂数据	信贷风控
决策树	易于理解和实现,训练数据少	容易过拟合,对连续字段难预测	搜索排序
朴素贝叶斯	有稳定分类效率,对数据缺失不敏感	决策存在错误率,对数据表达敏感	垃圾邮件过滤
逻辑回归	模型清晰,实施简单高效	容易欠拟合,精度不高	点击率预估

模型或树深度极深的决策树,人类未必能理解[16],其内嵌可解释性也未必优于深度神经网络。

1.1 线性回归

线性回归[17]通过拟合自变量和因变量之间的最佳线性关系来预测目标变量,其框架如下:

$$y=w_0+w_1x_1+w_2x_2+\cdots+w_nx_n+b \tag{1}$$

其中,w、b 分别表示学习到的权重和偏差,目标变量 y 就是 n 个特征变量的权重和。特征变量在整个变量中所占的比重越大,该特征对最终预测结果的影响力越大。

线性回归模型通常形式简单、易于建模、可解释性好,人们通常利用它来模拟黑盒模型的输出[18]。该方法利用迁移学习,将黑盒模型迁移到自解释模型中,从而实现对模型决策结果的解释。因为线性回归模型只考虑了自变量与因变量之间的线性相关关系,但也正因如此,它无法处理更复杂的关系,导致模型在测试集上的预测精度也可能比较低。如果模型的输入是复杂或难以理解的特征,那么模型将不可分解。同样,如果模型太大,以至于人类无法将模型视为一个整体,那么它的可模拟性就会受到质疑。

1.2 决策树

决策树是一种典型的机器学习自解释模型,它是一种树结构框架,由内部节点和叶节点组成。其中内部节点表示判断条件,而叶节点对应决策结果。决策树对数据的属性进行判断,得到分类或回归结果,是一种基于if-then决策规则的算法[19]。决策树的每条规则都给最后的结果提供了解释,让人们可以直观地理解模型做出的决策。

当决策树节点较少时,可作为一种自解释模型来模拟黑盒模型的输出。Zhang等人[20]通过学习决策树来量化解释卷积神经网络在语义层面上的预测逻辑,决策树告诉人们哪些物体部分被用于预测,以及它们对预测得分的贡献程度。但是随着树深度加深,叶节点增多,人们就很难理解树的决策规则[16]。

2 特定模型解释

特定模型(Model-Specific,MS)解释方法是专门为特定类型的模型设计的,该方法将待解释模型视为白盒,通过研究模型内部的结构和参数来获得解释。典型的特定模型解释方法包括激活最大化、基于梯度解释方法、基于类激活映射方法等,各方法的优缺点及适用场景如表3所示。这类解释方法通常使用可视化技术来有效地帮助人们理解DNN内部的工作机制,既直观且操作简单。但可视化技术存在一定局限性,人们对其可视化效果只能从肉眼判断或一般只是粗粒度的理解,无法量化特征对决策结果的重要程度。

2.1 激活最大化

为了深入理解和呈现DNN的内部表征,人们通常对DNN任意层神经单元计算内容进行可视化操作。其中最有效的一种方法是激活最大化(Activation Maximization,AM)[21],该方法的核心思想是寻找一个最大化特定层神经元激活值的输入模式,即寻找哪些输入会使激活函数值最大。其框架为:

表3 特定模型解释方法

方法	优点	缺点	适用场景
激活最大化	帮助人们理解模型内部工作逻辑	可能产生噪声图像,只用于连续型数据	解释连续型数据模型
基于梯度解释方法	可有效定位输入图像的决策特征	无法量化特征对决策结果的重要程度	解释神经网络
基于类激活映射方法	出色的定位物体的能力	只提供粗粒度的解释结果	识别分类物体

$$I^* = \underset{I}{\operatorname{argmax}} S_c(I) \tag{2}$$

其中，$S_c(I)$ 表示第 C 层神经元对输入 I 的激活值，利用反向传播法找到一个局部最优的 I 。从随机初始化开始，通过迭代优化，利用神经元激活值对图像的导数来调整图像。最后可视化生成的图像可以呈现每个神经元在其感受野中所捕获的内容。该方法虽然简单，但优化过程中不进行适当的正则化约束，会生成能最大化神经元激活却无法识别的图像。为解决此问题，可使用自然图像先验约束，来生成与自然图像相似的合成图像，其框架如式(3)所示：

$$I^* = \underset{I}{\operatorname{argmax}} S_c(I) - \lambda \|I\|_2^2 \tag{3}$$

另外可以使用由生成模型产生的自然图像先验，将潜在空间中的编码映射到图像空间，从而对优化过程进行正则化约束[22]。

激活最大化方法解释结果较精确，可以帮助人们理解DNN内部工作逻辑，但其优化过程中可能产生含有噪声的图像，导致输入 I 难以解释。此外，激活最大化方法只适用于连续型数据，难以解释含离散型数据模型，诸如自然语言处理模型。

2.2　基于梯度解释方法

人们常利用基于梯度解释方法的可视化技术来解释深度神经网络。方法主要有四种：反卷积(Deconvolutio)、导向反向传播(Guided Backpropagation)、积分梯度(Integrated Gradients)和平滑梯度(Smooth Gradients)，各方法总结如表4所示。该方法的核心思想是利用反向传播计算特定输出相对于输入的梯度来推导出特征重要性。

表4　基于梯度解释方法

方法	特点分析
反卷积	能大致看清目标轮廓，存在大量噪声
导向反向传播	几乎没有噪声，目标特征较集中
积分梯度	解释CNN内部结构，存在较少噪声
平滑梯度	定位图像的决策特征，无法量化贡献

最早Zeiler等人[23]利用反卷积实现特征可视化，来解释CNN每一层学到的东西。可视化效果图能大致看清目标轮廓，但仍存在大量噪声。随后Springenberg等人[24]将反向传播(Backpropagation)与反卷积网络相结合提出导向反向传播方法。与反卷积方法不同，使用导向反向传播几乎没有噪声，且目标特征较集中。Sundararajan等人[25]提出了一种积分梯度方法，它通过对梯度进行积分可量化输入的各个分量的重要性，有效地解决了DNN中神经元饱和问题导致无法利用梯度信息反映特征重要性的问题。Smilkov等人[26]提出了一种平滑梯度方法，该方法通过向输入图像中添加噪声，然后利用反向传播方法求解扰动图像的梯度，最后将求解得到的灵敏度图进行平均，并作为对最终决策结果的解释。

基于梯度解释方法可以观察解释CNN内部结构外，还能有效定位输入图像的决策特征。但是大多数方法的解释结果中依然存在清晰可见的噪声，且无法判断这些噪声是否真实反映模型的决策依据，同时梯度信息只能用于定位重要特征，却无法量化每个特征对决策结果的贡献程度[27]。

2.3　基于类激活映射方法

卷积神经网络里的最后一层卷积单元本身就具有出色的定位物体的能力，但这种定位能力在使用全连接层进行分类操作的过程中会丧失，位置信息难以用可视化的方法表现出来。因此，为了合理解释卷积神经网络的分类结果，必须充分利用最后一层卷积层。

Zhou等人[28]提出类激活映射(Class Activation Mapping，CAM)方法来解释卷积神经网络，并识别分类物体的位置。CAM将卷积神经网络中的全连接层替换成了全局平均池化(Global Average Pooling，GAP)层，得到最后一个卷积层中每个特征图的均值，并进行加权求和得到最终决策结果，其框架如图1所示。同时加权求和操作之后会得到模型的类激活图，通过以热力图的形式对类激活图进行可视化，从而得到分类结果的显著特征。

虽然CAM方法没有全连接层，可以减少参数，防止过拟合，且具有良好的定位能力来识别分类物体的位置。但运用CAM方法时需要修改原模型的网络结构，重新训练模型，这就导致在实际应用中会花费更多的时间与成本。

为解决CAM方法存在的问题，Selvaraju等人[29]提出了基于梯度加权类激活映射(Gradient-weighted Class Activation Mapping，Grad-CAM)方法对卷积神经网络生成视觉解释。与CAM的区别在于，Grad-CAM通过梯度的全局平均来计算权重，通过得到特征图对应的权重进行加权求和，最后以热力图的方式可视化神经网络的注意力，效果图如图2所示。

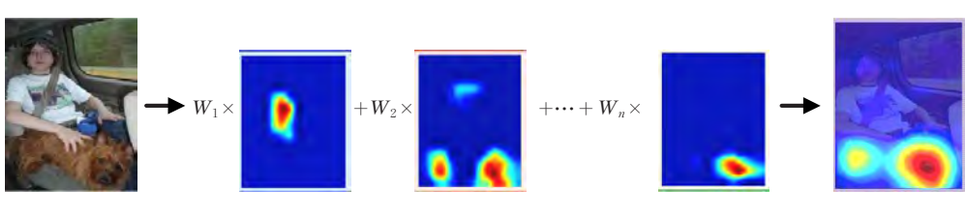

图1　CAM方法框架

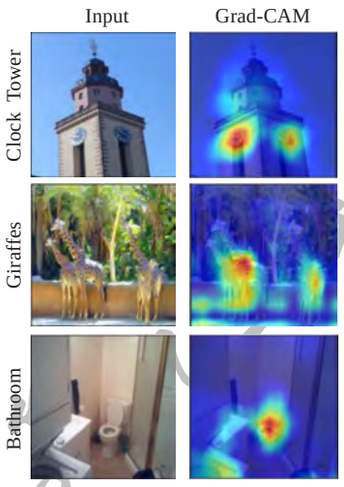

图2 Grad-CAM效果图

与CAM不同,Grad-CAM无需修改模型结构或重新训练模型,适用于各种CNN模型。但Grad-CAM只能提供粗粒度的解释结果,因此无法应用于医学诊断、金融等对解释结果高精度需求的应用场景。

3 不可知模型解释

不可知模型(Model-Agnostic,MA)解释方法是指解释与模型分离,并将待解释模型视为黑盒,通过分析模型的输入和输出来解释模型的预测。与特定模型解释方法深入研究模型结构和参数不同的是,不关心模型的中间过程,只分析模型的输入输出。典型的不可知模型解释方法包括LIME(Local Interpretable Model-Agnostic Explanation)、知识蒸馏等,各方法的优缺点及适用场景如表5所示。这类解释方法灵活性强,适用于任何类型的模型,可直接从其预测过程中提取重要知识,也可通过模型代理方法来降低模型操作复杂度。但该解释方法只能对待解释模型进行局部近似,因而只能捕获模型的局部特征,无法解释模型的整体决策行为。

表5 不可知模型解释方法

方法	优点	缺点	适用场景
LIME	操作简单,易于理解	只是对待解释模型的局部近似	局部近似解释
知识蒸馏	有效压缩模型大小	适用条件苛刻过程不可控	模型压缩

3.1 LIME

LIME[18]是一种局部近似解释的代表性方法,该方法的核心思想是:首先向输入样本中添加轻微的扰动,并观察待解释模型的输出变化,根据这种变化在原始输入中训练一个线性模型,并利用线性模型来局部近似待解释模型的预测,其中线性模型的权重系数表现为决策中该样本的每一维特征重要性。

由于LIME无法对RNN模型提出高保真的解释,Guo等人[30]提出了LEMNA(Local Explanation Method using Nonlinear Approximation),其核心思想是通过融合lasso正则的混合回归模型逼近复杂深度学习决策边界的局部区域,解决了现有的解释技术无法处理RNN模型特征依赖和非线性局部边界问题。

LIME方法操作简单,易于理解,并在图像分类任务中得到了很好的解释性效果。但针对相同的预测结果,模型给出的解释会不同,导致该方法缺乏稳定性。其次LIME只是对待解释模型的局部近似,无法解释模型的整体决策行为。针对每个输入样本,LIME方法均需要重新训练一个线性模型来拟合待解释模型的局部边界,因此此方法使用效率不高。

3.2 知识蒸馏

知识蒸馏是一种基于迁移学习的模型压缩方法,该方法将复杂的教师模型学习到的知识迁移到学生模型上,在保证学生模型性能的前提下,参数量大幅降低,进而实现对教师模型的全局解释。

最早Hinton等人[31]提出了一种知识蒸馏方法,该方法将复杂网络训练得到的知识压缩到单一网络中,来模拟复杂网络的决策过程,并且单一网络能达到复杂网络几乎同样的性能。Frosst等人[32]扩展了Hinton提出的知识蒸馏方法,提出利用决策树来模拟复杂深度神经网络模型的决策,对于相同的输入特征 $x_1,x_2,\cdots,x_n$,通过训练一个决策树来模拟黑盒模型的输出,使其有相似输出,即 $y_1\approx y_1',y_2\approx y_2',y_n\approx y_n'$,从而解释黑盒模型的决策,如图3所示。最近Cheng等人[33]通过量化和分析DNN中间层特征的知识来解释知识蒸馏的成功机理,并提出了三种类型的数学度量来评估DNN的特征表示。

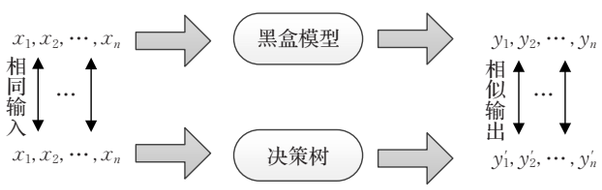

图3 利用决策树解释黑盒模型

知识蒸馏解释方法实现简单,可有效压缩模型大小,广泛应用于解释黑盒模型。但其适用条件苛刻,要求合适的教师和学生模型;过程不可控,无法保证知识迁移的效果;同时该方法只是对复杂模型的全局近似,其解释未必能真实反映复杂模型的决策行为。

4 因果可解释性

Pearl[34]指出当前机器学习系统几乎完全以统计学或黑盒子的方式运行,这对系统的性能造成严重的局限性。他结合因果推理理论,将因果可解释性分为三个不同层次来描述机器学习系统面临的阻碍,分别是统计可解释性、因果介入可解释性、反事实可解释性,并认为产生反事实解释是实现最高层次可解释性的方法。主流

机器学习方法主要侧重于统计可解释性,只能回答部分可解释性问题。因果可解释性旨在回答与因果介入可解释性和反事实可解释性相关的问题,可回答模型做出决策的真正原因。

为了产生更易于人类理解的解释,最近研究者提出了因果可解释性模型[35]。该模型将可解释性与因果关系紧密联系,以帮助人类理解算法做出决策的真正原因,能回答“为什么这个模型会做出这样的决定?”的问题。本章将从基于模型的解释、反事实解释和决策公平性三个方面对因果可解释性进行综述,其总结如表6所示。

表6 因果可解释性

方法	特点分析
基于模型的解释	估计模型组件对输出的因果影响
反事实解释	回答“为什么”的问题
决策公平性	辨认歧视问题
认知和因果推理	结合认知和因果推理构建模型

4.1 基于模型的解释

基于模型的解释方法是指解释模型中每个组成部分对最终决策的因果影响。通常可解释性框架不能回答诸如“深层神经网络的第 m 层的第 n 个卷积核对模型预测的影响是什么”的问题。于是提出因果解释框架来解决此问题,其主要思想是将DNN的结构建模为结构因果模型(Structural Causal Model,SCM),并通过执行因果推理来估计模型的每个组件对输出的因果影响。

Narendra等人[36]将DNN视为SCM,在模型的每个卷积核上应用函数来获得目标值,例如每个卷积核的方差或期望值。Harradon等人[37]进一步提出,为了产生有效的可解释性,有必要建立一个人类可理解的DNN因果模型,它允许进行各种因果干预。基于此假设,Harradon提出了一种可解释性框架,该框架可从深度神经网络中提取人类可理解的概念(例如猫的眼睛和耳朵),并在SCM中学习输入、输出和这些概念之间的因果结构,并对其进行因果推理,以获得对模型的更多见解。

4.2 反事实解释

反事实解释是一种基于实例的解释,在特征和预测结果之间建立了一个因果关系,旨在回答“为什么”的问题,如“为什么模型的决定是 Y ?”或者“是输入 X 导致模型预测了 Y 吗?”。

对于如何生成反事实解释,Kanehira等人[38]提出了一种使用多模态信息为视频分类任务生成反事实解释的方法,方法分三步产生视觉语言解释。首先,训练分类模型,并希望为其生成解释。然后,通过利用第一步训练后模型的输出和中间层特征来训练事后解释模型,并解释模型预测所有负类(实例不属于的类)的反事实分数。最后,解释模型通过最大化正类和负类之间的反事实分数来生成解释。Hendricks等人[39]提出用自然语言生成反事实解释的方法。该框架在原始输入生成的文本解释中检查反事实类的证据,然后会检查这些因素是否存在于反事实的图像中,并返回现有的因素。

反事实解释清晰明了,实现较易,无需访问数据或模型,只需访问模型的预测函数。但对于每个实例,通常会产生多个反事实的解释,而且该方法不能很好地处理多层次的分类特征。

4.3 决策公平性

目前,在银行贷款额度和筛选求职者等任务中,其初级决策大多是由人工智能系统负责。任务复杂度高和对系统认知片面,导致目前无法解释人工智能系统是否在公平地运行,且无法辨认是否存在歧视问题。Lipton[40]指出可解释的模型是保证决策公平性必不可少的一部分。近年来,将公平性纳入决策方法受到极大的关注。

研究人员提出了一些度量标准来定量衡量算法决策的公平性[41]。Bareinboim等人[42]首先引入三种从原因到结果变化传递的细粒度度量,即反事实直接效应(Ctf-DE)、间接效应(Ctf-IE)和虚假效应(Ctf-SE),其框架如下:

给定一个SCM四元组 $M, X=x_1$ 对 Y 的反事实直接效应定义为:

$$DE_{x_0,x_1}(Y|X)=P\left(Y_{x_1}, W_{x_0}|X\right)-P\left(Y_{x_0}|X\right) \tag{4}$$

$X=x_1$ 对 $Y=y$ 的间接效应被定义为:

$$IE_{x_0,x_1}(Y|X)=P\left(Y_{x_0}, W_{x_1}|X\right)-P\left(Y_{x_0}|X\right) \tag{5}$$

$X=x_1$ 对 $Y=y$ 的虚假效应被定义为:

$$SE_{x_0,x_1}(Y|X)=P\left(Y_{x_0}|x_1\right)-P\left(Y|x_0\right) \tag{6}$$

这些度量使人们第一次能够准确地检测和区分三种最自然的歧视类型,即直接、间接和虚假的歧视。

4.4 认知与因果推理

在人工智能领域,人们利用因果关系构造出可以演绎推理的模型,人们曾相信,在严谨的数理逻辑理论下,机器会超越人类智能。然而,如今的人工智能依赖的不是规则下的推理,而是用深度学习产生近乎直觉的智能。虽然深度学习的过程清晰,算法明确,但深度学习模型网络过深、复杂度高,导致模型做出的决策及中间过程让人类难以理解。

在传统的电影推荐系统中,人们将电影类别和影星等作为特征属性,统计出各影片对这些属性的评级;然后依据顾客对这些属性个人偏好的加权,推荐综合加权评级较高的电影。然而在矩阵分解推荐系统中,同样的任务,它不再依赖人工选取的属性特征和统计评级,而是通过巨量的顾客购买记录和商品交易数据,用机器学习自动产生出商品的属性分类、顾客的偏好;然后依顾客的购物历史,计算对这些特征属性的偏好加权,向顾客作出推荐。这个系统产生的属性分类效果,比前者更有效,更具备解释性。

认知意味着一种同步关系的感知,是一种经验的延续,而因果推理具备严谨的数理逻辑。利用人类认知结合因果推理应用到深度学习模型的构建中,可以有效提高深度学习模型的可解释性。

5 可解释性的应用

在某些特定领域,深度学习模型的黑盒特性会严重限制其在该领域的应用。但随着可解释性方法的不断提出和可解释性模型的建立,深度学习可解释性已逐渐应用到医疗诊断、金融和模型诊断领域。

5.1 医疗诊断

深度学习技术已广泛应用于医疗诊断系统,诊断任务的准确率甚至超过了人类专家,但对于医生和病人而言,一个医疗诊断系统必须是可解释的。于是研究者们将深度学习技术和可解释性结合起来,用于辅助医生进行临床诊断。大量数据表明,医学诊断任务的准确率得到了进一步的提高,同时提高了诊断效率,也避免了人为主观错误。

Zhang等人[7]提出了一种新颖的医学图像诊断网络(MDNet),来生成诊断报告和相应的网络关注区域,使网络诊断和决策过程在语义和视觉上都具备可解释性。Yang等人[43]构建了一个带注意力机制的RNN模型,用于分析医疗条件与重症监护室死亡率之间的关系,研究结果表明,使用可解释性技术有助于发现与医疗保健中某些结果相关的潜在影响因素。

5.2 金融领域

人工智能广泛应用于金融领域,银行贷款额度和产品推荐系统都由特定的算法完成,因此算法的透明性格外重要。一个可解释的金融系统不仅可以获得准确的预测结果,而且能够帮助人们信任模型做出决策背后的原因。

银行监管部门可依据具有可解释性的神经网络来预测银行是否有破产的可能。Wang等人[8]提出了一种具有可解释性的自组织模糊神经网络推理系统。当与其他的预测模型进行比较时,它能够更好地根据财务报表中的变量识别财务困境的内在特征。这种基于自动生成的模糊推理规则库系统运行准确,更值得注意的是,简化后的规则库具有很高的可解释性。

5.3 模型诊断

在搭建模型时,由于深度学习模型的复杂性,会导致模型出错的几率增加。而可解释性作为一种解释模型的重要工具,当模型表现不佳或给出错误决策时,可以用来分析和调试模型的错误行为。将可解释性与模型诊断相结合引起人们的广泛关注。

Cadamuro等人[44]提出一种循环诊断的模型诊断方法,不断地检测模型漏洞,直到找到对模型漏洞贡献最大的实例,从而推断模型出错的根本原因。另外,可通过可视化模型卷积核和特征图的方式,来解释和诊断模型,并对模型的改进做出有效的指导。

6 存在的问题与研究展望

6.1 存在的问题及解决思路

上文对深度学习模型可解释性研究取得的进展进行了综述,从研究现状可以看出,深度学习模型可解释性研究还存在以下问题,并给出相应的解决思路。

(1)缺乏统一的解释方法评估指标。

由于解释范围、解释方法原理不同等因素,导致评估解释方法优劣的指标无法达成统一,目前可以从认知科学、定性和定量分析三个角度进行评估。但由于人类认知缺乏局限性,而定性评估又存在复杂性和主观性,因此可从定量分析角度来统一解释方法评估指标。

解决思路:① 在特定数据集上,通过比较解释一致性和解释保真度等指标来定量分析原模型和解释模型的预测值是否接近;② 通过分析滤波器可解释性和位置不稳定性指标来定理评估用于图像分类且生成神经网络视觉解释的解释方法的可解释性。

(2)无法平衡模型准确性和可解释性。

模型准确性是指模型的拟合能力,一般规律下,模型准确性与可解释性相对立,与复杂性相统一。通常自解释模型结构简单、参数少,其可解释性好,但准确性不高。而复杂的深度学习模型拟合能力强,同时损失部分可解释性来满足准确性,似乎模型的准确性和可解释性之间始终存在一个平衡。但杜克大学Cynthia Rudin基于一场可解释性机器学习挑战赛指出:更复杂的模型本质上更精确并不一定是真的,准确性与可解释性之间可以兼得,未来可能设计出高精度且强可解释性的模型。

解决思路:① 有效结合多种解释方法,充分发挥其各自优势;② 将深度学习模型先拆分为功能模块后再组合,根据模块功能生成解释。

(3)无法完全保护隐私。

隐私保护需要信息隐藏,而可解释性会造成信息透明,两者天然矛盾。提高可解释性可能造成敏感数据披露,进而可能造成歧视等现象出现。因此如何平衡数据隐私保护与模型可解释性是需要关注的问题,而联邦学习[45]作为一种具有隐私保护特性的分布式机器学习框架可有效解决此问题。

解决思路:① 尝试在模型进行训练时加入随机噪声,利用差分隐私(Differential Privacy,DP)和联邦学习结合,将模型输出结果隐藏起来;② 尝试对模型参数进行加密操作,将安全多方计算(Secure Multi-party Computation,SMC)和联邦学习结合,从而对数据进行隐私保护。

6.2 研究展望

未来深度学习可解释性研究可从以下两个方向考虑。

(1)利用知识图谱引入人类知识。

目前大多数深度学习模型很少使用知识图谱技术来引入人类知识。而知识图谱语义丰富,是人类知识的总结,利用知识图谱将人类知识引入到深度学习模型中,有助于特征理解,并在一定程度上减少数据偏见。因此利用知识图谱引入人类知识到模型是值得关注的问题。

(2)利用人机交互推动人类专家的参与。

可解释深度学习需要多学科协作,人机交互领域的知识和经验可推动人和机器的协同作用。解释的最终目的是帮助用户建立一个完整和正确的学习算法,并增强对其输出的信任。作为解释的最终用户,人类的参与将是未来研究方向。因此利用人机交互来推动人类专家的参与是未来需要考虑的问题。

7 结束语

可解释深度学习是一个活跃的领域,是各行各界关注的重要话题。本文从自解释模型、特定模型解释、不可知模型解释、因果可解释性四个方面对主要解释方法进行了归纳、分析。同时列举出可解释性相关技术的应用,并讨论了可解释性研究当前存在的问题,展望其未来研究方向。对可解释深度学习的现有技术进行全面概述旨在帮助人们更好地理解可解释性框架以及不同解释方法的优缺点。虽然可解释深度学习技术发展迅速,但仍有一些关键挑战尚未解决,需要未来的解决方案来进一步推动这一领域的发展。

参考文献:

[1] YUAN Y,ZHOU X,PAN S,et al.A relation-specific attention network for joint entity and relation extraction[C]//Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence,2020.

[2] WANG D,ZHAO K,WANG Y.Based on deep learning in traffic remote sensing image processing to recognize target vehicle[J].International Journal of Computers and Applications,2020,11:1-7.

[3] HO N H,YANG H J,KIM S H,et al.Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention based recurrent neural network[J].IEEE Access,2020,8:61672-61686.

[4] CARUANA R,LOU Y,GEHRKE J,et al.Intelligible models for healthcare:predicting pneumonia risk and hospital 30-day readmission[C]//The 21st SIGKDD International Conference on Knowledge Discovery and Data Mining,Sydney,Australia.New York:Association for Computing Machinery,2015:1721-1730.

[5] BUOLAMWINI J,GEBRU T.Gender shades:intersectional accuracy disparities in commercial gender classification[C]//Proceedings of the 1st Conference on Fairness,Accountability and Transparency,2018:77-91.

[6] MOOSAVIDEZFOOLI S,FAWZI A,FAWZI O,et al.Universal adversarial perturbations[C]//IEEE Conference on Computer Vision and Pattern Recognition,2017:86-94.

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

[8] WANG D,QUEK C,NG G S.Bank failure prediction using an accurate and interpretable neural fuzzy inference system[J].AI Communications,2016,29(4):477-495.

[9] KIM J,ROHRBACH A.Textual explanations for self-driving vehicles[C]//Proceedings of European Conference on Computer Vision.Berlin,Germany:Springer,2018:577-593.

[10] GUNNING D,AHA D W.Darpa's explainable artificial intelligence program[J].Magazine,2019,40(2):44-58.

[11] MILLER T.Explanation in artificial intelligence:insights from the social sciences[C]//2018 IEEE International Conference on Smart Cloud,2018.

[12] DU M,LIU N,HU X.Techniques for interpretable machine learning[J].Communications of the ACM,2020,63(1):68-77.

[13] 苏炯铭,刘鸿福,项凤涛,等.深度神经网络解释方法综述[J].计算机工程,2020,46(9):1-15.

[14] 化盈盈,张岱墀,葛仕明.深度学习模型可解释性的研究进展[J].信息安全学报,2020,5(3):1-12.

[15] MOLNAR C.Interpretable machine learning[EB/OL].[2020-09-10].https://christophm.github.io/interpretable-ml-book/.

[16] GUIDOTTI R,MONREALE A,RUGGIERI S,et al.A survey of methods for explaining black box models[J].ACM Computing Surveys,2018,51(5):93.

[17] HAUFE S,MEINECKE F,GORGEN K,et al.On the interpretation of weight vectors of linear models in multivariate neuroimaging[J].NeuroImage,2014,87:96-110.

[18] 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 Int Conf on Knowledge Discovery and Data Mining.New York:ACM,2016:1135-1144.

[19] BRESLOW L A,AHA D W.Simplifying decision trees:a survey[J].The Knowledge Engineering Review,1997,12(1):1-40.

[20] ZHANG Q S,YANG Y,MA H,et al.Interpreting CNNs via decision trees[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition,Long Beach,CA,USA,2019:6254-6263.

[21] SIMONYAN K, VEDALDI A, ZISSERMAN A.Deep inside convolutional networks: visualising image classification models and saliency maps[C]//ICLR Workshop,2014.

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

[23] ZEILER M D,FERGUS R.Visualizing and understanding convolutional networks[C]//Proceedings of the 13th European Conf on Computer Vision.Berlin: Springer, 2014:818-833.

[24] SPRINGENBERG J T,DOSOVITSKIY A,BROX T,et al. Striving for simplicity: the all convolutional net[J].arXiv: 1412.6806,2014.

[25] SUNDARARAJAN M,TALY A,YAN Q.Gradients of counterfactuals[J].arXiv:1611.02639,2016.

[26] SMILKOV D,THORAT N,KIM B,et al.SmoothGrad: removing noise by adding noise[J].arXiv:1706.03825,2017.

[27] 纪守领,李进锋,杜天宇,等.机器学习模型可解释性方法、应用与安全研究综述[J].计算机研究与发展,2019,56(10):2071-2096.

[28] ZHOU B L,KHOSLA A,LAPEDRIZA A,et al.Learning deep features for discriminative localization[C]//Proceedings of the 28th IEEE Conf on Computer Vision and Pattern Recognition.Piscataway, NJ: IEEE, 2016: 2921-2929.

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

[30] GUO W B,MU D L,XU J,et al.LEMNA: explaining deep learning based security applications[C]//Proceedings of the 2018 ACM SIGSAC Conf on Computer and Communications Security.New York:ACM,2018:364-379.

[31] HINTON G,VINYALS O,DEAN J.Distilling the knowledge in a neural network[J].Computer Science,2015,14(7):38-39.

[32] FROSST N,HINTON G.Distilling a neural network into a soft decision tree[J].arXiv:1711.09784,2017.

[33] CHENG X,RAO Z,CHEN Y,et al.Explaining knowledge distillation by quantifying the knowledge[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition,Seattle,United States,2020.

[34] PEARL J.Theoretical impediments to machine learning with seven sparks from the causal revolution[J].arXiv: 1801.04016,2018.

[35] MORAFFAH R,KARAMI M,GUO R,et al.Causal interpretability for machine learning-problems,methods and evaluation[J].ACM SIGKDD Explorations Newsletter,2020.

[36] NARENDRA T,SANKARAN A,VIJAYKEERTHY D, et al.Explaining deep learning models using causal inference[J].arXiv:1811.04376,2018.

[37] HARRADON M,DRUCE J,RUTTENBERG B E.Causal learning and explanation of deep neural networks via autoencoded activations[J].arXiv:1802.00541,2018.

[38] KANEHIRA A,TAKEMOTO K,INAYOSHI S,et al.Multimodal explanations by predicting counterfactuality in videos[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition,Long Beach,CA,USA,2019.

[39] HENDRICKS L A,HU R,DARRELL T,et al.Generating counterfactual explanations with natural language[J]. arXiv:1806.09809,2018.

[40] LIPTON Z C.The mythos of model interpretability[J]. Communications of the ACM,2018,61(10):36-43.

[41] KUSNER M J,LOFTUS J R,RUSSELL C,et al.Advances in neural information processing systems[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2018.

[42] ZHANG J,BAREINBOIM E.Fairness in decision making the causal explanation formula[C]//Association for the Advance of Artificial Intelligence, New Orleans, Louisiana,USA,2018.

[43] YANG C L,RANGARAJAN A,RANKA S.Global model interpretation via recursive partitioning[C]//Proceedings of the IEEE 4th Int Conf on Data Science and Systems.Piscataway,NJ:IEEE,2018:1563-1570.

[44] CADAMURO G,GILAD-BACHRACH R,ZHU X.Debugging machine learning models[C]//Proceedings of the 3rd ICML Workshop on Reliable Machine Learning in the Wild.Tahoe City,CA: International Machine Learning Society,2016.

[45] 王健宗,孔令炜,黄章成,等.联邦学习算法综述[J].大数据,2020,6(6):64-82.