Journal of Computer Applications
计算机应用, 2022, 42(11): 3588 - 3602
ISSN 1001-9081
CODEN JYIIDU
2022-11-10
http://www.joca.cn

文章编号:1001-9081(2022)11-3588-15
DOI:10.11772/j.issn.1001-9081.2021122118

深度学习可解释性研究综述

雷　霞,罗雄麟*
(中国石油大学(北京)信息科学与工程学院,北京 102249)
(*通信作者电子邮箱luoxl@cup.edu.cn)

摘　要:随着深度学习的广泛应用,人类越来越依赖于大量采用深度学习技术的复杂系统,然而,深度学习模型的黑盒特性对其在关键任务应用中的使用提出了挑战,引发了道德和法律方面的担忧,因此,使深度学习模型具有可解释性是使它们令人信服首先要解决的问题。于是,关于可解释的人工智能领域的研究应运而生,主要集中于向人类观察者明确解释模型的决策或行为。对深度学习可解释性的研究现状进行综述,为进一步深入研究建立更高效且具有可解释性的深度学习模型确立良好的基础。首先,对深度学习可解释性进行了概述,阐明可解释性研究的需求和定义;然后,从解释深度学习模型的逻辑规则、决策归因和内部结构表示这三个方面出发介绍了几种可解释性研究的典型模型和算法,另外还指出了三种常见的内置可解释模型的构建方法;最后,简单介绍了忠实度、准确性、鲁棒性和可理解性这四种评价指标,并讨论了深度学习可解释性未来可能的发展方向。

关键词:深度学习;可解释性;决策归因;隐层表示;评价指标
中图分类号:TP18　　**文献标志码**:A

Review on interpretability of deep learning

LEI Xia, LUO Xionglin*
(*College of Information Science and Engineering*, *China University of Petroleum*, *Beijing* 102249, *China*)

Abstract: With the widespread application of deep learning, human beings are increasingly relying on a large number of complex systems that adopt deep learning techniques. However, the black-box property of deep learning models offers challenges to the use of these models in mission-critical applications and raises ethical and legal concerns. Therefore, making deep learning models interpretable is the first problem to be solved to make them trustworthy. As a result, researches in the field of interpretable artificial intelligence have emerged. These researches mainly focus on explaining model decisions or behaviors explicitly to human observers. A review of interpretability for deep learning was performed to build a good foundation for further in-depth research and establishment of more efficient and interpretable deep learning models. Firstly, the interpretability of deep learning was outlined, the requirements and definitions of interpretability research were clarified. Then, several typical models and algorithms of interpretability research were introduced from the three aspects of explaining the logic rules, decision attribution and internal structure representation of deep learning models. In addition, three common methods for constructing intrinsically interpretable models were pointed out. Finally, the four evaluation indicators of fidelity, accuracy, robustness and comprehensibility were introduced briefly, and the possible future development directions of deep learning interpretability were discussed.

Key words: deep learning; interpretability; decision attribution; latent representation; evaluation indicator

0　引言

近年来,基于深度学习模型的算法已逐步改变人类处理现实问题的方式,深度学习在社会和生活等各个领域的应用呈现高速增长的趋势。由于深度学习领域的研究,深度学习模型成功地应用在医疗[1-2]、自动驾驶[3-4]、图像处理分类和检测[5-6]、语音和音频处理[7-8]、网络安全[9-10]等现实生活的各种应用场景中,但是这种表现更多地依赖于模型复杂的体系结构和实验的调参技术,人们无法探知深度学习模型究竟从数据中学到了哪些知识,如何进行最终决策,以及缺乏完备的数学理论指导和改进深度学习模型的表达能力、训练能力和泛化能力[11-13]。

另外,深度学习模型的不可解释性存在很多的潜在危险,尤其在安全攻防领域[14-16]应用方面对可解释性的需求尤为明显。首先,不可解释性会降低模型的可信度,难以建立人与机器之间的信任;另一方面,也会带来难以解决的安全问题,作为一个具有大量参数的复杂模型,人们往往难以对深度学习模型的决策进行预判和解释。例如,即使一个深度

收稿日期:2021-12-18;**修回日期**:2022-02-12;**录用日期**:2022-02-23。　**基金项目**:国家自然科学基金资助项目(61703434)。
作者简介:雷霞(1989—),女,福建建瓯人,博士研究生,主要研究方向:机器学习、最优控制;　罗雄麟(1963—),男,湖南汨罗人,教授,博士,主要研究方向:控制理论、过程控制、化工系统工程、机器学习。

学习模型具有很好的性能,在物体识别任务上有很好的泛化能力,然而,Szegedy 等[17]发现通过对输入图像进行某种不可察觉的扰动就可以任意改变网络的预测,即对抗样本攻击。Nguyen 等[18] 提出 MAP-Elites (Multi-dimensional Archive of Phenotypic Elites)算法,采用训练好的、在 ImageNet 或 MNIST 数据集上有良好表现的卷积神经网络(Convolutional Neural Network,CNN),并利用演化算法的思想随机生成对于人类不可识别的图像,但深度学习模型以 99.99% 的可信度将其识别为特定物体。

因此,尽管深度学习模型可以在许多任务中取得优异的表现,但考虑到信任[19-21]、道德[22-25]、对人工智能(Artificial Intelligence, AI)的偏见[26-28],以及对抗性样本[29-32]在欺骗分类器决策的影响等问题,最近对深度学习可解释性的研究逐渐增多。为了提高人类对深度学习模型决策的信任度,促进决策过程的透明和公平,需要为深度学习模型提供一个可解释的解决方案。

鉴于深度学习可解释性研究的理论意义和重要的现实意义,本文对近年来深度学习可解释性的研究进展进行了系统性的综述,为进一步深入研究建立更高效且具有可解释性的深度学习模型确立良好的基础,图 1 给出了综述内容的全面概览图。

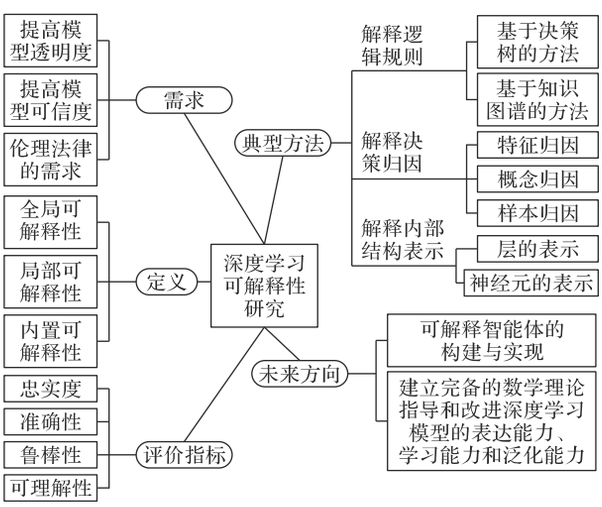

图 1　综述内容的概览

Fig. 1　Overview of survey

1　深度学习可解释性

1.1　可解释性研究的需求

随着深度学习模型在医疗保健、自动驾驶、信用评分和贷款评估等高风险领域的应用,除了关注模型的准确性之外,对深度学习模型可解释性的需求也越来越高,主要体现在以下三个方面:

1)提高模型透明度:深度学习模型的透明度是指模型所具有的表达能力和能够被人类理解的能力。透明度可以是算法本身的一部分,也可以使用外部手段,如使用代理模型进行解释提高透明度。利用黑箱模型给出最终决策让人们无法判定其公平性和合理性,因此通过对模型内部机制的理解提高其透明度是非常必要的。透明度对于评估模型预测的结果和分析模型受到对抗性样本攻击的原因有重要意义。

2)提高模型可信度:深度学习模型的可信度是对人类和终端用户在动态现实环境中对给定模型的预期工作的信心的衡量。尽管深度学习模型在一些测试集上表现出了良好的性能,但现实环境仍然要复杂得多,缺乏可靠决策依据的模型往往可能遭遇到失败,这对于一些要求高度可靠的预测系统来说可能会导致灾难性的结果。理解一个深度学习模型做出决策的原因和依据的决策特征,能让我们判断模型是否符合常理并分析模型发生错误的原因,对提高终端用户的信任度至关重要。因此,往往一个次优决策的具有可解释性的模型要比一个没有任何解释的高准确率模型要好。

3)伦理和法律的需求:考虑对深度学习模型做出解释以评估算法生成的决策是否符合道德和伦理的标准[33]有很重要的现实意义。比如,当深度学习模型应用于推荐系统时,保证推荐的内容符合道德和伦理的标准至关重要。文献[34]中提到法院应用深度学习模型来预测个人再次犯罪的可能性以决定谁该释放谁该拘留,这也引起了人们对道德的担忧。另外,为了保证预测模型不会因种族等其他因素而产生偏见,准确性不应该作为模型的唯一评价指标,公平性也同样至关重要,这也迫切地要求模型具有可解释性。另一方面,在欧盟的《通用数据保护条例》[35]也有提到,受算法决策影响的个人具有解释权。

1.2　可解释性的定义

由于不同研究者对可解释性研究侧重的角度不同,所提出的可解释性方法也各有不同,总体可分为内置可解释性和事后可解释性两大类。内置可解释性[36]的方法是指设计本身具有良好的可解释性的模型;而事后可解释性的方法是指利用可解释的方法对已设计好的模型进行解释,给出决策依据。

线性回归、朴素贝叶斯模型和决策树模型等都可以当作常用的内置可解释模型,由这些常用的可解释模型也衍生出了许多复杂深度学习模型的代理模型,进而得到事后可解释性方法。近年来,关于事后可解释性的方法不断被提出,其中主要包括全局可解释性和局部可解释性的方法[37]。没有统一的定义方式,下面分别从全局可解释性、局部可解释性和内置可解释性这三个角度给出如下定义:

定义 1　全局可解释性。给定样本集 X 和深度学习模型 M,ε 表示人类可理解的领域,若通过一个能逼近原模型 M 的决策过程的可解释全局模型 $m_g = f(X, M)$,可得到一个解释 $e_g(m_g, X) \in \varepsilon$,则称 e_g 为 M 的全局可解释方法。

定义 2　局部可解释性。给定一个样本 $\boldsymbol{x}$ 和深度学习模型 M,ε 表示人类可理解的领域,若通过一个可解释局部模型 $m_l = f(\boldsymbol{x}, M)$,可得到一个解释 $e_l(m_l, \boldsymbol{x}) \in \varepsilon$,则称 e_l 为 M 的局部可解释方法。进一步,若该解释 e_l 适用于一个与样本 $\boldsymbol{x}$ 相近的样本子集,则称 e_l 为半局部可解释方法。

定义 3　内置可解释性。给定样本集 X 和深度学习模型 M,ε 表示人类可理解的领域,若 M 本身具有局部可解释性或全局可解释性,即对任一样本 $\boldsymbol{x}$ 的决策存在一个解释 $e_l(M, \boldsymbol{x}) \in \varepsilon$,或对整体样本集 X 的决策存在一个解释 $e_g(M, X) \in \varepsilon$,称模型 M 具有内置可解释性。

人类可理解的领域 ε 可以是能给出形如“if-then”的逻辑

规则的解释，也可以是能为人类提供“它为什么这么做？它是怎么做到的？”等问题的回答的直观解释（如图2）。根据呈现方式的不同，它还可以分别为可视化解释、文本解释和多模态的解释。比如，对于一张图片的分类，一种解释可能是通过对当前输出决策的贡献更大特征区域的可视化呈现，也可以是为当前输出决策的主要依据做文本标注等。

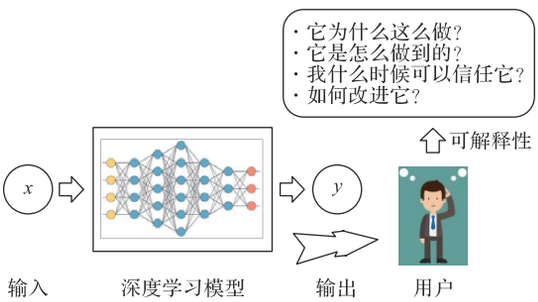

图 2 可解释性的含义

Fig. 2 Meaning of interpretability

2 深度学习可解释性研究进展

本章主要从解释深度学习模型的逻辑规则、决策归因和内部结构表示这三个方面出发介绍几种可解释性研究的典型方法。

2.1 解释深度学习模型的逻辑规则

2.1.1 基于决策树的可解释性方法

基于决策树或决策规则的解释往往是容易被人理解的，因此已有不少研究从深度学习模型中提取决策规则从而获得可理解的描述，同时满足提取的规则近似于原模型的决策结果。由于决策树可以被简化为决策规则集，在本文中没有明确区分基于决策树和决策规则这两种解释方法。

规则提取的解释方法大多是全局解释方法，可分为分解法和教学法。分解法是指将神经网络分解到神经元层面提取决策规则来模仿单个单元的行为。CRED（Continuous Rule Extractor via Decision tree induction）算法[38]利用决策树对神经网络进行分解，并将从每棵树中提取的规则进行合并，得到生成规则。该算法不依赖于网络结构，只提取数据中输入和输出变量之间的关系，同时适用于连续和离散的问题。

但是，CRED只适用于浅层的网络，DeepRED（Deep neural network Rule Extraction via Decision tree induction）[39]将CRED扩展到任意多个隐藏层的深度神经网络，该算法使用RxREN（Rule extraction by Reverse Engineering the Neural networks）[40]修剪不必要的输入，并应用算法C4.5[41]简化决策树，从而得到创建简约决策树的统计方法。

具体地说，DeepRED为每个类从输出层 $\boldsymbol{y}$ 到输入层 $\boldsymbol{x}$ 逐层反向提取规则，假设隐藏层的层数为 N，则应用算法C4.5得到隐藏层 $\boldsymbol{h}_k$ 基于前一个隐藏层 $\boldsymbol{h}_{k-1}$ 的决策规则 $R_{\boldsymbol{h}_{k-1}\to\boldsymbol{h}_k}(k=2,3,\cdots,N)$，最后合并所有规则 $R_{\boldsymbol{h}_N\to\boldsymbol{y}}$，$R_{\boldsymbol{h}_{N-1}\to\boldsymbol{h}_N},\cdots,R_{\boldsymbol{h}_1\to\boldsymbol{h}_2},R_{\boldsymbol{x}\to\boldsymbol{h}_1}$，得到了根据神经网络的输入来描述输出的规则集 $R_{\boldsymbol{x}\to\boldsymbol{y}}$。

虽然DeepRED能够构建与原始网络非常接近的完整树，但生成的树可能非常大，并且该方法的实现需要大量时间和内存，因此可扩展性受到限制。另一种教学法将深度学习模型视作一个黑盒子，直接将输入映射到输出来提取规则，而不是考虑神经网络的内部工作原理。DecText[42]就是采用经过黑盒子的数据来提取决策规则，该方法采用遗传算法对训练后的网络进行查询和原型提取，然后使用原型选择机制来选择原型的子集，最后，使用ID3或C5.0等标准归纳方法提取决策树。

给定一个已训练的神经网络和一个期望的输出向量，一个原型就是一个能被归为期望的输出类的输入向量。首先，采用遗传算法的原型提取方法，其中遗传算法的适应度函数为：

$$fitness=\sum_{i=1}^{n}\mathrm{abs}\left(\mathrm{ANN}\left(\boldsymbol{x}_i\right)-\boldsymbol{y}_i\right) \tag{1}$$

其中：$\mathrm{ANN}\left(\boldsymbol{x}_i\right)$ 表示输入 $\boldsymbol{x}_i$ 经过已训练的神经网络得到的输出。在原型被提取之后，通过概率神经网络和学习向量量化方法过滤掉那些不太可能服从给定训练集分布的原型，只保留那些满足训练集分布的原型。

为了克服决策树加深对可解释性造成的影响，Wu等[43]提出了区域树正则化的方法，该方法采用预定义的覆盖整个输入空间的区域集所对应的决策树集很好地逼近深度模型。全局树正则的定义如下：

$$\Omega^{\mathrm{global}}(\theta)=\mathrm{APL}\left(\{\boldsymbol{x}_i\}_{n=1}^{n},f(\cdot,\theta)\right) \tag{2}$$

其中：$\{\boldsymbol{x}_i\}_{i=1}^{n}$ 表示输入数据集；$f(\cdot,\theta)$ 表示多层感知机；$\mathrm{APL}(\cdot)$ 是指文献[44]中定义的平均决策路径长度（Average decision Path Length，APL），它刻画了模拟目标神经模型的单一全局决策树的复杂性。

但是，在实践中，全局树正则化的方法往往既不能得到可解释的优化，也不能得到性能优化，因此考虑找到一个在每个区域都很简单的高性能目标网络。当一些区域很复杂时，为了防止简单区域的正则化，选择只惩罚最复杂区域的平均决策路径长度，于是给出 L_0 区域树正则化定义：

$$\Omega^{\text{regional-}L_0}(\theta)=\max_{r\in\{1,2,\cdots,R\}}\mathrm{APL}\left(X_r,f(\cdot,\theta)\right) \tag{3}$$

其中：$X_1,X_2,\cdots,X_R$ 表示覆盖整个输入空间的区域集。通过在网络优化目标中加入区域树正则化项，在优化过程中抑制树的深度，得到更好的解释。

2.1.2 基于知识图谱的可解释性方法

2012年，Google推出了一款从Metaweb中衍生而来的产品——知识图谱（Knowledge Graph，KG），之后随着深度学习的发展，KG开始逐渐应用在人工智能领域的各个方面。KG[45]是指一个由表示实体的节点集 $\mathcal{E}$ 和表示关系的边集 $\mathcal{R}$ 构成的有向图，记为：

$$\mathcal{KG}=\left\{(h,r,t)\middle|h,t\in\mathcal{E},r\in\mathcal{R}\right\} \tag{4}$$

由于KG大多数属于异构图结构，对比其他的数据结构有更强的表达能力，因此，基于KG的可解释性通常比基于决策树的解释方法包含更多信息，更容易让人类理解。本节主要从基于路径的方法和基于嵌入的方法这两个方面对KG在深度学习可解释性中的研究进行一个概述。

1）基于路径的方法。

近年来，将KG引入到推荐系统中的研究引起了越来越

多的关注。通过探索KG内的相互联系,可以发现用户$\mathcal{U}=\{u_t\}_{t=1}^{M}$与项目$\mathcal{I}=\{i_t\}_{t=1}^{I}$之间的连接$p(u,i)=\left[u\xrightarrow{r_1}\cdots\xrightarrow{r_{L-1}}i\right]$作为路径,为用户与项目的交互提供丰富而互补的信息。在基于路径的方法的文献中,文献[46-47]引入了元路径的概念以明确地指导推荐倾向,元路径是指一个实体类型的序列,用于捕获KG中携带的用户-项目关联。然而,由于关系通常被排除在元路径之外,因此很难明确路径的整体语义,而且元路径需要预定义领域知识。

为了建模实体的顺序依赖关系和连接用户-项目对的路径的复杂关系,同时还能在推断用户兴趣时能区分不同路径的不同贡献提高模型可解释性,Wang等[45]提出了一种新的解决方案,称为知识感知路径递归网络(Knowledge-aware Path Recurrent Network, KPRN),该模型通过组合实体和关系的语义来生成路径表示,然后采用长短期记忆(Long Short-Term Memory, LSTM)网络来建模实体和关系的顺序依赖关系。最后,执行池操作来聚合路径的表示,以获得用户-项目对的预测信号。更重要的是,用一种新的加权池化操作来区分用户与物品连接的不同路径的贡献大小,使模型具有一定的可解释性。

KPRN模型分为三层:首先,嵌入层将KG的实体以及对应的关系映射到一个向量作为输入;然后,解码层用LSTM根据这些路径对下游任务进行解码,学习这些路径的时序依赖;最后,经过池化层的处理后,根据不同的路径分数对物品进行打分$\hat{y}_{ui}=f_\theta\big(u,i|p(u,i)\big)$。模型的学习任务是一个二元分类问题,其中一个观察到的用户-项目对被分配一个目标值$y_{ui}=1$,否则为0,目标函数的定义如下:

$$\mathcal{L}=-\sum_{(u,i)\in\mathcal{O}^+}\log\hat{y}_{ui}+\sum_{(u,j)\in\mathcal{O}^-}\log\left(1-\hat{y}_{uj}\right)\tag{5}$$

其中:$\mathcal{O}^+=\left\{(u,i)|y_{ui}=1\right\}$,$\mathcal{O}^-=\left\{(u,j)|y_{uj}=0\right\}$分别表示正和负用户-项目关系对。当这种模型训练好之后,可以使用这种模型对用户进行物品推荐的同时,追溯推荐的原因,赋予推荐系统推理能力和可解释性。

但KPRN在大规模KG中完全探索每个用户-项目对的所有路径是不现实的。文献[48]中提出一种称为策略导向路径推理(Policy-Guided Path Reasoning, PGPR)的方法,它用强化学习的方法去代替有监督学习,通过一个智能体自动在图上探索解释的路径,使这种方法得到的解释更加灵活。跟大多数现有方法不同的是,它不只利用KG来获得更准确的推荐,而且使用知识执行显式推理,以便通过可解释的因果推理过程生成并支持推荐。

PGPR的目标是通过与知识图环境交互,学习从用户导航到潜在感兴趣项目的策略,然后在路径推理阶段采用经过训练的策略向用户提出建议。文献[48]中首先给出了可解释推荐的知识图推理(Knowledge Graph Reasoning for Explainable Recommendation, KGRE-Rec)问题的形式化定义,对于给定的知识图谱$\mathcal{G}$和用户$u\in\mathcal{U}$,KGRE-Rec问题的目标是找到一个推荐项目集$\{i_t\}_{t\in[I]}\subseteq\mathcal{I}$满足每个用户-项目对$(u,i_t)$关联一条推理路径$p(u,i_t)$。

然后,将KGRE-Rec问题形式化为马尔可夫决策过程(Markov Decision Process, MDP),记为:

$$Mdp=\left\langle s_t,a_t,P_{s_t,a_t},R_T\right\rangle\tag{6}$$

其中:状态$s_t=(u,e_t,h_t),e_t,h_t\in\mathcal{E}$,状态$s_t$的全体动作空间记为:

$$A_t=\{a_t\}=\left\{(r,e)\middle|e\notin\{e_0,e_1,\cdots,e_{t-1}\}\right\}\tag{7}$$

同时引入一种基于评分函数的用户-条件动作剪枝策略,评分函数$f\big((r,e)|u\big)$为将任意边(r,e)映射为在用户u的条件下的实值分数,状态s_t的用户-条件剪枝动作空间定义为:

$$\tilde{A}_t(u)=\left\{(r,e)\middle|\mathrm{rank}\Big(f\big((r,e)|u\big)\Big)\leqslant\alpha,(r,e)\in A_t\right\}\tag{8}$$

其中:α是给定的动作空间大小的上界。

另外,采用考虑仅对终端状态$s_T=(u,e_T,h_T)$给予软奖励:

$$R_T=\begin{cases}\max\left(0,\dfrac{f(u,e_T)}{\max\limits_{i\in\mathcal{I}}f(u,i)}\right), & e_T\in\mathcal{I}\\ 0, & \text{其他}\end{cases}\tag{9}$$

其中:$f(\cdot,\cdot)$表示评分函数;u和i分别表示用户和项目;e_T表示终端实体。

最后,从状态$s_t=(u,e_t,h_t)$采取行动$a_t=(r_{t+1},e_{t+1})$转移到$s_{t+1}=(u,e_{t+1},h_{t+1})$的概率为:

$$P_{s_t,a_t}\left[s_{t+1}\middle|s_t=(u,e_t,h_t),a_t=(r_{t+1},e_{t+1})\right]=1\tag{10}$$

其中:s_t表示在t时刻包含用户u和实体e_t,h_t的状态;a_t表示在t时刻包含关系r_{t+1}和实体e_{t+1}的行动。

基于上述的MDP,PGPR的目标是学习一个随机策略π,使任何初始用户u的预期累积奖励最大化:

$$J(\theta)=\mathrm{E}_\pi\left[\sum_{t=0}^{T-1}\gamma^tR_{t+1}\middle|s_0=(u,u,\phi)\right]\tag{11}$$

PGPR方法的所有实验是在亚马逊电子商务数据集[49]上进行的,该数据集由亚马逊的产品评论和元信息组成。KGRE-Rec问题的目标是推荐测试集中用户购买的物品,以及每个用户-物品对的推理路径。与之前的方法相比,PGPR方法在所有数据集上的归一化折损累计增益(Normalized Discounted Cumulative Gain, NDCG)、命中率、召回率和精度都优于所有其他基线。

2)基于嵌入的方法。

基于KG的可解释性方法的另一个研究方向是利用KG嵌入模型[50-51],将KG中的元素映射到一个正则向量空间中,并通过计算实体之间的表示距离来揭示实体之间的相似性,这有助于提升算法的性能。然而,KG嵌入方法缺乏发现多跳关系路径的能力。Ai等[52]提出了协同过滤(Collaborative Filtering, CF)方法在KG嵌入基础上进行个性化推荐,然后提出了一种软匹配算法来寻找用户与商品之间的解释路径。

首先,假设产品知识可以表示成一组三元组$\mathcal{S}=\{(e_h,e_t,r)\}$,其中$r$是从实体$e_h$到实体$e_t$的关系。因为一个实体可以与一个或多个其他实体通过一个或多个关系关联,Ai等[52]提出将CF的实体和关系的建模分离,通过将每个实体投影到一个低维隐空间$\mathbb{R}^d$,并将每个关系视为一个转换函

数 trans(·),将两个实体 $e_h \in \mathbb{R}^d$ 和 $e_t \in \mathbb{R}^d$ 的关系建模为一个从 e_h 到 e_t 的线性投影,即

$$e_t = \text{trans}(e_h, r) = e_h + r \tag{12}$$

由于求解式(12)的所有解在实际中是不可行的,Ai 等[52]采用基于嵌入式的生成框架来学习,优化目标为:

$$P(e_t|\text{trans}(e_h, r)) = \frac{\exp(e_t \cdot \text{trans}(e_h, r))}{\sum_{e_t' \in E_t} \exp(e_t' \cdot \text{trans}(e_h, r))} \tag{13}$$

其中:E_t 是与 e_t 拥有相同类型的所有可能实体的集合。

然后,为了生成推荐解释的关键是在知识图中找到从用户到项目的合理逻辑推理序列。Ai 等[52]定义从实体 e_h 到 e_t 的解释路径为关系 $\mathcal{R}_\alpha = \{r_\alpha | \alpha \in [1, m]\}$ 和 $\mathcal{R}_\beta = \{r_\beta | \beta \in [1, n]\}$,满足

$$e_u + \sum_{\alpha=1}^{m} r_\alpha = e_i + \sum_{\beta=1}^{n} r_\beta \tag{14}$$

其中:当 $\alpha = 1$ 时,$e_u \in E_h^{r_\alpha}$;当 $\alpha \in [2, m]$ 时,$E_t^{r_{\alpha-1}} = E_h^{r_\alpha}$;当 $\beta = 1$ 时,$e_i \in E_h^{r_\beta}$;当 $\beta \in [2, n]$ 时,$E_t^{r_{\beta-1}} = E_h^{r_\beta}$。

然而,根据观察到的关系找到有效的解释路径通常是困难的,于是提出在解释构建的隐空间中进行实体软匹配,通过扩展 softmax 函数来计算实体的概率:

$$P(e_x|\text{trans}(e_u, \mathcal{R}_\alpha)) = \frac{\exp(e_x \cdot \text{trans}(e_u, \mathcal{R}_\alpha))}{\sum_{e' \in E_t^{r_m}} \exp(e' \cdot \text{trans}(e_u, \mathcal{R}_\alpha))} \tag{15}$$

其中:$E_h^{r_\alpha}$ 和 $E_t^{r_m}$ 分别是关于关系 r_α 和 r_m 的首实体集和尾实体集,$e_u \in E_h^{r_\alpha}$ 且 $e_x \in E_t^{r_m}$,$\mathcal{R}_\alpha = \{r_\alpha | \alpha \in [1, m]\}$,$\text{trans}(e_u, \mathcal{R}_\alpha) = e_u + \sum_{\alpha=1}^{m} r_\alpha$。

因此,可以用关系集 $\mathcal{R}_\alpha = \{r_\alpha | \alpha \in [1, m]\}$ 和 $\mathcal{R}_\beta = \{r_\beta | \beta \in [1, n]\}$ 构造一个对任意用户 e_u 和物品 e_i 的解释路径,并计算这个解释路径的概率:

$$P(e_x|e_u, \mathcal{R}_\alpha, e_i, \mathcal{R}_\beta) = P(e_x|\text{trans}(e_u, \mathcal{R}_\alpha))P(e_x|\text{trans}(e_i, \mathcal{R}_\beta)) \tag{16}$$

为了找到 (e_u, e_i) 的最佳解释,它按 $P(e_x|e_u, \mathcal{R}_\alpha, e_i, \mathcal{R}_\beta)$ 对所有路径进行排序,并从中挑选出最佳路径,使用预定义模板生成自然语言解释。

然而,这种策略的一个问题是,解释不是根据推理过程产生的,而是后来由用户和物品嵌入之间的经验相似匹配产生的。另一方面,之前的工作模型将查询映射到向量空间中的一个单点,但是在现实中一个复杂的查询可能隐含大量实体。Ren 等[53]提出一个基于嵌入的框架 Query2box,用于在大规模和不完全的 KG 中推理具有∧、∨和∃操作符的任意查询。Query2box 是用“箱子”不再是一个点来进行嵌入,它的目标是想融入存在正一阶逻辑推理能力到这些嵌入之中。其中,“箱子”的定义是:

$$Box_p \equiv \{\boldsymbol{v} \in \mathbb{R}^d : \text{cen}(p) - \text{off}(p) \leq \boldsymbol{v} \leq \text{cen}(p) + \text{off}(p)\} \tag{17}$$

其中:≤表示 element-wise 不等式;$\text{cen}(p) \in \mathbb{R}^d$ 表示箱子的中心;$\text{off}(p) \in \mathbb{R}^d \geqslant 0$ 是箱子的正偏移量。

另外,箱子的一组点对应于查询的一组回答实体,同时,Ren 等[53]证明了合取∧可以天然地表示为箱子间的交集,同时还证明了处理析取∨需要嵌入的维度和 KG 实体数成比例。但是,通过将查询转换为析取范式,Query2box 能够以可扩展的方式处理带有∧、∨和∃的任意逻辑查询。Ren 等[53]在三个标准的 KG(FB15k、FB15k-237 和 NELL995)上进行了实验,展示了 Query2box 的有效性,并表明该方法比目前的技术水平实现了高达 25% 的相对改进。

表 1 对基于 KG 的深度学习可解释模型的研究方法进行了简单的对比分析。

表 1 基于 KG 的可解释模型研究的概述

Tab. 1 Overview of researches on interpretability models based on knowledge graph

解释方法	典型方法	优缺点
基于路径的方法	KPRN[45],PGPR[48]	结构简单,可解释性强,但效率不高,不适合复杂逻辑推理
基于嵌入的方法	CF[52],Query2Box[53]	准确率较高,可适用于更高级的逻辑推理,但结构比较复杂,解释不够直观

2.2 解释深度学习模型的决策归因

2.2.1 特征归因

特征归因是根据输入特征对输出的影响,得到输入特征对于决策的重要性大小。下面将特征归因的解释方法主要分为基于扰动的方法、基于反向传播的方法和基于代理模型的方法三种。

1)基于扰动的方法。

基于扰动的可解释性方法是指通过探究输入数据的扰动对输出的影响,从而试图解释输入特征对相应类输出决策的重要性大小的方法。Zeiler 等[54]使用反卷积网络 DeConvNet 将 CNN 各隐藏层的特征进行可视化,另外,通过遮挡输入图像的不同区域并观察输出结果的变化,找到对结果影响最大的特征。模型通过训练以及反卷积操作后,提取效果最好的特征,并投影到像素空间进行可视化。通过可视化,能够发现当输入特征存在一定变形时,输出特征仍能够保持不变。同时,每层的可视化结果反映了网络的层次化特点,每层可以分别学习到图像的轮廓、颜色和纹理等。另一方面,通过可视化分析每层的特征以及特征随模型训练而发生变化也能更好地改进模型结构。

Koh 等[55]使用影响函数的方法得到一个测试样本 $\boldsymbol{x}_{\text{test}}$ 的输出的损失函数 $L(\boldsymbol{x}_{\text{test}}, \boldsymbol{x})$ 受原先训练样本 $\boldsymbol{x}$ 的扰动 $\boldsymbol{x} \rightarrow \boldsymbol{x}_\delta = \boldsymbol{x} + \delta$ 的变化为:

$$\frac{\partial L(\boldsymbol{x}_{\text{test}}, \boldsymbol{x})}{\partial \boldsymbol{x}} = -\frac{1}{n} \nabla_\theta L(\boldsymbol{x}_{\text{test}}, \theta^*)^{\mathrm{T}} \boldsymbol{H}_{\theta^*}^{-1} \left[\nabla_{\boldsymbol{x}} \nabla_\theta L(x, \theta^*)\right] \tag{18}$$

其中:$\theta^* = \arg\min_\theta \left\{\frac{1}{n} \sum_{i=1}^{n} L(\boldsymbol{x}_i, \theta)\right\}$;$L(\cdot)$ 表示损失函数;$\boldsymbol{H}_\theta = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta^2 L(\boldsymbol{x}_i, \theta^*)$ 是正定的海森矩阵。

从而,由式(18)得到模型预测结果主要依据的样本特征,并且通过实验展示了模型对决策特征的归因,同时上述

理论还可以用于生成对抗样本和修正错误的标注。

Zintgraf等[56]提出了一种预测差异分析方法,该方法通过记录去除一个输入特征$\boldsymbol{x}_i$对C类的预测概率值的变化:

$$p(C|X) - p(C|X_{\setminus i}) \tag{19}$$

其中:$X_{\setminus i}$表示除$\boldsymbol{x}_i$以外的所有输入特征的集合,观察个体数据特征与特定模型决策的正相关性和负相关性,进而突出显示给定输入图像中提供支持或反对相应类的证据的区域。

为了在特征未知时评估预测结果$p(C|X_{\setminus i})$,Zintgraf等[56]通过边缘化特征来模拟特征的缺失:

$$p(C|X_{\setminus i}) = \sum_{\boldsymbol{x}_i} p(\boldsymbol{x}_i|X_{\setminus i})\, p(C|X_{\setminus i}, \boldsymbol{x}_i) \tag{20}$$

然而,对$p(\boldsymbol{x}_i|X_{\setminus i})$进行建模很容易因为有大量的特征而变得不可行。因此,Zintgraf等[56]通过假设特征$\boldsymbol{x}_i$独立于其他特征来近似方程(20):

$$p(C|X_{\setminus i}) \approx \sum_{\boldsymbol{x}_i} p(\boldsymbol{x}_i)\; p(C|X_{\setminus i}, \boldsymbol{x}_i) \tag{21}$$

其中:先验概率$p(\boldsymbol{x}_i)$通常由该特征的经验分布来近似。但是由于采用边缘分布$p(\boldsymbol{x}_i)$近似$p(\boldsymbol{x}_i|X_{\setminus i})$的精确度太低,Zintgraf等[56]采用一个准确度更高的近似方法:

$$p(\boldsymbol{x}_i|X_{\setminus i}) \approx p(\boldsymbol{x}_i|\hat{X}_{\setminus i}) \tag{22}$$

其中:$\hat{X}$表示一个大小为$l \times l$包含点$\boldsymbol{x}_i$的区域。

当得到类别概率$p(C|X_{\setminus i})$的估计值时,就可以将其与$p(C|X)$进行比较,最后采用一种证据权重的评估方法,

$$WE_i(C|X) = \mathrm{lb}\big(odds(C|X)\big) - \mathrm{lb}\big(odds(C|X_{\setminus i})\big) \tag{23}$$

其中:$odds(C|X) = p(C|X)/\big(1 - p(C|X)\big)$。

然后通过条件采样、多元分析、深度可视化在ImageNet和医学影像(MRI脑扫描)两个数据集上实现可视化结果,说明了一种可以突出显示给定输入图像中提供支持或者反对某个类的证据的区域,为分类器决策过程提供新的视角。

Petsiuk等[57]提出了一种更通用的解释方法——基于随机输入采样的解释(Randomized Input Sampling for Explanation, RISE)方法,该方法通过将输入图像与随机掩码$R_i(i = 1, 2, \cdots, m)$逐元相乘得到的掩码图$I \odot R_i(i = 1, 2, \cdots, m)$作为输入,然后对随机掩码进行加权平均得到显著图,其中权重是模型的输出$f(I \odot R_i)$并通过随机掩码R的期望进行正则化:

$$S_{x,f}(\boldsymbol{x}_{ij}) \approx \frac{1}{\mathrm{E}(R) \cdot m} \sum_{i=1}^{m} f(I \odot R_i) \cdot R_i(\boldsymbol{x}_{ij}) \tag{24}$$

在几个基准数据集上的大量实验表明,RISE与之前的相关方法相比表现出了更好的性能。

表2对已有的基于扰动的解释方法的相关研究做了简单的概述和对比分析。

表2　关于基于扰动的方法的已有研究的总结

Tab. 2　Summary of existing researches on perturbation-based methods

典型方法	实验数据集(应用网络)	解释方法	优缺点
DeConvNet[54]	Caltech-101, Caltech-256, PASCAL VOC 2012 (AlexNet)	可视化卷积神经网络各隐藏层的特征,并通过遮挡输入图像的不同区域并观察输出结果的变化,找到对模型决策影响最大的特征	通过可视化呈现隐层学习到的特征,解释直观,但并未对模型整体的决策做解释
影响函数[55]	MNIST(Inception)	使用影响函数的方法得到模型预测结果主要依据的样本特征,并且通过实验展示了模型对决策特征的归因	理论严谨,计算得到改变一个训练数据之后对模型参数和模型预测的影响,但解释不够直观
预测差异分析[56]	ImageNet(AlexNet GoogLeNetVGG)	通过找到每个输入特征的相关值来观察各个特征与模型决策之间的正相关和负相关,进而突出显示给定输入图像中提供支持或反对相应类的证据的区域	可同时得到正相关和负相关的解释,可视化呈现解释直观,但计算较复杂,效率不高
RISE[57]	PASCAL VOC07, MSCOCO2014 ImageNet (ResNet50, VGG16)	基于随机输入采样的方法通过将输入图像与随机掩码逐元相乘得到的掩码图作为输入,然后对随机掩码进行加权平均得到解释图	在自动因果度量方面优于之前的解释方法,但不能解释视频和其他领域中复杂网络所做的决策

2)基于反向传播的方法。

以下主要将基于反向传播的方法分为梯度反向传播、类激活映射(Class Activation Mapping, CAM)、分层关联传播这三类典型的方法做介绍。

梯度反向传播:基于梯度的可解释性方法是指利用神经网络中信息流的反向传递来理解输入的变化对输出的影响,以解释输入特征对相应类输出决策的重要性大小的方法。由于损失函数关于输入的梯度反映了损失函数变化最快的方向,因此使用梯度来解释分类决策是一种自然的想法,如在线性模型中,梯度就是模型的权重系数,能直接反映样本特征重要性,权重绝对值越大,则该特征对最终预测结果的贡献越大,反之则越小。这也是线性模型通常被认为是可解释的一个重要原因。下面具体介绍一些常见的方法。

Simonyan等[58]提出了利用反向传播推断特征重要性的解释方法,通过计算模型的输出类别相对于输入图像的梯度来求解该输入图像所对应的分类显著图,从而可视化一个特定类的输出决策依据。Springenberg等[59]结合了文献[58]和文献[54]中的方法提出了导向反向传播方法,在梯度反向传播过程中只考虑正的误差信号,这种方法有助于解释深度网络中每个神经元对输入图像的影响。

与只计算输出针对当前输入的梯度不同,Sundararajan等[60]提出了一种集成梯度方法,该方法通过计算输入从某些起始值按比例放大到当前值的梯度的积分代替单一梯度,具体如下:

$$IntegratedGrads_i(\boldsymbol{x}) = \int_0^1 \frac{\partial f(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} \mathrm{d}\alpha \tag{25}$$

其中：$\{\gamma(\alpha)|0 \leqslant \alpha \leqslant 1\}$表示反事实图像集，反事实图像是指将图像像素从零按比例α缩放至实际图像中的值。它有效地解决了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。

然而，这些可视化的方法通过反向传播所得到的显著图通常含有很多噪声。为此，Smilkov 等[61]提出了一种平滑梯度的反向传播解释方法，通过向待解释样本中添加高斯噪声$N(0,\sigma^2)$对相似的样本进行采样，然后利用反向传播方法求解每个采样样本的决策显著图，最后将所有求解得到的显著图进行平均并将其作为对模型针对该样本的决策结果的解释，解决了视觉噪声问题。

类激活映射（CAM）：Zhou 等[62]提出了 CAM 解释方法，该方法利用全局平均池化层来替代传统 CNN 模型中除最后一个 softmax 层以外的所有全连接层，并通过将输出层的权重ω_k^C投影到最后一个卷积层的特征图$\boldsymbol{F}^k$得到类显著图以定位输入样本中区分类的重要区域。设y^C是关于类C的得分，则

$$y^C = \sum_k \omega_k^C \sum_i \sum_j F_{ij}^k \tag{26}$$

于是，可得类显著图$\boldsymbol{L}^C$：

$$L_{ij}^C = \sum_k \omega_k^C F_{ij}^k \tag{27}$$

然而，CAM 方法需要修改网络结构并重训练模型，因此，Selvaraju 等[63]对 CAM 方法进行了改进，提出了一种将梯度信息与特征映射相结合的梯度加权 CAM 方法 Grad-CAM。给定一个输入样本，Grad-CAM 首先计算目标类别相对于最后一个卷积层中每一个特征图的梯度$\frac{\partial y^C}{\partial F_{ij}^k}$并对梯度进行全局平均池化：

$$\omega_k^C = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^C}{\partial F_{ij}^k} \tag{28}$$

其中：Z表示特征图的像素点的个数，以获得每个特征图的重要性权重，然后，得到一个粗粒度的梯度加权类显著图$\boldsymbol{L}^C$用于定位图像中区分类的重要区域。与 CAM 相比，Grad-CAM 无需修改网络架构或重训练模型，避免了模型的可解释性与准确性之间的权衡，因而可适用于多种任务以及任何基于 CNN 结构的模型。

尽管上述 CAM 解释方法计算效率高，解释结果视觉效果好且易于理解，但缺乏像素级别梯度可视化解释方法显示细粒度特征重要性的能力。文献[64]中提出的 Grad-CAM++ 方法能提供更细粒度的解释结果，它只考虑梯度有正误差信号时，反向传播通过 ReLU 层，此时取权重

$$\omega_k^C = \sum_i \sum_j \alpha_{ij}^{kC}\,\mathrm{ReLU}\left(\frac{\partial y^C}{\partial F_{ij}^k}\right) \tag{29}$$

其中：α_{ij}^{kC}表示对于特定类C和激活映射k的梯度权重，其表达式为：

$$\alpha_{ij}^{kC} = \frac{\dfrac{\partial^2 y^C}{\left(\partial F_{ij}^k\right)^2}}{2\dfrac{\partial^2 y^C}{\left(\partial F_{ij}^k\right)^2} + \sum_i \sum_j F_{ij}^k \dfrac{\partial^3 y^C}{\left(\partial F_{ij}^k\right)^3}} \tag{30}$$

另外，y^C是关于类C的得分，F_{ij}^k是特征图$\boldsymbol{F}^k$在第i行第j列的像素值。类似地，可以得到用于定位图像中区分类的重要区域的梯度加权类显著图$\boldsymbol{L}^C$。图 3 对 Grad-CAM 和 Grad-CAM++做了简单说明。

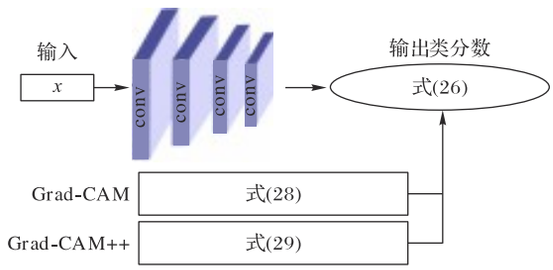

图3 Grad-CAM 和 Grad-CAM++的说明

Fig. 3 Description of Grad-CAM and Grad-CAM++

分层关联传播：基于梯度的可解释方法有时可能会失效，如考虑一个分段连续函数：

$$y = \begin{cases} x_1 + x_2, & x_1 + x_2 < 1 \\ 1, & x_1 + x_2 \geqslant 1 \end{cases} \tag{31}$$

当$x_1 + x_2 \geqslant 1$时，函数饱和此时梯度总是零，DeepLIFT[65]指出了这个问题，强调了在目标输入之外引入一个参考输入来解释重要性。在输出层，每个单元i的相关性是指在初始网络输入$\boldsymbol{x}$处激活的单元与在参考输入$\bar{\boldsymbol{x}}$处激活的单元的相对影响。

$$r_i^{(N+1)} = \begin{cases} y_i(\boldsymbol{x}) - y_i(\bar{\boldsymbol{x}}), & \text{单元}i\text{是目标单元} \\ 0, & \text{其他} \end{cases} \tag{32}$$

另外，分层相关性传播的公式为：

$$r_i^{(k)} = \sum_j \frac{z_{ij} - \bar{z}_{ij}}{\sum_i z_{ij} - \sum_i \bar{z}_{ij}} r_j^{(k+1)} \tag{33}$$

其中：$z_{ij} = \omega_{ij}^{(k+1.k)} x_i^{(k)}$和$\bar{z}_{ij} = \omega_{ij}^{(k+1.k)} \bar{x}_i^{(k)}$分别表示当网络输入为$\boldsymbol{x}$和$\bar{\boldsymbol{x}}$时神经元$i$和神经元$j$的加权激活函数。最后关于输入层的归因记为$R_i^C(\boldsymbol{x}) = r_i^{(1)}$。Deep-LIFT 是 LRP（Layerwise Relevance Propagation）[66]的一个更一般的情况，在 LRP 中，通常选零作为参考输入。

显著性方法旨在解释深度神经网络的预测，但是当解释对与模型预测无关的因素敏感时，解释方法就会缺乏可靠性。Kindermans 等[67]引入了输入不变性的概念，它要求归因方法满足模型对输入转换的不变性，并通过几个例子说明不满足输入不变性的显著性方法会导致错误归因。

3）基于代理模型的方法。

基于代理模型的可解释性方法是指通过简单的可解释模型作为代理模型对初始模型的局部决策或整体决策行为做出解释。

Ribeiro 等[68]提出与模型无关的局部可解释性的解释方法 LIME（Local Interpretable Model-agnostic Explanation，LIME），该方法通过扰动输入样本并构造一个局部线性模型作为输入的邻域内完整模型的简化代理，来判断对于输出结果有着最大的影响的可理解的特征。记内置可解释模型如决策树、线性模型等模型的集合为G，$\pi_x(z)$表示样本$\boldsymbol{x}$和其经过扰动得到的邻近样本$\boldsymbol{z}$之间的距离度量，Ribeiro 等[68]取线性模型$g \in G$，$L(f,g,\pi_x)$表示在定义的π_x下用g近似f的不可信度，则 LIME 生成的解释可记为：

$$\xi(\boldsymbol{x}) = \arg\min_{g \in G} L(f, g, \pi_{\boldsymbol{x}}) + \Omega(g) \tag{34}$$

其中:$\Omega(g)$表示模型g的复杂度。

由于线性模型的系数权重的大小反映了针对输入样例所做的决策依据的每一维特征重要性的大小,从而以一种可解释的且令人信服的方式解释任意分类器的预测值,并将该方法用于提取对网络输出高度敏感的图像区域。

由于LIME往往无法准确地解释如循环神经网络(Recurrent Neural Network, RNN)这种包含序列数据依赖关系的神经网络,Guo等[69]提出了一种适用于安全应用的高保真度解释方法LEMNA,利用一个简单的回归模型逼近复杂的深度学习决策边界的局部区域。与LIME不同的是,LEMNA假设待解释模型的局部边界是非线性的,首先通过训练混合回归模型来近似RNN针对每个输入实例的局部决策边界,然后通过引入融合Lasso正则来处理RNN模型中的特征依赖问题,有效地弥补了LIME等方法的不足,从而提高了解释的保真度。

虽然LIME和LEMNA较简单,但随机扰动和特征选择方法导致生成的解释不稳定,Zafar等[70]提出了一个确定性局部可解释模型不可知论解释(Deterministic Local Interpretable Model-agnostic Explainations, DLIME)方法,该方法使用凝聚层次聚类(Hierarchical Clustering, HC)和K-最近邻(K-Nearest Neighbour, KNN)算法来代替随机扰动,首先使用HC将训练数据分组聚类,并使用KNN来选择与待解释样例最近的邻域。当KNN选择了一个聚类时,在选定的聚类上训练一个线性模型来生成解释,该方法生成的模型解释比传统的LIME算法更稳定。另外,由于扰动样本由均匀分布产生,忽略了特征之间复杂的相关性,Shi等[71]引入一种使用修正扰动采样操作(Modified Perturbed Sampling Operation for LIME, MPS-LIME)对图像数据提取超像素信息的替代方法。通过将超像素转换为无向图,将传统的超像素选取操作转化为团集构造问题。各种实验表明,MPS-LIME对黑箱模型的解释在可理解性、保真度和效率方面取得了更好的性能。

Bramhall等[72]使用二次近似框架QLIME,将LIME提出的线性关系重新定义为二次关系,扩展了它在非线性情况下的灵活性,提高了特征解释的准确性。该模型使用的数据来自一家全球人力资源公司,其目标是成功预测候选人的工作安置问题。实验结果表明,QLIME增加了模型的可解释性,而且在使用均方误差作为比较度量方式的前提下,QLIME比LIME在预测类标签的均方误差方面有所改进。

2.2.2　概念归因

目前大部分深度学习模型在低级特征如像素值层面运算,而无法与人类能轻易理解的高级概念相对应。Kim等[73]引入概念激活向量(Concept Activation Vector, CAV),并使用方向导数来量化用户定义的概念对分类结果的敏感度,得到一种以人类友好的概念来解释神经网络内部状态的全局可解释性方法。

该方法首先定义感兴趣的概念$\mathcal{C}$,例如,斑马可以归因为正概念$P_{\mathcal{C}}$如条纹,另外,可以收集一组随机的照片作为负概念N。通过训练一个二元线性分类器$v_{\mathcal{C}}^k$来区分两个集合$\{z_k(\boldsymbol{x}): \boldsymbol{x} \in P_{\mathcal{C}}\}$和$\{z_k(\boldsymbol{x}): \boldsymbol{x} \in N\}$的层激活函数$z_k$,称$v_{\mathcal{C}}^k$为概念$\mathcal{C}$的CAV。然后,用类似于基于梯度的方法的方向导数的方法来评估特定层k的预测关于给定概念$\mathcal{C}$方向变化的敏感性。关于一个输入$\boldsymbol{x}$对概念$\mathcal{C}$的第C类的概念敏感性可以由一个概念向量$\boldsymbol{v}_{\mathcal{C}}^k \in \mathbb{R}^m$的方向导数$S_{C,\mathcal{C},k}(\boldsymbol{x})$得到:

$$S_{C,\mathcal{C},k}(\boldsymbol{x}) = \lim_{\varepsilon \to 0} \frac{y_k^C\left(z_k(\boldsymbol{x}) + \varepsilon \boldsymbol{v}_{\mathcal{C}}^k\right) - y_k^C\left(z_k(\boldsymbol{x})\right)}{\varepsilon} = \nabla y_k^C\left(z_k(\boldsymbol{x})\right) \cdot \boldsymbol{v}_{\mathcal{C}}^k \tag{35}$$

进而,若X_C表示所有标签为C的输入集合,则记模型关于整个输入类的概念敏感度为TCAV得分:

$$TCAV_{C,\mathcal{C},k} = \frac{\left|\left\{\boldsymbol{x} \in X_C: S_{C,\mathcal{C},k}(\boldsymbol{x})\right\} > 0\right|}{\left|X_C\right|} \tag{36}$$

但是,由于人们在选择概念时带有主观性,如果没有正确地选择输入概念,TCAV可能会产生无意义的CAV。与TCAV方法不同的是,Ghorbani等[74]提出了一种叫作自动概念解释(Automatic Concept Interpretation, ACE)的全局解释方法,通过在不同的数据中聚合相关的局部图像片段,在没有人工监督的情况下对训练好的分类器进行全局解释。为了提取类的所有概念,ACE首先对给定的类图像使用多个分辨率进行分割,然后将相似的片段作为相同概念的例子进行分组,最后,基于概念的TCAV分数为特定分类提供重要性评分并通过实验表明提取的概念适用于深度学习模型中的决策。

如果训练数据实例中包含多个类,即使类之间的相关性很低,诸如TCAV之类的方法也会遇到概念混淆的问题。此外,数据集中的偏差可能会影响概念,以及输入数据中的颜色。Goyal等[75]通过提出因果概念效应模型CaCE改进了TCAV方法,该模型研究了高层次概念的存在或缺失对深度学习模型预测的因果效应。

设已有概念$\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_m$,$g$表示生成模型,$\mathrm{do}(\mathcal{C}_i = 1)$,$\mathrm{do}(\mathcal{C}_i = 0)$分别表示概念$\mathcal{C}_i$存在或缺失的操作,$\mathrm{E}_g\left[\cdot \mid \mathrm{do}(\mathcal{C}_i = 1)\right]$表示do-operator操作[76]。于是,概念$\mathcal{C}_i$的存在或缺失对深度学习模型预测的因果效应定义为:

$$\mathrm{CaCE}(\mathcal{C}_i, f) = \mathrm{E}_g\left[f(X) \mid \mathrm{do}(\mathcal{C}_i = 1)\right] - \mathrm{E}_g\left[f(X) \mid \mathrm{do}(\mathcal{C}_i = 0)\right] \tag{37}$$

Goyal等[75]提出GT-CaCE(Ground Truth CaCE, GT-CaCE)的方法,通过对数据生成过程进行精确干预的情况下就可以准确地计算CaCE。另外,还阐述了一种使用变分自编码器(Variational Auto-Encoder, VAE)估算CaCE的方法,称为VAE-CaCE。在四个数据集上的实验结果表明,即使数据集存在偏差或相关性,CaCE方法的聚类和性能也得到了改善。

Yeh等[77]引入ConceptSHAP[76]来定义具有高完整性评分的概念的重要性。首先,给出两种完整性η的定义量化特定概念集$C = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_m\}$在解释模型行为方面的充分程度。然后,提出了一种概念发现方法,该方法在保证具有高完整性得分的同时引入了两个正则化项来提高已发现概念的可解释性,使用博弈论的概念对集合进行聚合,为每个发现的概念定义一个重要分数ConceptSHAP。

$$\psi_i(\eta)=\sum_{\tilde{C}\subseteq C\setminus C_i}\frac{(m-|\tilde{C}|-1)!|\tilde{C}|!}{m!}\left[\eta(\tilde{C}\cup\{C_i\})-\eta(\tilde{C})\right] \tag{38}$$

通过在合成数据集和真实世界的文本和图像数据集上的实验表明，该方法与TCAV相比在寻找能够完整解释决策和可解释的概念方面更有效。

2.2.3 样本归因

基于样本的解释方法是选择数据集的特定样本来解释机器学习模型的行为或底层数据分布。

基于原型的解释是样本归因中一种典型的解释方法，原型是从数据中选择具有代表性的样本，而批评是那些原型无法很好地表示的样本。随着现代数据集规模的增长，能够从数据集中选择一些“有代表性”的样本，向领域专家提供这些样本，具有越来越大的解释价值。Bien等[78]讨论了在分类设置中选择原型的方法，对于一个类C_i的原型应该由接近类C_i的许多训练点而远离其他类的训练点的点组成。设

$$B(\boldsymbol{x})=\{\boldsymbol{x}'\in\mathbb{R}^n\,|\,d(\boldsymbol{x},\boldsymbol{x}')<\varepsilon\} \tag{39}$$

则选择原型就是寻找最小的样本点子集$P\subseteq X$使$\{B(\boldsymbol{x}_i)\,|\,\boldsymbol{x}_i\in P\}$覆盖$X$，然后将集合覆盖问题表示成整数规划问题，并引出了两种标准的近似算法。最后，通过在USPS邮政编码数字数据集上进行实验演示了生成原型的解释价值，并表明作为一个分类器，它的性能相当好。

基于样本的解释被广泛用于提高高度复杂分布的可解释性，然而，仅靠原型不足以代表复杂性的要点。为了让用户构建更好的心理模型并理解复杂的数据分布，还需要用批评来解释哪些样本没有被原型捕获。在贝叶斯模型批评框架的推动下，Kim等[79]开发了能够有效学习原型和批评的MMD- critic。MMD-critic的目标是最小化选择的原型分布和数据分布之间的差异，其中最大平均差异的计算公式是：

$$\mathrm{MMD}^2(\boldsymbol{x})=\frac{1}{m^2}\sum_{i,j\in[m]}k(\boldsymbol{z}_i,\boldsymbol{z}_j)-\frac{2}{mn}\sum_{i\in[n],j\in[m]}k(\boldsymbol{x}_i,\boldsymbol{z}_j)+\frac{1}{n^2}\sum_{i,j\in[n]}k(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{40}$$

其中：$k(\cdot,\cdot)$是一个核函数，用于测量两点的相似性；m是原型$\boldsymbol{z}$的个数；n是初始数据集样本点的个数。

与现有方法相比，MMD-critic作为最接近的原型分类器表现出了较好的性能。一项人类受试者的初步研究表明，当批评与原型一起出现时，人类能够更好地执行预测任务，从而使数据分布得到很好的解释。

除此之外，不少学者还提出了一些利用原型构建可解释性的深度学习模型的方法。Li等[80]构建了一个包含自动编码器和原型层的网络架构，原型层的每个单元存储一个权重向量，类似于编码的训练输入。根据编码的输入和学到的原型之间的接近程度进行预测。除了交叉熵损失和自动编码器重构误差外，它们还包括两个可解释性正则化项，鼓励每个原型至少与一个编码输入相似，反之亦然。网络经过训练后，这些原型可以自然地用作解释。

与文献[80]不同的是，Chen等[81]引入一种深度网络架构的原型零件网络ProtoPNet，该模型不需要解码器来可视化原型，每个原型都是某个训练图像块的隐表示，自然而忠实地成为原型的可视化。另外，解码器的去除也有助于网络的训练，以获得更好的解释和更高的准确性。

2.3 解释深度学习模型的内部结构表示

深度学习模型的内部结构表示的解释旨在了解流经这些网络的数据的作用和结构，其中包括解释隐层的隐表示和单个神经元的行为。

2.3.1 隐层的表示

为了研究深度神经网络的每一层学习到哪些特征，Zeiler等[54]使用反卷积网络DeConvNet将CNN各隐藏层的特征进行可视化，从而直观地呈现出各隐藏层学习到的特征。通过实验能够发现每层的可视化结果反映了网络的层次化特点，低层学习到的特征基本上是颜色、边缘等通用特征，而随着层数的增加学习到的特征开始变得复杂，进一步学习到纹理、轮廓等比较有区别性的特征。另一方面，通过可视化分析每层的特征以及特征随模型训练而发生变化也能更好地改进模型结构。

另一方面，可以通过测试隐层学习的特征向量用于解决与网络最初训练的问题不同的任务的性能来解释其有效性和通用性。Razavian等[82]发现对目标图像进行分类的CNN的中间层输出产生的特征向量可以直接重新用于解决许多其他不同的识别任务，包括场景识别、细粒度识别、属性检测和图像检索等，这突出了隐层学习的隐变量表示的有效性和通用性。Razavian等[82]使用OverFeat网络针对不同识别任务进行了一系列实验，经过训练后可以在ILSVRC13上进行图像分类。实验结果表明即使像SVM这样简单的模型都能够直接将隐层学习的隐表示应用于目标问题，并且在不训练全新深度网络的情况下比先前的方法表现出更好的性能。

由于深度学习模型第一层学习的隐表示具有通用性，最后一层学习到的特征具有特殊性，于是需要进一步研究特征是如何从通用特征过渡到特定特征的。Yosinski等[83]定义了一个特定的方法来量化各层学习到的隐表示的可迁移性，并通过实验量化了CNN的每一层神经元的通用性和特殊性。结果表明，可迁移性受到两个不同问题的负面影响：1)较高层的神经元对其原始任务的特定性是以牺牲对目标任务的性能为代价的；2)相邻层上的共适应神经元之间的网络分裂而导致的优化困难。另外，特征的可转移性随着基本任务和目标任务之间距离的增加而降低，最后一个令人惊讶的结果是，使用从几乎任何层转移的特征来初始化网络，可以促进泛化，即使在对目标数据集进行微调后，这种泛化也会持续存在。

2.3.2 神经元的表示

单个隐层内的信息可以进一步细分为单个神经元或单个卷积滤波器，这些单个单元的作用可以通过创建输入模式的可视化来最大化单个单元的响应来定性地理解，或者通过测试一个单元解决迁移问题的能力来定量地理解。

2014年，Zhou等[84]提出一种数据驱动的方法来估计CNN中每个样本中不同神经元的感受野即其对原图像的感受范围的形状和大小。通过实验发现随着层的加深，每个神经元的感受野大小逐渐增加，激活区域变得更语义化而且可

以进行目标定位。

2018 年,DeepMind[85]表示不管是去掉高语义或者去掉低语义的神经元,对网络的整体分类准确度的影响都是无差异的,所以神经元的语义没有意义,也不影响网络的泛化能力。随后,Zhou 等[86]指出文献[85]中只是分析了神经元对整体分类准确度的影响,而忽略了对不同类别的分类结果的影响。他们指出去掉高语义的神经元,会对某些特定类别的分类有毁灭性影响。

2020 年,Bau 等[87]把之前的一系列解析一个神经元价值的工作整合起来,通过分析在激活或关闭神经元时网络所产生的变化,量化分析了场景分类网络和生成网络里面一个神经元的价值,并且将该分析框架应用于解释对抗性攻击和图片编辑。

总之,通过对单个神经元的系统分析可以对深层网络的黑盒内部产生深刻的见解,了解网络已经学习到的知识结构,并建立帮助人类与这些强大模型交互的系统。

最后,表 3 对第 2 章中从三类解释目标出发常见的几种解释方法及其相关文献做了概述总结。

表 3　可解释性文献的概述总结

Tab. 3　Overview summary of literatures about interpretability

解释目标	解释方法	典型方法
解释逻辑规则	决策树	分解法(CRED[38],DeepRED[39]);教学法(DecText[42],区域树正则化[43])
	KG	基于路径(KPRN[45],PGPR[48]);基于嵌入(CF[52],Query2box[53])
解释决策归因	特征归因	基于扰动(DeConvNet[54],影响函数[55],预测差异分析[56],RISE[57])
		梯度反向传播(Saliency Maps[58],导向反向传播[59],集成梯度[60],平滑梯度[61])
		类激活映射(CAM[62],Grad-CAM[63],Guided Grad-CAM[64])
		分层关联传播(Deep-LIFT[65],LRP[66],输入不变性[67])
		基于代理模型(LIME[68],LEMNA[69],DLIME[70],MPS-LIME[71],QLIME[72])
	概念归因	TCAVs[73],ACE[74],CaCE[75],ConceptSHAP[76]
	样本归因	Prototype selection[78],MMD-critic[79],ProtoPNet[81]
解释内部结构表示	层的表示	DeConvNet[54],文献[82-83]
	神经元的表示	文献[84,86],DeepMind[85],文献[87]

3　内置可解释模型的构建

3.1　基于决策树的模型

由于决策树模型可以被线性化为一系列由 if-then 形式组成的决策规则,所以浅层的决策树模型是通常被认为是可解释的,于是,由此衍生出了许多可解释的深度学习模型。Letham 等[88]引入贝叶斯规则列表(Bayesian Rule List, BRL)得到一个生成模型,对可能的决策列表产生后验分布,在保持准确性的同时提高可解释性。Yang 等[89]进一步通过改进理论边界、计算重用和高度调优的语言库提高了 BRL 的可伸缩性。

另外,Zhou 等[90]提出了一种基于决策树的内置可解释性的深度学习方法 gcForest。该方法采用一种深度树集成方法,比深度神经网络具有更少的超参数,并且可以根据数据自动确定模型复杂度。另外,gcForest 所需的训练数据集较小,这使 gcForest 训练起来更容易,也使其可解释性理论分析更简单。该算法具有很强的鲁棒性,即使遇到不同领域的不同数据,也能取得很好的结果。

目前大多数的可解释模型是基于使用实际标签的数据或基于黑盒模型的预测,但是得到的全局可解释模型可能与黑盒模型的局部解释不一致。Pedapati 等[91]构造了一个透明的全局模型,同时与黑盒模型的局部解释保持准确性和一致性。Pedapati 等[91]引入了一个自然的局部一致性度量,量化黑箱模型的局部解释和预测是否也与代理全局透明模型一致。同时,从黑盒模型的稀疏局部对比解释中创建自定义布尔特征,然后训练一个全局透明模型,并通过实验表明,与其他已知策略相比,这些模型具有更高的局部一致性,而且在性能上仍然接近那些通过访问原始数据而训练出来的模型。

3.2　广义加性模型

1986 年,Hastie 等[92]提出了广义可加模型 GAM,其形式如下:

$$g(\mathrm{E}(\boldsymbol{y})) = \beta + \sum_{i=1}^{n} f_i(\boldsymbol{x}_i) \tag{41}$$

其中:$g(\cdot)$是连接函数;$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$是输入的 n 个特征;f_i是一个单变量函数,满足 $\mathrm{E}(f_i) = 0$;$\boldsymbol{y}$是目标变量。进一步,为了提高准确性,可以将成对的相互作用添加到 GAM 中,形成一个模型 $\mathrm{GA^2M}$[93],形式如下:

$$g(\mathrm{E}(\boldsymbol{y})) = \beta + \sum_{i=1}^{n} f_i(\boldsymbol{x}_i) + \sum_{i \neq j} f_{i,j}(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{42}$$

由于对 GAM 使用低阶光滑样条函数进行拟合能减少过度拟合且适合分析,Caruana 等[93]将具有成对交互的高性能广义可加模型 $\mathrm{GA^2M}$ 应用于真实的医疗保健问题,获得了具有高精确度的可解释模型。

另外,往往 GAM 需要数百万棵决策树来使用加法算法提供准确的结果,作为 GAM 的一种改进的方法,Agarwal 等[94]提出了神经可加性模型(Neural Additive Model, NAM)(如图 4),它将深度神经网络的某些表达性与广义可加性模型固有的可理解性结合起来。

NAM 学习神经网络的线性组合,每个神经网络关注一个单一的输入特征,同时这些网络是联合训练的,可以学习它们的输入特征和输出之间任意复杂的关系。通过在回归和分类数据集上的实验表明,NAM 比常见的可解释模型如逻辑回归和浅层决策树更准确,在精确度上与现有的最先进的广义加性模型相似,但可以更容易地应用于现实世界的

问题。

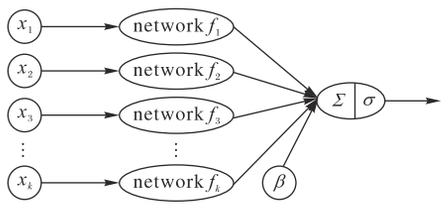

图 4 神经可加性模型

Fig. 4 Neural additive model

3.3 生成可解释模型

生成可解释模型是指通过设计生成人类可理解的模型如视觉问答系统[95-96]等，作为深度神经网络显式训练的一部分。在完成系统的主要任务的同时，它还可以生成可视化解释、文本解释以及同时包含这两者的多模态解释。

Hendricks 等[97]提出一个使用自然语言进行深层视觉解释的框架，联合分类和解释模型，对图像给出的预测标签的依据做出可视化解释。该模型基于长时递归卷积神经网络(Long-term Recurrent Convolutional Network, LRCN)[98]，它由一个卷积网络和两个堆叠的 LSTM 组成，前者提取高级视觉特征，后者根据视觉特征生成描述。与 LRCN 不同的是，Hendricks 等[97]通过同时包含相关性损失和区别性损失来确保生成的描述满足特定图像实例中呈现的视觉内容的同时，包含适当的信息来解释图像为何属于特定类别。CUB 数据集上的结果显示，该模型能够生成与图像一致的解释，而且比此前的字幕方法生成的描述更具鉴别性。

由于之前的可解释模型大多是单模态的，文献[99]中提出一种生成包含视觉和文本解释的多模态解释方法，表明两种模态之间有互相促进提升解释质量的优势。该系统建立在 2016 年视觉问答(Visual Question Answering，VQA)[96]挑战的获胜者的基础上，并进行了一些简化和添加。该模型定义了活动识别任务(Activity Recognition Task, ACT-X)和视觉问答任务(Visual Question Answering Task, VQA-X)的数据集，除了问答任务和内部注意图，该系统还训练了一个额外的解释生成器，以及优化为视觉解释的第二注意图。无论是视觉解释还是文字解释，实验表明在用户信任和解释质量的评估上都有很好的得分。

另一方面，由于 VQA 模型往往只捕捉到训练集合中表面的语言相关性，不能推广到不同 QA 分布的测试集。理想的 VQA 模型应具有以下两个不可缺少的特性：1)视觉可解释性，模型在做出决策时应该依赖于正确的视觉区域；2)问题敏感型，该模型应该对所讨论的语言变化敏感。为此，文献[100]中提出一个模型不可知的反事实样本合成训练方案 CSS。CSS 通过掩盖图像中的关键对象或问题中的单词，并分配不同的真实答案，生成大量反事实训练样本。在使用原始和生成的样本训练之后，VQA 模型被迫集中于所有关键的对象和单词，这显著提高了视觉解释和问题敏感的能力，同时模型的性能得到了进一步提升。

4 评价指标

不同类型的解释之间的可解释性往往很难进行比较，需要针对不同解释方法的目的提出一些不同的评价方法。例如，对于基于决策树和逻辑规则的解释方法，通常将提取的规则模型的大小作为解释的复杂度的评判标准[101-102]，如规则的数量、每条规则的前因数量、决策树的深度等。本章主要介绍忠实度、准确性、鲁棒性和可理解性这四种的评价指标。

1)忠实度：指作为一个待解释模型的代理可解释模型与原模型的接近程度。文献[103]中定义了待解释模型 M 和其可解释模型 m 的局部近似程度为 $\sigma(M,m)=1/\kappa(M,m)$，记

$$\kappa(M,m)=\frac{1}{n}\sum_{i=1}^{n}|y_i-y_i'| \tag{43}$$

其中：$\boldsymbol{y}=[y_1,y_2,\cdots,y_n]$ 是 $\boldsymbol{x}$ 经过模型 M 得到的预测向量；$\boldsymbol{y}'=[y_1',y_2',\cdots,y_n']$ 是 $\boldsymbol{x}$ 经过模型 m 得到的预测向量。$\kappa(M,m)$ 越小，M 和 m 的局部近似程度越高，则认为 m 是 M 的一个较好的可解释模型。

2)准确性：指可解释方法得到的特征归因的正确性。Hooker 等[104]通过删除重要归因的输入特征并对编辑后的数据实例进行预测，进而观察由此产生的性能下降情况，从而可以评估所得的特征归因的准确性。但是如果不对模型进行再训练，修改后的输入可能会落在训练数据流形之外，因此，很难区分准确性下降是由于数据落入分布之外还是由于良好的特征归因。另一方面，重新训练导致模型与被解释的原始模型不同，因此应该采用仍服从原来分布的输入对原始模型进行方法评估。

Yang 等[105]引入一个名为基准归因方法(Benchmarking Attribution Method, BAM)的框架来评估特征归因的正确性和它们的相对重要性。BAM 数据集是通过复制称为公共特征的像素组生成的，这些像素组代表 MSCOCO 数据集[106]中的对象类别，并将它们粘贴到 MiniPlaces 数据集[107]中。由于专注于粘贴对象的归因方法在增强重要特征的特征归因方面做得并不好，Yang 等[105]还提出了三个定量评价归因方法的指标：1)模型对比评分 MCS，用来比较不同模型之间的相对特征重要性；2)输入相关率 IDR，用来学习公共特征对单个实例的相关性；3)输入独立率 IIR：用来学习两个功能相似的输入之间特征的差异性。

3)鲁棒性：指可解释模型相对于输入的局部扰动引起的输出的变化的稳定性程度。Alvarez-Melis 等[108]给出了量化解释生成模型 $G(\boldsymbol{x})$ 的稳定性的表达式：

$$s(\boldsymbol{x})=\underset{\boldsymbol{x}'\in N(\boldsymbol{x})}{\arg\max}\frac{\|G(\boldsymbol{x})-G(\boldsymbol{x}')\|}{\|\boldsymbol{x}-\boldsymbol{x}'\|} \tag{44}$$

其中：$N(\boldsymbol{x})=\{\boldsymbol{x}'\in X\,|\,\|\boldsymbol{x}-\boldsymbol{x}'\|\leqslant\varepsilon\}$ 表示 $\boldsymbol{x}$ 的 ε-邻域。该指标能反映当输入发生变化时解释的变化大小，还能看出对相邻近的样例的解释是否具有一致性。

4)可理解性：指人类对可解释方法合理性和容易理解的程度的评估，也就是解释符合人类期望的程度。Mohseni 等[109]引入一个以人为基础的评估基准来评估由可解释算法生成的特征显著性图解释，这种以人为基础的基准能够快速、可复制和客观地执行显著性解释的评估实验。与此同时，这种方法的一个根本缺陷可能是在解释中加入了人为偏

见。然而,人类对来自一个大群体的单个数据点的标签可以抵消固有偏见的影响。Holzinger等[110]引入系统因果关系量表(System Causability Scale, SCS)来理解面向用户的人机界面的解释需求,同时描述了一个将SCS工具应用于弗雷明汉风险工具的医疗场景,以了解人机界面的特定特征的影响和重要性。

5　未来的发展方向

目前大多数的可解释性研究主要是对深度学习模型的行为和做出决策的潜在原因的解释,但是关于如何在不损害网络性能的情况下主动地使深度学习模型可解释仍然是一个有待解决的问题。同时,这些研究中大多数处理的是数据驱动的可解释性,以克服黑盒算法的不透明性,针对目标驱动的可解释性研究如可解释的智能体等的贡献仍然缺失,发展具有可解释性的人工智能体,取得人类用户的"信任",从而产生高效的人机协作,进而融入一个人机共生共存的社会是未来人工智能研究的一个美好愿景[111]。此外,目前仍缺乏完备的数学理论指导和改进深度学习模型的表达能力、学习优化能力和泛化能力,为深度学习模型提供理论保证的道路仍然任重而道远。

6　结语

由于对透明人工智能系统需求和兴趣的日益增长,本文进行了一个对深度学习可解释性研究的全面回顾。首先,阐明了可解释性研究的需求和定义,然后,详细介绍了从三种解释目标出发的可解释性研究的几种典型方法并指出了各模型提出的原因以及具有的优缺点,同时还指出了三种类型的内置可解释模型的构建方法,随后还给出了几种常见的对可解释性的评价指标。最后对未来的研究方向进行了阐述,指出了其未来巨大的应用潜力。总之,随着对深度学习可解释性研究的不断深入,未来势必将发挥越来越重要的作用。

参考文献 (References)

[1] LEE S M, SEO J B, YUN J, et al. Deep learning applications in chest radiography and computed tomography [J]. Journal of Thoracic Imaging, 2019, 34(2): 75-85.

[2] CHEN R P, YANG L, GOODISON S, et al. Deep-learning approach identifying cancer subtypes using high-dimensional genomic data[J]. Bioinformiatics, 2020, 36(5): 1476-1483.

[3] GRIGORESCU S, TRASNEA B, COCIAS T, et al. A survey of deep learning techniques for autonomous driving [J]. Journal of Field Robotics, 2020, 37(3): 362-386.

[4] FENG D, HAASE-SCHÜTZ C, ROSENBAUM L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(3): 1341-1360.

[5] SAHBA A, DAS A, RAD P, et al. Image graph production by dense captioning[C]// Proceedings of the 2018 World Automation Congress. Piscataway: IEEE, 2018: 1-5.

[6] BENDRE N, EBADI N, PREVOST J J, et al. Human action performance using deep neuro-fuzzy recurrent attention model[J]. IEEE Access, 2020, 8: 57749-57761.

[7] BOLES A, RAD P. Voice biometrics: deep learning-based voiceprint authentication system [C]// Proceedings of the 12th System of Systems Engineering Conference. Piscataway: IEEE, 2017: 1-6.

[8] PANWAR S, DAS A, ROOPAEI M, et al. A deep learning approach for mapping music genres[C]// Proceedings of the 12th System of Systems Engineering Conference. Piscataway: IEEE, 2017: 1-5.

[9] DE LA TORRE PARRA G, RAD P, CHOO K K R, et al. Detecting Internet of Things attacks using distributed deep learning [J]. Journal of Network and Computer Applications, 2020, 163: No. 102662.

[10] CHACON H, SILVA S, RAD P. Deep learning poison data attack detection [C]// Proceedings of the IEEE 31st International Conference on Tools with Artificial Intelligence. Piscataway: IEEE, 2019: 971-978.

[11] MHASKAR H N, POGGIO T. Deep vs. shallow networks: an approximation theory perspective[J]. Analysis and Applications, 2016, 14(6): 829-848.

[12] LIAO Q L, POGGIO T. Theory of deep learning Ⅱ: landscape of the empirical risk in deep learning: CBMM Memo No. 066[EB/OL]. (2017-06-23)[2021-09-23]. https://cbmm. mit. edu/sites/default/files/publications/CBMM% 20Memo% 20066_1703. 09833v2. pdf.

[13] ZHANG C Y, LIAO Q L, RAKHLIN A, et al. Musings on deep learning: properties of SGD, CBMM Memo Series 067[EB/OL]. (2017-12-26)[2021-09-23]. https://cbmm. mit. edu/sites/default/files/publications/CBMM-Memo-067-v4. pdf.

[14] CINÀ A E, TORCINOVICH A, PELILLO M. A black-box adversarial attack for poisoning clustering [J]. Pattern Recognition, 2022, 122: No. 108306.

[15] SEMWAL P, HANDA A. Cyber-attack detection in cyber-physical systems using supervised machine learning[M]// CHOO K K R, DEHGHANTANHA A. Handbook of Big Data Analytics and Forensics. Cham: Springer, 2022: 131-140.

[16] ENGSTROM L, TRAN B, TSIPRAS D, et al. Exploring the landscape of spatial robustness [C]// Proceedings of the 36th International Conference on Machine Learning. New York: JMLR.org, 2019: 1802-1811.

[17] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2014-02-19)[2021-05-16]. https://arxiv. org/pdf/1312. 6199. pdf.

[18] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 427-436.

[19] HENGSTLER M, ENKEL E, DUELLI S. Applied artificial intelligence and trust — the case of autonomous vehicles and medical assistance devices [J]. Technological Forecasting and Social Change, 2016, 105: 105-120.

[20] LUI A, LAMB G W. Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector[J]. Information and Communications Technology Law, 2018, 27(3): 267-283.

[21] WELD D S, BANSAL G. The challenge of crafting intelligible intelligence[J]. Communications of the ACM, 2019, 62(6): 70-79.

[22] BOSTROM N, YUDKOWSKY E. The ethics of artificial intelligence[M]// FRANKISH K, RAMSEY W M. The Cambridge Handbook of Artificial Intelligence Cambridge: Cambridge University Press, 2014: 316-334.

[23] ETZIONI A, ETZIONI O. Incorporating ethics into artificial intelligence[J]. The Journal of Ethics, 2017, 21(4): 403-418.

[24] STAHL B C, WRIGHT D. Ethics and privacy in ai and big data: implementing responsible research and innovation [J]. IEEE Security and Privacy, 2018, 16(3): 26-33.

[25] KESKINBORA K H. Medical ethics considerations on artificial intelligence [J]. Journal of Clinical Neuroscience, 2019, 64: 277-282.

[26] CHEN L Y, CRUZ A, RAMSEY S, et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening[J]. PLoS ONE, 2019, 14(8): No. e0220113.

[27] CHALLEN R, DENNY J, PITT M, et al. Artificial intelligence, bias and clinical safety[J]. BMJ Quality and Safety, 2019, 28(3):231-237.

[28] SINZ F H, PITKOW X, REIMER J, et al. Engineering a less artificial intelligence[J]. Neuron, 2019, 103(6): 967-979.

[29] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial machine learning at scale[EB/OL]. (2017-02-11)[2021-07-09]. https://arxiv. org/pdf/1611. 01236. pdf.

[30] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. (2015-03-20)[2021-05-16]. https://arxiv. org/pdf/1412. 6572. pdf.

[31] SU J W, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks [J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[32] HUANG S, PAPERNOT N, GOODFELLOW I, et al. Adversarial attacks on neural network policies[EB/OL]. (2017-02-08)[2020-05-16]. https://arxiv. org/pdf/1702. 02284. pdf.

[33] GOODMAN B, FLAXMAN S. European Union regulations on algorithmic decision-making and a "right to explanation"[J]. AI Magazine, 2017, 38(3): 50-57.

[34] CHOULDECHOVA A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments[J]. Big Data, 2017, 5(2): 153-163.

[35] VOIGT P, VON DEM BUSSCHE A. The EU General Data Protection Regulation (GDPR): A Practical Guide[M]. Cham: Springer, 2017: 141-187.

[36] ALVAREZ-MELIS D, JAAKKOLA T. Towards robust interpretability with self-explaining neural networks [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. , 2018: 7786-7795.

[37] GUIDOTTI R, MONREALE A, RUGGIERI S, et al. A survey of methods for explaining black box models[J]. ACM Computing Surveys, 2018, 51(5): No. 93.

[38] SATO M, TSUKIMOTO H. Rule extraction from neural networks via decision tree induction [C]// Proceedings of the 2001 International Joint Conference on Neural Networks. Piscataway: IEEE, 2001: 1870-1875.

[39] ZILKE J R, LOZA MENCÍA E, JANSSEN F. DeepRED-rule extraction from deep neural networks[C]// Proceedings of the 2016 International Conference on Discovery Science, LNCS 9956. Cham: Springer, 2016: 457-473.

[40] 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.

[41] SALZBERG S L. C4. 5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc. , 1993[J]. Machine Learning, 1994, 16(3): 235-240.

[42] BOZO O. Extracting decision trees from trained neural networks [C]// Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002:456-461.

[43] WU M, PARBHOO S, HUGHES M C, et al. Regional tree regularization for interpretability in deep neural networks [C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020: 6413-6421.

[44] WU M, HUGHES M C, PARBHOO S, et al. Beyond sparsity: tree regularization of deep models for interpretability [C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 1670-1678.

[45] WANG X, WANG D X, XU C R, et al. Explainable reasoning over knowledge graphs for recommendation [C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019: 5329-5336.

[46] YU X, REN X, GU Q Q, et al. Collaborative filtering with entity similarity regularization in heterogeneous information networks [C/OL]// Proceedings of the 2nd IJCAI Workshop on Heterogeneous Information Network Analysis. [2021-09-22]. http://hanj. cs. illinois. edu/pdf/hina13_xyu. pdf.

[47] GAO L, YANG H, WU J, et al. Recommendation with multi-source heterogeneous information [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. California: ijcai. org, 2018: 3378-3384.

[48] XIAN Y K, FU Z H, MUTHUKRISHNAN S, et al. Reinforcement knowledge graph reasoning for explainable recommendation[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2019: 285-294.

[49] HE R N, McAULEY J. Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering [C]// Proceedings of the 25th International Conference on World Wide Web. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee, 2016: 507-517.

[50] BORDES A, USUNIER N, GARCIAD-DURÁN A, et al. Translating embeddings for modeling multi-relational data [C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. , 2013: 2787-2795.

[51] LIN Y K, LIU Z Y, SUN M S, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2015: 2181-2187.

[52] AI Q Y, AZIZI V, CHEN X, et al. Learning heterogeneous knowledge base embeddings for explainable recommendation[J]. Algorithms, 2018, 11(9): No. 137.

[53] REN H Y, HU W H, LESKOVEC J. Query2box: reasoning over knowledge graphs in vector space using box embeddings[EB/OL]. (2020-02-29)[2021-05-16]. https://arxiv. org/pdf/2002. 05969. pdf.

[54] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks [C]// Proceedings of the 2014 European Conference on Computer Vision, LNCS 8689. Cham: Springer, 2014: 818-833.

[55] KOH P W, LIANG P. Understanding black-box predictions via influence functions [C]// Proceedings of the 34th International Conference on Machine Learning. New York: JMLR. org, 2017: 1885-1894.

[56] ZINTGRAF L M, COHEN T S, ADEL T, et al. Visualizing deep neural network decisions: prediction difference analysis [EB/OL].

(2017-02-15)[2021-05-16]. https://arxiv.org/pdf/1702.04595.pdf.

[57] PETSIUK V, DAS A, SAENKO K. RISE: randomized input sampling for explanation of black-box models[C]// Proceedings of the 2018 British Machine Vision Conference. Durham: BMVA Press, 2018: No. 1064.

[58] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps[EB/OL]. (2014-04-19)[2021-05-06]. https://arxiv.org/pdf/1312.6034.pdf.

[59] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for simplicity: the all convolutional net[EB/OL]. (2015-04-13)[2021-06-07]. https://arxiv.org/pdf/1412.6806.pdf.

[60] SUNDARARAJAN M, TALY A, YAN Q Q. Gradients of counterfactuals[EB/OL]. (2016-11-15)[2021-06-11]. https://arxiv.org/pdf/1611.02639.pdf.

[61] SMILKOV D, THORAT N, KIM B, et al. SmoothGrad: removing noise by adding noise[EB/OL]. (2017-06-12)[2021-06-23]. https://arxiv.org/pdf/1706.03825.pdf.

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

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

[64] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks[C]// Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2018: 839-847.

[65] SHRIKUMAR A, GREENSIDE P, KUNDAJE A. Learning important features through propagating activation differences[C]// Proceedings of the 34th International Conference on Machine Learning. New York: JMLR.org, 2017: 3145-3153.

[66] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLoS ONE, 2015, 10(7): No. e0130140.

[67] KINDERMANS P J, HOOKER S, ADEBAYO J, et al. The (un)reliability of saliency methods[M]// SAMEK W, MONTAVON G, VEDALDI A, et al. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, LNCS 11700. Cham: Springer, 2019: 267-280.

[68] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?" explaining the predictions of any classifier[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144.

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

[70] ZAFAR M R, KHAN N M. DLIME: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems[EB/OL]. (2019-06-24)[2021-07-03]. https://arxiv.org/pdf/1906.10263.pdf.

[71] SHI S, ZHANG X F, FAN W. A modified perturbed sampling method for local interpretable model-agnostic explanation[EB/OL]. (2020-02-18)[2021-08-16]. https://arxiv.org/pdf/2002.07434.pdf.

[72] BRAMHALL S, HORN H, TIEU M, et al. QLIME — a quadratic local interpretable model-agnostic explanation approach[J]. SMU Data Science Review, 2020, 3(1): No. 4.

[73] KIM B, WATTENBERG M, GILMER J, et al. Interpretability beyond feature attribution: quantitative Testing with Concept Activation Vectors (TCAV)[C]// Proceedings of the 35th International Conference on Machine Learning. New York: JMLR.org, 2018: 2668-2677.

[74] GHORBANI A, WEXLER J, ZOU J, et al. Towards automatic concept-based explanations[C/OL]// Proceedings of the 33rd Conference on Neural Information Processing Systems. [2021-09-21]. https://proceedings.neurips.cc/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf.

[75] GOYAL Y, FEDER A, SHALIT U, et al. Explaining classifiers with Causal Concept Effect (CaCE)[EB/OL]. (2020-02-28)[2021-08-19]. https://arxiv.org/pdf/1907.07165.pdf.

[76] PEARL J. Causality[M]. 2nd ed. Cambridge: Cambridge University Press, 2009.

[77] YEH C-K, KIM B, ARIK S Ö, et al. On completeness-aware concept-based explanations in deep neural networks[C]// NeurIPS 2020: Proceedings of the 2020 Advances in Neural Information Processing Systems 33. Berlin: Springer, 2020: 20554-20565.

[78] BIEN J, TIBSHIRANI R. Prototype selection for interpretable classification[J]. The Annals of Applied Statistics, 2011, 5(4): 2403-2424.

[79] KIM B, KHANNA R, KOYEJO O. Examples are not enough, learn to criticize! criticism for interpretability[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2016: 2288-2296.

[80] LI O, LIU H, CHEN C F, et al. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 3530-3537.

[81] CHEN C F, LI O, TAO C F, et al. *This* looks like *that*: deep learning for interpretable image recognition[C/OL]// Proceedings of the 33rd Conference on Neural Information Processing Systems. [2021-09-21]. https://proceedings.neurips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf.

[82] RAZAVIAN A S, AZIZPOUR H, SULLIVAN J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2014: 512-519.

[83] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks?[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 3320-3328.

[84] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Object detectors emerge in deep scene CNNs[EB/OL]. (2015-04-15)[2021-06-16]. https://arxiv.org/pdf/1412.6856.pdf.

[85] MORCOS A S, BARRETT D G T, RABINOWITZ N C, et al. On the importance of single directions for generalization[EB/OL]. (2018-05-22)[2021-05-16]. https://arxiv.org/pdf/1803.06959.pdf.

[86] ZHOU B L, SUN Y Y, BAU D, et al. Revisiting the importance of individual units in CNNs via ablation[EB/OL]. (2018-06-07)[2021-05-16]. https://arxiv.org/pdf/1806.02891.pdf.

[87] BAU D, ZHU J Y, STROBELT H, et al. Understanding the role of individual units in a deep neural network[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(48): 30071-30078.

[88] LETHAM B, RUDIN C, McCORMICK T H, et al. Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model[J]. The Annals of Applied Statistics, 2015, 9(3): 1350-1371.

[89] YANG H Y, RUDIN C, SELTZER M. Scalable Bayesian rule lists[C]// Proceedings of the 34th International Conference on Machine Learning. New York: JMLR. org, 2017: 3921-3930.

[90] ZHOU Z H, FENG J. Deep forest: towards an alternative to deep neural networks[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. California: ijcai. org, 2017: 3553-3559.

[91] PEDAPATI T, BALAKRISHNAN A, SHANMUGAN K, et al. Learning global transparent models consistent with local contrastive explanations [C/OL]// Proceedings of the 34th Conference on Neural Information Processing Systems. [2021-09-21]. https://proceedings. neurips. cc/paper/2020/file/24aef8cb3281a2422a59b51659f1ad2e-Paper. pdf.

[92] HASTIE T, TIBSHIRANI R J. Generalized additive models[J]. Statistical Science, 1986, 1(3):297-310.

[93] CARUANA R, LOU Y, GEHRKE J, et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission [C]// Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 1721-1730.

[94] AGARWAL R, MELNICK L, FROSST N, et al. Neural additive models: interpretable machine learning with neural nets[C/OL]// Proceedings of the 35th Conference on Neural Information Processing Systems. [2022-01-21]. https://proceedings. neurips. cc/paper/2021/file/251bd0442dfcc53b5a761e050f8022b8-Paper. pdf.

[95] ANTOL S, AGRAWAL A, LU J S, et al. VQA: visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 2425-2433.

[96] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2016: 457-468.

[97] HENDRICKS L A, AKATA Z, ROHRBACH M, et al. Generating visual explanations [C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9908. Cham: Springer, 2016: 3-19.

[98] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 2625-2634.

[99] PARK D H, HENDRICKS L A, AKATA Z, et al. Multimodal explanations: justifying decisions and pointing to the evidence[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8779-8788.

[100] CHEN L, YAN X, XIAO J, et al. Counterfactual samples synthesizing for robust visual question answering [C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10797-10806.

[101] ODAJIMA K, HAYASHI Y, TIANXIA G, et al. Greedy rule generation from discrete data and its use in neural network rule extraction[J]. Neural Networks, 2008, 21(7): 1020-1028.

[102] ZHANG Q, YANG Y, MA H, et al. Interpreting CNNs via decision trees [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 6261-6270.

[103] LEI X, FAN Y K, LI K C, et al. High-precision linearized interpretation for fully connected neural network[J]. Applied Soft Computing, 2021, 109: No. 107572.

[104] HOOKER S, ERHAN D, KINDERMANS P J, et al. A benchmark for interpretability methods in deep neural networks [C/OL]// Proceedings of the 33rd Conference on Neural Information Processing Systems. [2021-09-21]. https://proceedings. neurips. cc/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper. pdf.

[105] YANG M J, KIM B. Benchmarking attribution methods with relative feature importance[EB/OL]. (2019-11-04)[2021-05-01]. https://arxiv. org/pdf/1907. 09701. pdf.

[106] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]// Proceedings of the 2014 European Conference on Computer Vision, LNCS 8693. Cham: Springer, 2014: 740-755.

[107] ZHOU B L, LAPEDRIZA A, KHOSLA A, et al. Places: a 10 million image database for scene recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452-1464.

[108] ALVAREZ-MELIS D, JAAKKOLA T S. Towards robust interpretability with self-explaining neural networks [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. , 2018: 7786-7795.

[109] MOHSENI S, BLOCK J E, RAGAN E D. A human-grounded evaluation benchmark for local explanations of machine learning [EB/OL]. (2020-06-28)[2021-05-16]. https://arxiv. org/abs/1801. 05075v2. pdf.

[110] HOLZINGER A, CARRINGTON A, MÜLLER H. Measuring the quality of explanations: the System Causability Scale (SCS) [J]. KI - Künstliche Intelligenz, 2020, 34(2): 193-198.

[111] EDMONDS M, GAO F, LIU H X, et al. A tale of two explanations: enhancing human trust by explaining robot behavior [J]. Science Robotics, 2019, 4(37): No. aay4663.

This work is partially supported by National Natural Science Foundation of China (61703434).

LEI Xia, born in 1989, Ph. D. candidate. Her research interests include machine learning, optimal control.

LUO Xionglin, born in 1963, Ph. D., professor. His research interests include control theory, process control, chemical system engineering, machine learning.