计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190485
57(6):1208-1217,2020

深度学习可解释性研究进展

成科扬[1,2] 王 宁[1] 师文喜[2,3] 詹永照[1]
[1](江苏大学计算机科学与通信工程学院 江苏镇江 212013)
[2](社会安全风险感知与防控大数据应用国家工程实验室(中国电子科学研究院) 北京 100041)
[3](新疆联海创智信息科技有限公司 乌鲁木齐 830001)
(kycheng@ujs.edu.cn)

Research Advances in the Interpretability of Deep Learning

Cheng Keyang[1,2], Wang Ning[1], Shi Wenxi[2,3], and Zhan Yongzhao[1]
[1](*School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu* 212013)
[2](*National Engineering Laboratory for Public Safety Risk Perception and Control by the Big Data (China Academy of Electronic Sciences), Beijing* 100041)
[3](*Xinjiang Lianhaichuangzhi Information Technology Co. LTD, Urumqi* 830001)

Abstract The research on the interpretability of deep learning is closely related to various disciplines such as artificial intelligence, machine learning, logic and cognitive psychology. It has important theoretical research significance and practical application value in too many fields, such as information push, medical research, finance, and information security. In the past few years, there were a lot of well studied work in this field, but we are still facing various issues. In this paper, we clearly review the history of deep learning interpretability research and related work. Firstly, we introduce the history of interpretable deep learning from following three aspects: origin of interpretable deep learning, research exploration stage and model construction stage. Then, the research situation is presented from three aspects, namely visual analysis, robust perturbation analysis and sensitivity analysis. The research on the construction of interpretable deep learning model is introduced following four aspects: model agent, logical reasoning, network node association analysis and traditional machine learning model. Moreover, the limitations of current research are analyzed and discussed in this paper. At last, we list the typical applications of the interpretable deep learning and forecast the possible future research directions of this field along with reasonable and suitable suggestions.

Key words artificial intelligence; deep learning; interpretability; neural network; visualization

摘 要 深度学习的可解释性研究是人工智能、机器学习、认知心理学、逻辑学等众多学科的交叉研究课题,其在信息推送、医疗研究、金融、信息安全等领域具有重要的理论研究意义和实际应用价值.从深度学习可解释性研究起源、研究探索期、模型构建期3方面回顾了深度学习可解释性研究历史,从可视化分析、鲁棒性扰动分析、敏感性分析3方面展现了深度学习现有模型可解释性分析研究现状,从模型代理、逻辑推理、网络节点关联分析、传统机器学习模型改进4方面剖析了可解释性深度学习模型构建研究,同时对当前该领域研究存在的不足作出了分析,展示了可解释性深度学习的典型应用,并对未来可能的研究方向作出了展望.

收稿日期:2019-07-10;**修回日期**:2019-11-13
基金项目:国家自然科学基金项目(61972183,61672268);社会安全风险感知与防控大数据应用国家工程实验室主任基金项目
This work was supported by the National Natural Science Foundation of China (61972183, 61672268) and the Director Foundation Project of National Engineering Laboratory for Public Safety Risk Perception and Control by the Big Data.

关键词 人工智能;深度学习;可解释性;神经网络;可视化

中图法分类号 TP391

随着大型数据库的可用性和深度学习方法的不断改进,人工智能系统在越来越多复杂任务上的性能已经达到甚至超过了人类的水平.目前,基于深度学习算法的系统已经广泛应用于图像分类[1]、情绪分析[2]、语音理解[3]等领域,实现了代替人工作出决策的过程.然而,尽管这些算法在大部分的任务中发挥着卓越的表现,但由于产生的结果难以解释,有些情况下甚至不可控[4].与此同时,如果一个模型完全不可解释,那么其在众多领域的应用就会因为无法展现更多可靠的信息而受到限制.

从用户的角度而言,深度学习系统不仅需要向用户展现推荐的结果,还需要向用户解释推荐的原因.如在新闻推送的应用[5]方面,针对不同的用户群体,需要推荐不同类型的新闻,满足他们的需求.此时不仅要向用户提供推荐的新闻,还要让用户知道推荐这些新闻的意义.因为一旦用户认为推荐的内容不够精准,那么他们就会认为深度学习系统在某些方面存在偏差.在经济学方面,对于股价的预测以及楼市的预测深度学习有可能会表现得更好.但是由于深度学习的不可解释以及不安全性,应用中可能会更偏向于使用传统可被解释的机器学习.

从系统开发人员的角度来说,深度学习一直以来是作为一个黑盒在实验室的研究过程中被直接使用的,大多数情况下其确实可以取得一些良好的结果.而且,通常情况下,深度学习网络的结果比传统机器学习的结果更精准.但是,关于如何获得这些结果的原因以及如何确定使结果更好的参数问题并未给出解释.同时,当结果出现误差的时候,也无法解释为什么会产生误差、怎么去解决这个误差.如耶鲁大学科研人员曾尝试使用基于深度学习的 AI 进行程序 Debug,导致 AI 直接将数据库删除的结果.

从监管机构立场来看,监管机构更迫切希望作为技术革命推动力的深度学习具有可解释性.2017 年监督全球金融稳定委员会(Financial Stability Board)称,金融部门对不透明模型(如深度学习技术)的广泛应用可能导致的缺乏解释和可审计性表示担忧,因为这可能导致宏观层级的风险[6].该委员会于 2017 年底发布了一份报告,强调 AI 的进展必须伴随对算法输出和决策的解释.这不仅是风险管理的重要要求,也是建立公众及金融服务监管机构更大信任的重要条件.

现今,随着深度学习广泛而深入的应用,其可解释的重要性越发突显,如基于深度学习的医疗诊断由于不可解释性无法获知其判断依据从而无法可信使用、司法量刑风险得分因为其不可解释而发生偏差造成判断错误、无人驾驶造成车祸却因为系统不可解释而难以分析其原因等.

由此可见,深度学习的可解释性研究意义重大,其可以为人们提供额外的信息和信心,使其可以明智而果断地行动,提供一种控制感;同时使得智能系统的所有者能够清楚地知道系统的行为和边界,人们可以清晰看到每一个决策背后的逻辑推理,提供一种安全感;此外,也可监控由于训练数据偏差导致的道德问题和违规行为,能够提供更好的机制来遵循组织内的问责要求,以进行审计和其他目的.

当前,学术界和工业界普遍认识到深度学习可解释性的重要性,《Nature》《Science》《MIT Technology Review》近来都有专题文章讨论这一问题,AAAI 2019 设置了可解释性人工智能讨论专题,David Gunning 则领导了美国军方 DAPRA 可解释 AI 项目,试图建设一套全新且具有可解释性的深度学习模型.

本文将对深度学习可解释性研究的源起、发展历史进行分析,并从深度学习的可解释性分析和构建可解释性深度模型 2 个方面对现有研究方向进行归纳总结,同时对可解释深度学习未来的发展作出展望.

1 深度学习可解释性研究历史

1.1 深度学习可解释性研究起源

近年来,随着深度学习应用领域的不断拓展,作为制约深度学习应用的瓶颈,可解释性问题越来越受到研究者的重视.

早在 1982 年 Fukushima 等人开发了一种名为 Neocognitron 的人工神经网络[7],该网络采用分层的多层设计允许计算机“学习”识别视觉模式.经过多层重复激活的强化策略训练使其性能逐渐增强,由于层数较少,学习内容固定,具有最初步的可解释性能,并且可以看作是深度学习可视化的开端之作.1991 年 Garson[8] 提出了基于统计结果的敏感性分析方法,从机器学习模型的结果对模型进行分析,试图

得到模型的可解释性.早期的研究,启迪了后来研究者的思路.自此,越来越多的研究者加入到了深度学习可解释性研究,研究进入了蓬勃的发展期.

1.2 深度学习可解释性研究探索期

在这一阶段,研究者们从实验和理论2方面都进行了探索研究,研究取得了显著进展.

在实验研究方面,主要包括深度学习模型内部隐层可视化和敏感性分析等实验.

1) 在内部隐层可视化实验方面.2012年Google研究人员在基于TensorFlow的深度模型可视分析工作中,将人的视觉感知能力和深度学习算法的计算能力相结合,对深度学习的可解释性进行探索和分析[9];2014年Zeiler等人[10]介绍了一种新颖的CNN隐层可视化技术,通过特征可视化查看精度变化,从而知道CNN学习到的特征是怎样的,深入了解中间特征层的功能和分类器的操作.

2) 在敏感性分析实验方面的代表性工作则包括:

① 基于连接权的敏感性分析实验.如1991年Garson[8]提出通过度量输入变量对输出变量的影响程度的方法.

② 基于统计方法的敏感性分析法.如2002年Olden等人[11]通过大量的重复采样、随机打乱输出值,得到基于给定初始值随机训练网络的权重和重要性分布,通过统计检验的方法来进行敏感性分析.

③ 基于样本影响力的敏感性分析实验.如2017年Koh等人[12]通过影响力函数来理解深度学习黑盒模型的预测效果,即将样本做微小的改变,并将参数改变对样本微小改变的导数作为该样本的影响力函数[13].

除了实验方面,研究者们也对深度学习可解释性理论进行了探索性研究:

2018年Lipton[14]首次从可信任性、因果关联性、迁移学习性、信息提供性这4个方面分析了深度学习模型中可解释性的内涵,指出"可解释的深度学习模型作出的决策往往会获得更高的信任,甚至当训练的模型与实际情况发生分歧时,人们仍可对其保持信任;可解释性可以帮助人类理解深度学习系统的特性,推断系统内部的变量关系;可解释性可以帮助深度学习模型轻松应对样本分布不一致性问题,实现模型的迁移学习;可解释性可为人们提供辅助信息,即使没有阐明模型的内部运作过程,可解释模型也可以为决策者提供判断依据."同时,作者指出构建的可解释深度学习模型至少应包含"透明性"和"因果关联性"的特点.

早期对深度学习可解释性的探索研究取得了丰富的成果,然而,基于黑盒模型进行解释始终存在解释结果精度不高、计算机语言难以理解等局限.所以,构建可解释性模型开始成为新的研究方向.

1.3 深度学习模型可解释性模型构建期

与对深度学习黑盒模型进行解释相比,直接构建的可解释性模型往往具有更强的可解释性.

2012年起,人们开始尝试引入知识信息以构建可解释的深度模型.主要的尝试方案有2种:1)将表示为知识图谱的离散化知识转换为连续化向量,从而使得知识信息的先验知识能够成为深度模型的输入;2)将知识信息中的逻辑规则作为深度学习优化目标的约束,指导模型的学习过程[15].2017年Hu等人[16]提出的teacher-network网络中,通过将深度神经网络与一阶逻辑规则相结合,显著提高了分类的效果,表现出良好的可解释性.

值得一提的是,Hinton等人提出的胶囊网络模型(CapsNet)[17],其在2017年发表的"Dynamic Routing Between Capsules"[18]一文中,详细介绍了CapsNet架构.由于该模型采用动态路由来确定神经网络的边权,这就从一定程度上提供了边权确定的解释性,更重要的是,与传统卷积神经网络(convolutional neural networks, CNN)相比,CapsNet具有用少量训练数据就能实现泛化的特点.同时,样本对象的细节特征,如对象的位置、旋转、厚度、倾斜度、尺寸等信息会在网络中被学习保存下来,不会被丢失.这些CapsNet所独具的优点展现了其作为成功的可解释深度学习模型的特点.

深度学习的可解释性研究从提出到展开,也就短短数年,但已取得了令人瞩目的诸多成果.为此我们有必要对相关研究现状及其成果做一个系统的梳理.

2 深度学习可解释性研究现状

可解释性研究目前主要存在2方面研究方法:1)从深度学习模型本身进行入手,调整模型内部参数,对系统得到的结果进行分析,判断参数对于结果的影响.或是通过对输入变量添加扰动,探测表征向量来评估系统中不同变量的重要性,推测系统作出决策的依据.2)直接构建本身就具有可解释性的模型,旨在学习更结构化和可解释的模型.

2.1 深度学习模型可解释性分析研究

深度学习作为黑盒模型,对于输出结果往往无

法指出系统得到决策的依据.因此,通过对黑盒模型内部结构进行剖析,可以清晰地看到决策背后的逻辑推理,得到具有可解释性的结果.

2.1.1 基于可视化的可解释性研究

2012 年 Krizhevsky 等人[19]利用大型卷积网络模型在 ImageNet 基准数据集上的测试展示了令人印象深刻的分类性能.然而,并没有较为合理的方法说明模型表现出色的原因或者展示作出判断的依据.因此,Zeiler 等人[10]提出了一种新颖的 CNN 隐层可视化技术,从信息提供性方面入手,通过特征可视化,查看精度变化,从而知道 CNN 学习到怎样的特征.这些隐层可视化可应用于中间每一层结构,通过对隐层运用一些可视化方法来将其转化成人类可以理解的有实际含义的图像.方法利用反卷积的相关思想实现了隐层特征可视化来帮助理解 CNN 的每一层究竟学到了什么东西,从而能够找到优于 Donahue 等人[20]的模型架构,具体过程如图 1 所示:

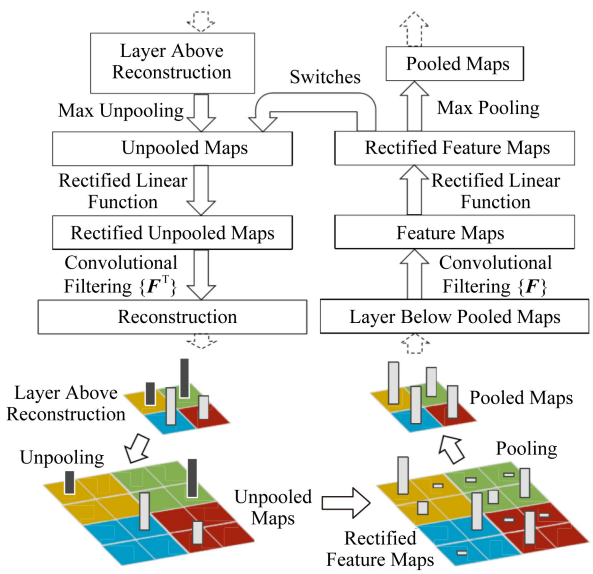

Fig. 1 Visualization of the hidden layer process

图 1 可视化隐层过程

2.1.2 基于鲁棒性扰动测试的可解释性研究

基于鲁棒性扰动的方法主要是通过对输入数据添加扰动元素[21].有些模型不能直接解释实现过程,但是可以对其他属性作出评估.例如通过对输入数据添加扰动元素,测试添加的特征是否为主要特征,是否会影响最后得出的结果.所以,解释这些黑盒模型的工作普遍集中在理解固定模型如何导致特定预测.例如通过在测试点周围局部拟合更简单的模型[22]或通过扰乱测试点来了解模型预测的变化[23-25].其中比较有代表性的工作是 Koh 等人[12]在 2017 年提出的通过影响力函数来理解深度学习黑盒模型的预测效果.通过学习算法跟踪模型[26]的预测并返回其训练数据,从而识别对给定预测负责的训练点[27].即使在理论失效的非凸和非可微模型上,影响函数的近似仍然可以提供有价值的信息.在线性模型和卷积神经网络上,由于计算出了对训练样本施加轻微扰动之后对特定测试样本损失函数的影响,所以这个方法也可以应用到对抗样本的生成中[28],只需要在一部分影响力函数较大的样本中添加一些微小的扰动,就足以干扰其他样本的判定结果.文章证明影响函数可用于理解模型行为[29]、调试模型、检测数据集错误,甚至创建视觉上可区分的训练集攻击等多个任务.

在判断扰动对模型结果产生的影响方面,Fisher 等人[30]提出了模型分类依赖性(model class reliance, MCR)方法,通过提取模型的重要特征来对模型进行解释.变量重要性(variable importance, VI)工具用于描述协变量在多大程度上会影响预测模型的准确率[31].通常情况下,在一个模型中重要的变量在另一个模型中却不是那么重要,或者说,一个分析人员使用的模型依赖的协变量信息可能与另一个分析人员使用的协变量完全不同.通过设定模型 VI 的上下限,根据变量的扰动对模型判定结果的影响来判断模型对于变量的依赖程度[32-33].方法通过对 COMPAS 犯罪模型中性别和种族等变量的变化进行度量,解释模型对于不同变量的依赖程度.

2.1.3 基于敏感性分析的可解释性研究

敏感性分析(sensitivity analysis, SA)是一类非常重要的,用于定量描述模型输入变量对输出变量的重要性程度的方法,在经济、生态、化学、控制等领域都已经有了非常成熟的应用.其基本思想就是令每个属性在可能的范围内变动,研究和预测这些属性的变化对模型输出值的影响程度.典型的敏感性分析方法有基于连接权、基于统计和基于扰动分析 3 类.

在基于连接权的方法中比较有代表性的工作是 Garson[8]提出的通过设置权重来估测输入变量对输出变量的影响程度.然而这种方法放到深度网络中由于忽略了非线性激活函数误差会一步一步积累[34],所以逐渐不再被使用.

在基于统计方法的敏感性分析方法中有代表性的是 Olden 等人[11]提出的使用随机初始的权重构建一组神经网络并记录其中预测性能最好的神经网络的初始权重[35],通过大量重复采样得到基于给定

初始值随机训练网络的权重和重要性分布,从而判断不同变量对于模型的影响程度,可以被看作是在基于统计方法的敏感性分析方法中的代表作.

在评估输入样本扰动敏感性方面,Hunter 等人[36]开发一种新的扰动流形模型及其对应的影响程度测量方法,以评估各种扰动对输入样本或者网络可训练参数的敏感性影响.这种方法是对使用信息几何解决分类问题的局部影响测量方法的全新扩展.它的贡献在于其度量方法是由信息几何驱动的,可以直接进行计算而不需要优化任何目标函数.并且该方法提出的敏感性影响测量适用于各种形式的外部和内部扰动,可用于 4 个重要的模型构建任务:检测潜在的"异常值"、分析模型架构的敏感性、比较训练集合测试集之间的网络敏感性以及定位脆弱区域.

2.2 可解释性深度学习模型构建研究

深度学习模型可解释性分析的研究只是尝试通过一定的技术手段去分析和解释深度学习模型,犹如管中窥豹、盲人摸象,所以另一些研究者试图直接创建具有可解释性的深度学习模型,使其对数据处理的过程、表示或其他方面更易于人们理解.

2.2.1 基于模型代理的可解释性建模

常用的深度网络通常使用大量的基本操作来推导它们的决策.因此,解释这种处理所面临的基本问题是找到降低所有这些操作的复杂性的方法,或是将已有的深度学习系统学习另外的可解释的系统,以此提高可解释性,代理模型法就是这样一类方法.

Ribeiro[37]的局部可理解的与模型无关的解释技术(local interpretable model-agnostic explanation, LIME)即为一种代理模型方法.该方法首先通过探测输入扰动获得深度模型的响应反馈数据,然后凭此数据构建局部线性模型[38],并将该模型用作特定输入值深度模型的简化代理.Ribeiro 表示,该方法可作用于识别对各种类型的模型和问题域的决策影响最大的输入区域.LIME 这样的代理模型可以根据其对原始系统的吻合程度来运行和评估.代理模型也可以根据其模型复杂度来测量,例如 LIME 模型中的非零维度的数量.因为代理模型在复杂性和可靠性之间提供了可量化的关系,所以方法可以相互对照,使得这种方法很具有吸引力.

另一种代理方法是 Carnegie Mellon 大学 Hu 等人[39]提出的反复蒸馏方法,该方法体现了可解释方法中的迁移学习性,通过将逻辑规则的结构化信息转移到神经网络的权值[40]中.网络包括教师网络(teacher network)和学生网络(student network)两个部分,教师网络负责将逻辑规则所代表的知识建模,学生网络利用反向传播方法加上教师网络的约束,迫使网络模拟一个规则化教师的预测,并且在训练过程中迭代地演进[41].这 2 个模型.教师网络在每次训练迭代中都要构建,也就是说训练过程中教师网络和学生网络是一个联合训练[42]的过程.它将深度神经网络与一阶逻辑规则[43]相结合,将逻辑规则整合到神经模型中,将逻辑规则的结构化信息转换为神经网络的权重.通过使用后验正则化原理[44]构建的教师网络完成这种规则信息的转移,具体过程如图 2 所示.方法能够在具有高精度分类效果的同时,又能体现逻辑规则的解释性,该方法可用于 CNN 和循环神经网络(recurrent neural network, RNN)等不同网络上,用于情感分析、命名实体识别,在深度神经网络模型的基础上实现效果提升.

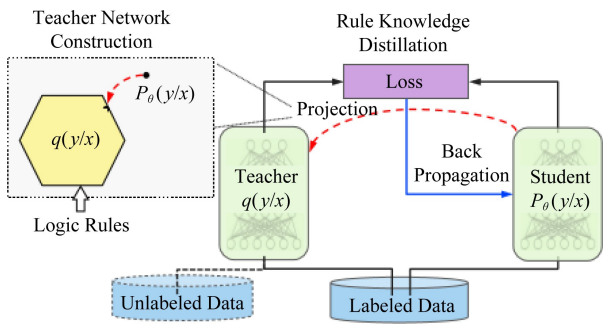

Fig. 2 Teacher-Student network

图 2 教师-学生网络

2.2.2 基于逻辑推理的可解释性建模

由于逻辑推理[45]能够很好地展现系统的可解释性[46],并且逻辑推理体现了可解释方法中的因果关联性.Garcez 等人[47]提出了一种面向连接性论证的网络框架,它允许推理和学习论证推理.方法使用神经符号学习系统将论证网络转换成标准的神经网络,实现了基于权重的论证框架,具有学习能力[48].文中将算法分为正面论点和反面论点,2 种论点都被设置在了论证网络中,通过学习进行累积论证,随着时间的推移,某些论证将会加强,某些论证将会削弱,论证结果可能会发生变化,其展现出该网络的学习过程.

此外,Yao 等人[49]提出了另一种新颖的推理模型,该模型通过深度强化学习来激活逻辑规则.模型采用记忆网络的形式,存储关系元组,模仿人类认知活动中的"图像模式".方法将推理定义为修改或从内存中恢复的顺序决策,其中逻辑规则用作状态转

换函数.模型在使用不到1000的训练样本且没有支持事实的情况下,实现了在文本数据集bAbI-20上仅0.7%的平均错误率.

2.2.3 基于网络节点关联分析的可解释性建模

2017年Hinton等人[18]提出了一种被称为"胶囊"的新型神经单元.胶囊网络极大地体现了可解释方法中的因果关联性特点,它改进了传统的CNNs网络,胶囊网络中神经元节点间的权重路由关系可以解释检测到的特征之间的空间关系.每一组神经元组成一个胶囊,通过每一个胶囊中的神经元的活动向量(activity vector)来表示实体类型的实例化参数.活动向量的长度表示实体出现的概率,方向表示实例化的参数.活跃的低层胶囊预测结果通过转移矩阵发送到相邻活跃度相对较高的胶囊之中.当多个预测信息一致时,胶囊将被激活.该方法使用协议路由机制,该机制会为那些能更好拟合高层胶囊的实例化参数的低层胶囊分配更高权重.其中,协议路由机制使得每个胶囊能够编码一个特定语义的概念,可以清晰地知道每一个"胶囊"所做的工作.在一定程度上,胶囊网络可以看作是一种特定语义化的网络结构,从而使得构建的胶囊网络成为了一种具有能够解释并识别对象空间结构信息的可解释模型[50].在MNIST上的实验表明,使用胶囊网络,能够有效甄别重叠的不同数字.

2.2.4 基于传统机器学习的可解释性建模

已有的深度学习系统具有强大的预测能力,结果精准但缺乏可解释性;传统的机器学习系统结构较为简单,预测精度不如深度学习系统,但往往具备可解释性.所以利用传统可解释机器学习方法构建可解释深度学习模型,成为了一种新的尝试方向.

以决策树为例,众所周知决策树具有较好的可解释性.自20世纪90年代起,便有研究者将决策树与多层神经网络相联系进行研究.该工作主要利用决策树的可解释性的优点对神经网络决策进行过程简化,使深度学习网络具有信息提供性的特征.方法之一是基于决策树的深度神经网络规则提取器(DeepRED)[51],它将为浅层网络设计的基于决策树的连续规则提取器(CRED)[52]算法扩展到任意多个隐层,并使用神经网络逆向提取规则方式(RxREN)[53]来修剪不必要的输入.然而,尽管DeepRED能够构建完全忠实于原始网络的树,但生成的树可能非常大,并且该方法的实现需要大量的时间和内存,因此在可伸缩性方面受到限制.为解决此问题,2018年南京大学周志华等人[54]提出了一种全新的深度学习方法"gcForest"(multi-grained cascade forest).该方法采用一种深度树集成方法(deep forest ensemble method),使用级联结构让gcForest做表征学习.需要指出,由于模型的构建是基于可解释的决策树,gcForest的超参数比一般深度神经网络少得多并且其可解释性强、鲁棒性高.因此,在大多数情况下,即使遇到不同领域的不同数据,也能取得很好的结果.同时,gcForest所需的训练数据集较小,这不仅使gcForest训练起来很容易,也使其可解释性理论分析更为简单.

总体而言,目前可解释性深度学习模型的构建可以从可信任性、因果关联性、迁移学习性、信息提供性4个方面对其进行分析.可信任性是具有可解释性深度学习模型的基础,其可以为人们提供额外的信息和信心,使人们可以明智而果断地行动.使得智能系统的所有者清楚地知道系统的行为和边界,人们可以清晰地看到每一个决策背后的逻辑推理,提供一种安全感,使得深度学习模型更好地服务于大众.因果关联性主要从逻辑推理和特征关联2方面体现,例如面向连接性的网络框架与Hinton提出的胶囊网络.迁移学习性主要通过将结构化信息转移到神经网络的权值中,使神经网络具有可解释性.信息提供性主要是使模型向人们提供可以被理解的知识,主要包括与传统机器学习相结合的深度学习模型或是深度学习模型的可视化等方法.

3 可解释性深度学习的应用

当前,深度学习应用广泛,但在某些特定领域,由于深度学习模型的不可解释性限制了深度学习模型的应用.随着深度学习可解释性研究的深入,特别是具有可解释性深度学习模型的建立,越来越多的关系到重大生产活动、人类生命安全的关键领域也开始能够放心接受深度学习所带来的红利.

在推荐系统方面,新加坡国立大学Catherine等人[55]提出知识感知路径递归网络(KPRN),对用户和物品之间的交互特征在知识图谱中存在的关联路径进行建模,为用户提供可解释性推荐.在基于外部知识的基础上,Wang等人[56]又提出基于翻译的推荐模型,利用共同学习推荐系统和知识图谱补全模型,提高推荐的解释性.加州大学圣地亚哥分校Wang等人[57]借鉴混合专家模型(mixtures-of-experts)的思想提出了一种全新的深度学习推荐系统框架,利用用户序列行为中相邻物品间的关系来

解释用户在特定时间点的行为原因,进而基于用户的近期行为对其下一次行为进行预测,实现对用户群体的精准推送.

在社区安全方面,可解释深度学习应用于犯罪风险评估,可根据罪犯的受教育程度、前科、年龄、种族等一系列参数判断再次犯罪的概率,对社会管理起到协助作用.Bogomolov 等人[58]采用图形卷积神经网络来对毒品进行检测,对输入向量与网络中的神经元关系进行解释,通过训练完成的图形卷积神经网络的测试结果给出结论,并根据药效团特征来证明他们的结论.

在医疗方面,Luo 等人[59]从重症监护中的多参数智能监测(MIMICII)数据集中提取了特征,使用局部可解释的模型不可知解释(LIME)技术,实现了对难以解释的复杂 RF 模型决策过程中重要特征的简单解释.这些解释符合当前的医学理解,并且推动了基于深度学习医学诊断的发展进程.Zhang 等人[60]提出了 AuDNNsynergy 深度学习模型来进行药物组合克服耐药性,通过整合多组样本学习数据和化学结构数据来预测药物组合产生新药物,并对其中的深度模型的学习过程进行解释分析.

综上,深度学习已广泛应用于推荐系统、医疗、安全等各个领域,而深度学习良好的表现也使其成为这些领域不可或缺的工具.可解释性深度学习的出现,将显著提高系统的可靠性,使其可知、可控、可被人们信任,在更多的领域发挥更大的作用.

4 总结与展望

以深度学习为代表的各种机器学习技术方兴未艾,取得了举世瞩目的成功.机器和人类在很多复杂认知任务上的表现已经不分伯仲.然而,在解释模型为什么奏效及如何运作方面,目前学界的研究还处于非常初级的阶段.从当前研究现状看,研究者们普遍意识到深度学习可解释性的重要性,并已展开了诸多十分有意义的研究.但目前对深度学习可解释性的研究尚处于起步阶段,对于可解释性的本质、研究手段认识都还未能形成统一认识和找到最佳方案,未来可解释性深度学习领域的研究将会持续火热下去.基于对当前研究实践的分析和理解,我们认为未来深度学习的可解释性研究将可从 4 个方面着手深入:

1) 嵌入外部人类知识.目前,大多数深度学习模型使用数据驱动的方法,而较少关注知识驱动的观点.因此,将人类知识,如以知识图谱形式与深度学习技术相结合构建具有解释性的深度学习模型,可以作为一个研究方向.此外,可以利用可视化分析直观地验证模型是否正确遵循人类嵌入的知识和规则,以确保深度学习按照人类的意愿进行工作.

2) 深度学习的渐进式视觉分析.大多数现有可解释的深度学习方法主要侧重于在模型训练完成后进行理解和分析模型预测,但由于许多深度学习模型的训练非常耗时,因此迫切需要使用渐进的可视化分析技术,在保证模型准确率的情况下,同步进行可视化分析,保证模型的可解释性.这样不仅可以在模型训练过程中渐进式进行同步分析.专家可以利用交互式可视化来探检查新传入的结果并执行新一轮的探索性分析,而无需等待整个培训过程完成,并且保证了模型每一层的可解释性.

3) 提高深度学习的扰动可解释性.深度学习模型通常容易受到对抗性扰动的影响导致输出错误的预测.有时对抗性示例的修改非常轻微,以至于根本无法注意到修改,但模型仍然会出错.这些对抗性示例通常用于攻击深度学习模型.在这方面,保持深度学习模型的鲁棒性在实际应用中至关重要,当模型具有可解释性时,即使轻微的扰动人们也可以知道扰动变量对于模型的影响以及影响程度,并且可以在基于人类知识的情况下向人们进行解释.因此,关于可解释深度学习的一个研究机会是结合人类知识来提高深度学习模型的鲁棒性.

4) 以人为中心进行模型解释性升级.理想的深度学习可解释模型,应该能够根据受众背景知识的不同作出不同的解释,即以人为中心进行解释.同时,这种解释应是机器一边解决问题,一边给出答案背后的逻辑推理过程.面对这样的需求,未来深度学习可解释模型,其输出的整体可解释性将由各个多元的子可解释性组合而成,这对目前的机器学习从理论到算法都将是一个极大的挑战.

参 考 文 献

[1] Haralick R M, Shanmugam K, Dinstein I. Textural features for image classification [J]. IEEE Transactions on Systems Man & Cybernetics, 1973, 3(6): 610-621

[2] Zhang Ying, Wang Chao, Guo Wenya, et al. Multi-source emotion tagging for online news comments using bi-directional hierarchical semantic representation model [J]. Journal of Computer Research and Development, 2018, 55(5): 933-944 (in Chinese)

(张莹, 王超, 郭文雅, 等. 基于双向分层语义模型的多源新闻评论情绪预测[J]. 计算机研究与发展, 2018, 55(5): 933-944)

[3] Xin Yu, Yang Jing, Tang Chuheng, et al. An overlapping semantic community detection algorithm based on local semantic cluster [J]. Journal of Computer Research and Development, 2015, 52(7): 1510-1521 (in Chinese)
(辛宇, 杨静, 汤楚蘅, 等. 基于局部语义聚类的语义重叠社区发现算法[J]. 计算机研究与发展, 2015, 52(7): 1510-1521)

[4] Xiong Hongkai, Gao Xing, Li Shaohui, et al. Interpretable, structured and multimodal deep neural networks [J]. Pattern Recognition and Artificial Intelligence, 2018, 31(1): 1-11 (in Chinese)
(熊红凯, 高星, 李劭辉, 等. 可解释化、结构化、多模态化的深度神经网络[J]. 模式识别与人工智能, 2018, 31(1): 1-11)

[5] Koenigstein N, Dror G, Koren Y. Yahoo! music recommendations: Modeling music ratings with temporal dynamics and item taxonomy [C] //Proc of the 5th ACM Conf on Recommender Systems. New York: ACM, 2011: 165-172

[6] Heaton J B, Polson N G, Witte J H. Deep learning for finance: Deep portfolios [J]. Applied Stochastic Models in Business and Industry, 2017, 33(1): 3-12

[7] Fukushima K, Miyake S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition [M] //Competition and Cooperation in Neural Nets. Berlin:Springer, 1982: 267-285

[8] Garson G D. Interpreting neural-network connection weights [J]. AI Expert, 1991, 6(4): 46-51

[9] Wu Fei, Liao Binbing, Han Yahong. Interpretability of deep learning [J]. Aviation Weapons, 2019, 26(1): 39-46 (in Chinese)
(吴飞, 廖彬兵, 韩亚洪. 深度学习的可解释性[J]. 航空兵器, 2019, 26(1): 39-46)

[10] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2014: 818-833

[11] Olden J D, Jackson D A. Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks [J]. Ecological Modelling, 2002, 154(1/2): 135-150

[12] Koh P W, Liang Percy. Understanding black-box predictions via influence functions [C] //Proc of the 34th Int Conf on Machine Learning. Cambridge, MA: MIT Press, 2017: 1885-1894

[13] Adler P, Falk C, Friedler S A, et al. Auditing black-box models for indirect influence [C] //Proc of 2016 IEEE 16th Int Conf on Data Mining (ICDM). Volume 54. Piscataway, NJ: IEEE, 2016: 95-122

[14] Lipton Z C. The mythos of model interpretability [J]. ACM Queue, 2018, 61(10): 96-100

[15] Steiner T, Verborgh R, Troncy R, et al. Adding realtime coverage to the Google knowledge graph [C] //Proc of Int Semantic Web Conf. Berlin: Springer, 2012: 65-68

[16] Hu Zhiting, Yang Zichao, Liang Xiaodan, et al. Toward controlled generation of text [C] //Proc of the 34th Int Conf on Machine Learning. Cambridge, MA: MIT Press, 2017: 2503-2513

[17] Sabour S, Frosst N, Hinton G E. Matrix capsules with EM routing [C] //Proc of the 6th Int Conf on Learning Representations. La Jolla, CA: ICLR, 2018: 1884-2020

[18] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules [C] //Proc of the 31st Annual Conf on Neural Information Processing Systems. Long Beach, CA: NIPS, 2017: 3857-3867

[19] Krizhevsky A, Sutskever I, Hinton G E, et al. ImageNet classification with deep convolutional neural networks [J]. Neural Information Processing Systems, 2012, 141(5): 1097-1105

[20] Donahue J, Jia Y, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2014: 647-655

[21] Donahue J, Jia Yangqing, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2014: 647-655

[22] Yao Xin. Evolving artificial neural networks [J]. Proceedings of the IEEE, 1999, 87(9): 1423-144

[23] Bertsimas D, Sim M. Robust discrete optimization and network flows [J]. Mathematical Programming, 2003, 98(1/2/3): 49-71

[24] Eitel A, Springenberg J T, Spinello L, et al. Multimodal deep learning for robust RGB-D object recognition [C] //Proc of Int Conf on Intelligent Robots and Systems (IROS). Piscataway, NJ: IEEE, 2015: 681-687

[25] Milanese M, Tempo R. Optimal algorithms theory for robust estimation and prediction [J]. IEEE Transactions on Automatic Control, 1985, 30(8): 730-738

[26] Reitmayr G, Drummond T. Going out: Robust model-based tracking for outdoor augmented reality [C] //Proc of IEEE/ACM Int Symp on Mixed & Augmented Reality. New York: ACM, 2006: 109-118

[27] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences [C] //Proc of the 34th Int Conf on Machine Learning. Cambridge, MA: MIT Press, 2017: 3145-3153

[28] Ray R B, Ray R, Hepatitis C. Virus core protein: Intriguing properties and functional relevance [J]. FEMS Microbiology Letters, 2001, 202(2): 149-156

[29] Wojnowicz M, Cruz B, Zhao Xun, et al. Influence sketching: Finding influential samples in large-scale regressions [C] // Proc of IEEE Int Conf on Big Data. Piscataway, NJ: IEEE, 2016: 3601-3612

[30] Fisher A, Cynthia R, Francesca D. All models are wrong but many are useful: Variable importance for blackbox, proprietary, or misspecified prediction models, using model class reliance [J]. Journal of Machine Learning Research, 2019, 20(177): 1-81

[31] Strobl C, Boulesteix A L, Kneib T, et al. Conditional variable importance for random forests [J]. BMC Bioinformatics, 2008, 9(1): 296-307

[32] Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests [J]. Statistics and Computing, 2017, 27(3): 659-678

[33] Datta A, Sen Shayak, Zick Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems [C] //Proc of IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2016: 598-617

[34] Anas A, Liu Yu. A regional economy, land use, and transportation model: Formulation, algorithm design, and testing [J]. Journal of Regional Science, 2007, 47(3): 415-455

[35] Shet V, Singh M, Bahlmann C, et al. Predicate logic based image grammars for complex pattern [J]. International Journal of Computer Vision, 2011, 93(2): 141-161

[36] Hunter A, Kennedy L, Henry J, et al. Application of neural networks and sensitivity analysis to improved prediction of trauma survival [J]. Computer Methods and Programs in Biomedicine, 2000, 62(1): 11-19

[37] Ribeiro B. Prediction of the LIME availability on an industrial kiln by neural networks [C] //Proc of IEEE Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 1998: 1987-1991

[38] Ribeiro M T, Sameer S, Carlos G. Model-agnostic interpretability of machine learning [J]. arXiv preprint arXiv:1606.05386, 2016

[39] Hu Zhiting, Ma Xuezhe, Liu zhengzhong, et al. Harnessing deep neural networks with logic rules [C] //Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 2410-2420

[40] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 427-436

[41] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models [J]. arXiv preprint, arXiv:1401.4082, 2014

[42] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [J]. arXiv preprint, arXiv: 1312.6199, 2013

[43] Muggleton S, King R D, Stenberg M J E. Protein secondary structure prediction using logic-based machine learning [J]. Protein Engineering, Design and Selection, 1992, 5(7): 647-657

[44] Protheroe R J, Biermann P L. A new estimate of the extragalactic radio background and implications for ultra-high-energy gamma-ray propagation [J]. Astroparticle Physics, 1996, 6(1): 45-54

[45] Cangelosi D, Blengio F, Versteeg R, et al. Logic learning machine creates explicit and stable rules stratifying neuroblastoma patients [J]. BMC Bioinformatics, 2013, 14(7): 1-20

[46] Tran S N, Garcez A S A. Deep logic networks: Inserting and extracting knowledge from deep belief networks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 29(2): 246-258

[47] Garcez A S D, Gabbay D M, Lamb L C. Towards a connectionist argumentation framework [C] //Proc of the 16th European Conf on Artificial Intelligence. Amsterdam: IOS Press, 2004: 987-988

[48] Garcez A S D, Gabbay D M, Lamb L C. Value-based argumentation frameworks as neural-symbolic learning systems [J]. Journal of Logic and Computation, 2005, 15(6): 1041-1058

[49] Yao Yiqun, Xu Jiaming, Shi Jing, et al. Learning to activate logic rules for textual reasoning [J]. Neural Networks, 2018, 106(6): 42-49

[50] Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems [J]. arXiv preprint, arXiv:1603.04467, 2016

[51] Zilke J R, Mencia E L, Janssen F. DeepRED-Rule extraction from deep neural networks [C] //Proc of Int Conf on Discovery Science. Berlin: Springer, 2016: 457-473

[52] Sato M, Tsukimoto H. Rule extraction from neural networks via decision tree induction [C] //Proc of Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2001: 1870-1875

[53] Augasta M G, Kathirvalavakumar T. Reverse engineering the neural networks for rule extraction in classification problems [J]. Neural Processing Letters, 2012, 35(2): 131-150

[54] Feng Ji, Zhou Zhihua. Autoencoder by forest [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence.Palo Alto,CA: AAAI, 2018: 1-15

[55] Catherine R, Cohen W. Personalized recommendations using knowledge graphs: A probabilistic logic programming approach [C] //Proc of the 10th ACM Conf on Recommender Systems. New York: ACM, 2016: 325-332

[56] Cao Yi, Wang Xiang, He Xiangnan, et al. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences [C] //Proc of the World Wide Web Conf. New York: ACM, 2019: 151-161

[57] Wang Chengkang, Wan Mengting, McAuley J. Recommendation through mixtures of heterogeneous item relationships [C] // Proc of the 27th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2018: 1143-1152

[58] Bogomolov A, Lepri B, Staiano J, et al. Once upon a crime: Towards crime prediction from demographics and mobile data [C] //Proc of the 16th Int Conf on Multimodal Interaction. New York: ACM, 2014: 427-434

[59] Luo Yuan, Ahmad F S, Shah S J, et al. Tensor factorization for precision medicine in heart failure with preserved ejection fraction [J]. Journal of Cardiovascular Translational Research, 2017, 10(3): 305-312

[60] Wen Ming, Zhang Zhimin. Deep-learning-based drug-target interaction prediction [J]. Journal of Proteome Research, 2017, 16(4): 1401-1409

Cheng Keyang, born in 1982. PhD, associate professor. Member of CCF. His main research interests include computer vision and machine learning.

Wang Ning, born in 1996. Master candidate at Jiangsu University. His main research interests include deep learning and convolutional neural network.

Shi Wenxi, born in 1987. PhD, senior engineer. His main research interests include video surveillance and city security.

Zhan Yongzhao, born in 1962. PhD, professor. Senior member of CCF. His main research interests include artificial intelligence and pattern recognition.