计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：可解释人工智能研究综述
作者：赵延玉，赵晓永，王磊，王宁宁
网络首发日期：2023-03-14
引用格式：赵延玉，赵晓永，王磊，王宁宁．可解释人工智能研究综述[J/OL]．计算机工程与应用. https://kns.cnki.net/kcms/detail/11.2127.TP.20230313.1550.016.html

可解释人工智能研究综述

赵延玉[1]，赵晓永[1,2]，王磊[1]，王宁宁[1]

1.北京信息科技大学 信息系统研究所，北京 100129
2.北京信息科技大学 北京材料基因工程高精尖创新中心，北京 100129

摘 要：随着机器学习和深度学习的发展，人工智能技术已经逐渐应用在各个领域。然而采用人工智能的最大缺陷之一就是它无法解释预测的依据。模型的黑盒性质使得在医疗、金融和自动驾驶等关键任务应用场景中人类还无法真正信任模型，从而限制了这些领域中人工智能的落地应用。推动可解释人工智能（Explainable Artificial Intelligence , XAI）的发展成为实现关键任务应用落地的重要问题。目前，国内外相关领域仍缺少有关可解释人工智能的研究综述，也缺乏对因果解释方法的关注以及对可解释性方法评估的研究。因此，本文首先从解释方法的特点出发，将主要可解释性方法分为三类：独立于模型的方法、依赖于模型的方法和因果解释方法，分别进行总结分析，然后对解释方法的评估进行总结，列举出可解释人工智能的应用，最后讨论当前可解释性存在的问题并进行展望。

关键词：可解释性；人工智能；机器学习；深度学习；评估
文献标志码:A 中图分类号:TP301 doi：10.3778/j.issn.1002-8331.2208-0322

Review of Explainable Artificial Intelligence

ZHAO Yanyu[1], ZHAO Xiaoyong[1,2], WANG Lei[1], WANG Ningning[1]

1. Information Systems Institute, Beijing Information Science & Technology University, Beijing, 100129, China
2. Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science & Technology University, Beijing 100129, China

Abstract: With the development of machine learning and deep learning, artificial intelligence technology has been gradually applied in various fields. However, one of the biggest drawbacks of adopting AI is its inability to explain the basis for predictions. The black-box nature of the models makes it impossible for humans to truly trust them yet in mission-critical application scenarios such as healthcare, finance, and autonomous driving, thus limiting the grounded application of AI in these areas. Driving the development of explainable artificial intelligence (XAI) has become an important issue for achieving mission-critical applications on the ground. At present, there is still a lack of research reviews on XAI in related fields at home and abroad, as well as a lack of studies focusing on causal explanation methods and the evaluation of explainable methods. Therefore, this paper firstly starts from the characteristics of explanatory methods and divides the main explainable methods into three categories: model-independent methods,

基金项目：国家重点研发计划（2019YFB1705402）。
作者简介：赵延玉（1999-），女，硕士研究生，主要研究方向为模型的可解释性，E-mail：18551712062@163.com；**赵晓永**（1981-），男（通信作者），副教授，硕导，博士后，主要研究方向为数据科学与人工智能，E-mail：zhaoxiaoyong@bistu.edu.cn；**王磊**（1982-），女，副教授，硕导，博士，主要研究方向为服务科学与人工智能，E-mail：wanglei575882@163.com；**王宁宁**（1988-），女，副教授，博士，信息空间与人工智能，E-mail:wangningninggustb@163.com.

model-dependent methods, and causal explanation methods from the perspective of explanation types, and summarizes and analyzes them respectively, then summarizes the evaluation of explanation methods, lists the applications of explainable AI, and finally discusses the current problems of explainability and provides an outlook.

Key words：explainability; artificial intelligence; machine learning; deep learning; evaluation

在大数据时代，人工智能（Artificial Intelligence, AI）逐渐应用到各个领域，特别在语音识别、自然语言处理和计算机视觉等领域，这主要得益于以深度学习为代表的机器学习技术的发展和进步。

然而，随着人工智能技术的不断发展，我们研究和使用的模型越来越复杂化，从手工规则和启发式方法到线性模型和决策树再到集成和深度模型以及最近的元学习模型。尽管这些新方法更为精确，但它所构建的人工智能系统错综复杂，模型的透明度也越来越低，一般用户几乎无法理解其工作原理，进而限制了它们在现实任务中尤其是错误预测代价很高的领域的应用。例如在医疗诊断、金融服务和自动驾驶等对安全性有着极高要求的领域，在智能系统做出决策或提出建议时，是否会面临生命危险、服务灾难以及道德困境，这些因素形成了新的决策、治理和监管问题。要解决这些问题，就需要打开黑盒模型，可解释人工智能就是理解 AI 模型的有效技术。

目前有很多关于模型可解释性的研究方法，相应产生了一些综述研究。在 Adadi 等人[1]的研究中，与侧重于可解释性的具体解释维度不同，提倡研究领域的多学科性质，并从不同的角度介绍了可解释性的主要方面和应用领域。Zhang 等人[2]的研究主要集中在视觉可解释性，如特征可视化和热图等。而在最新的研究综述中 Zhang 等人[3]主要对神经网络的解释方法进行总结，并从参与类型、解释类型和解释范围三个维度对可解释性方法进行分类。苏炯铭等人[4]对卷积神经网络、循环神经网络、生成对抗网络等典型网络的解释方法进行分析梳理。化盈盈等人[5]从可解释性原理的角度对深度学习模型的解释方法进行分类。曾春燕等人[6]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。Speith[7]等人将可解释性方法分为基于功能的方法、基于结果的方法、概念方法和混合方法四类，并讨论了它们各自的优势和局限性。Minh[8]等人将 XAI 方法分为建模前可解释性、可解释性模型和建模后可解释性，并比较每种方法的优缺点及面临的挑战。

近年来，除了关注可解释人工智能方法的研究，如何评估和量化可解释性也是一个待研究的重要问题，因为不同解释方法在同一问题上的解释结果可能是不同的，或者同一种解释方法的解释结果是不稳定的[9]。IsLam 等人[10]提出了一种与模型无关的可解释性量化方法，并将领域知识纳入到模型中，使预测更容易解释两个不同的领域。在 Carvalho 等人[11]的研究中通过衡量可解释性属性的指标来评估可解释性。Mohseni 等人[12]提出了基于图像和文本的解释评估基准，该基准基于人类注意力的显著图进行定量评估，实验结果表明了该方法在定量评估模型解释方面的有效性。

虽然有关可解释人工智能的研究在不断深入，但研究成果依然存在局限性，目前对于可解释人工智能的研究综述较少，而且一些综述中未考虑到解释方法评估的重要性和从因果关系的角度对模型做出决策的原因进行分析总结。因此本文加入了对可解释人工智能的简单概述以及从解释方法的特点出发对当前的研究现状进行分类，主要归纳为独立于模型的解释方法、依赖于模型的解释方法和因果解释方法，并对不同方法进行优缺点比较。独立于模型的解释方法主要关注模型的输入和输出，依赖于模型的解释方法依赖模型的内部结构，因果解释方法从因果的角度解释模型做出决策的原因。同时总结现有的解释评估方法，并列举出可解释人工智能的具体应用，最后讨论可解释性当前存在的问题，展望未来的研究趋势与方向。

1 可解释人工智能概述

1.1 可解释人工智能的定义

可解释人工智能对于用户理解、信任和有效管理新一代人工智能系统至关重要。在 2015 年，美国国防部高级研究计算局（DARPA）制定了可解释人工智能（XAI）计划，其目标是使最终用户能够更好地理解和信任人工智能系统学习到的模型和决策[13]。我国于 2022 年提出了“可解释、可通用的下一代人工智能方法”重大研究计划，旨在打破现有深度学习“黑箱算法”的现状，建立一套可适用于不同领域、不同场景的通用方法体系[14]。

对于可解释性，不同的研究学者赋予不同的定义。

Gunning 等人认为可解释性是通过提供解释，使其行为更容易被人类解释[15]。Das 等人[16]将可解释性定义为衡量人类对机器学习模型做出的决策的背后原因的理解程度。Duval 等人[17]认为 XAI 系统能够解释 AI 系统做了什么，现在正在做什么以及接下来将发生什么。由上述不同学者的定义可以看出，对于可解释人工智能虽然难以形成统一的定义，但其核心目的都是为了帮助人类增强理解，改进预测。因此，本文从面向受众的解释目的出发，将 XAI 定义为："人类可以从 AI 系统获得有关其预测的决策依据和推断，从而打破人工智能的黑盒子，进而建立用户与 AI 系统之间的信任"，说明 XAI 是以用户需求为驱动的。如图 1 是 XAI 工作流程。

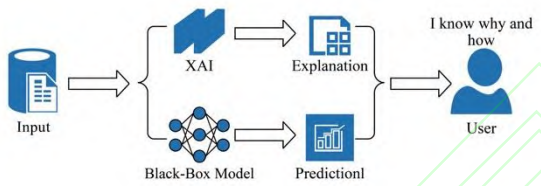

图 1　XAI 工作流程

Fig.1　XAI workflow

1.2　可解释人工智能的重要性

可解释人工智能的重要性可以总结为两个方面。

从应用研究角度来说，对于模型调试、检测偏差、人机协作、法规遵从等至关重要，特别是对于人类关键应用而言，让使用者理解、验证、编辑和信任模型非常重要[18]。

从知识发现的角度来说，可以为人们提供一种新的科学范式，模型本身也意味着知识，人们希望知道模型究竟从数据中学到了哪些知识从而产生最终的决策，进而可以将从模型学到的知识中再运用到模型的性能提升中。

1.3　不同视角下的 XAI 分类方法

基于业界现有研究成果，实现可解释人工智能的方法可以根据以下几个角度进行分类，如图 2 所示。

从模型相关性角度，可将其分为与模型无关的解释方法和特定于模型的解释方法[19]。与模型无关的解释方法通常是通过分析特征的输入和输出来解释模型决策的依据，可以应用于任何机器学习模型；特定模型的解释方法是指该方法仅限于特定的模型，每个解释方法都基于特定模型的内部结构。

从解释时机角度，分为内在自解释模型和训练后解释两种方法[20]。内在自解释模型是指通过训练结构简单、可解释性好的模型或将可解释性结合到具体的模型结构中使模型本身具备可解释能力；训练后解释是指模型训练后进行解释的方法，将解释与机器学习模型分开，具有很大的灵活性。

从解释范围角度，可以分为局部解释方法和全局解释方法[21]。全局可解释性是指帮助人们理解复杂模型背后的整体逻辑以及内部的工作机制；局部可解释性是指帮助人们理解机器学习模型针对每一个输入样本的决策过程和决策依据。

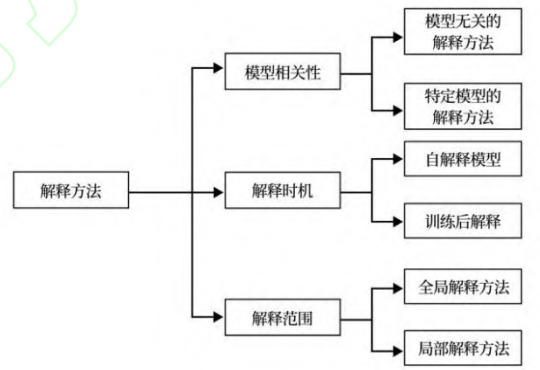

图 2　解释方法分类

Fig.2 Classification of Explanation Methods

2　可解释人工智能方法分类

本节主要对划分的三大类的可解释性研究方法进行分类阐述，表 1 展示了对三大类解释方法的细粒度划分。

表 1　解释方法细粒度分类

Table 1　Fine-grained classification of Explanation Methods

解释方法	细分方法	解释算法
独立于模型的解释方法	特征相关解释	部分依赖图[23]、个体条件期望[24]、累计局部效应[25]、特征交互[26]、置换特征重要性[27,28]、Shapley Values[29]
	基于样本的解释	对抗样本[30~32]、原型[33]、有影响力的样本[34,35]
	代理模型解释	全局代理模型[19]、LIME 及其变种[36~41]
依赖于模型的解释方法	自解释模型	线性回归[42]、逻辑回归[43]、决策树[44]、朴素贝叶斯[45]、RuleFIt[46]
	特定模型的解释	DeepLIFT[47]、LRP[48~50]、激活最大化[51,52]、基于梯度的方法[53,56]、类激活映射[57,60]
因果解释方法	-	反事实解释[61,69]

2.1 独立于模型的解释方法

独立于模型的解释方法将解释过程与底层的机器学习模型分离，可以应用于任何模型，其主要依赖于分析特征输入和输出，而不关注模型的内部细节，如权重或结构信息。此类方法的最大优点是松耦合，可以很方便地替换底层的机器学习模型或采用新的解释方法，且通常是在模型训练后进行解释，解释时不牺牲原模型的预测能力。此外不同类型的解释能产生特征表示和满足不同的信息需求，具有表示灵活性和解释灵活性[22]。

本小节将独立于模型的解释方法又细分为三种类型：特征相关解释、基于样本的解释和代理模型解释。

2.1.1 特征相关解释

在可解释性研究中，首先要回答是什么驱动了模型的预测，即哪些特征在模型的决策中发挥了重要作用。

特征相关解释方法主要有：部分依赖图(Partial Dependence Plot, PDP)、个体条件期望(Individual Condition Expectation, ICE)、累积局部效应(Accumulated Local Effect, ALE)、特征交互(Feature Interaction)和置换特征重要性(Permutation Feature Importance)等。

PDP[23]描述了特征是如何影响模型预测的，同时显示特征和标签之间是否具有线性相关性等。具体来说，先假定其他特征保持不变，然后将感兴趣的特征全部改成某个取值，利用训练模型对这些数据进行预测，求所有样本的预测的平均值，然后遍历该特征的其他不同取值，得到平均预测值。通过绘制 PDP 图，可以判定特征的贡献度，是一种全局解释方法。然而该方法的缺点是，当特征之间存在明显的线性关系时，对于原数据特征的修改会产生无意义的数据，输出不符合现实逻辑的决策结果。另外 PDP 只能反映特征变量的平均水平，而忽视了数据异质对结果产生的影响。

ICE[24]刻画的是每个个体的预测值与单一变量之间的关系，是一种局部解释方法。其原理是：对某一个体，保持其他变量不变，随机置换选定的特征变量的取值，黑盒模型输出预测结果，最后绘制该个体的单一特征变量与预测值之间的关系图。通过 ICE 图，可视化每个个体对每个特征的预测依赖关系，每个特征下的每个个体都会分别产生一条线，而部分依赖图是个体条件期望图的平均值，因此每个特征只有一条线。ICE 图避免了数据异质的问题，但它也只能反映单一特征变量与目标的关系。

由于 PDP 与 ICE 方法的前提都是待解释变量与其他变量之间相互独立，但在大多数场景下，都不能满足这样的前提。ALE[25]正是为了解决上述缺陷而提出。ALE 通过计算特征的条件分布来计算预测的差异，而不是预测的平均。首先，选定特征，取特征的上下界取值计算局部效应，从而解决了特征之间存在相关性的问题，预测值的差异被累积中心化，形成 ALE 曲线，是一种全局解释方法。但该方法实现复杂且不直观，当特征强相关时，解释困难。

特征交互方法[26]考虑了在预测模型中存在特征交互的情况。估计交互强度的一种方法是衡量预测的变化在多大程度上取决于特征的交互作用。这项衡量被称为 H 统计量。该方法是一种全局解释方法，统计信息无量纲，会检测特征之间的交互，其缺点是计算量大。

基于 Breiman[27]在随机森林中引入的位移特征重要性度量，Fisher 等[28]提出了与模型无关的特征重要性解释方法，将特征重要性表示为特征信息被破坏时模型误差的增加幅度，可同时解释主特征和特征交互作用对模型性能的影响，提供了对模型行为高度压缩的全局洞察力，是一种全局解释方法。置换特征重要性的基本思想是将待研究的数据打乱，然后观察预测的准确率变化了多少，根据变化量来决定特征重要性。相比于部分依赖图展示的特征如何影响模型预测，特征重要性展示的是哪些变量对预测的影响最大。然而该方法的缺陷是会受到数据质量的影响。

Shapley 值法最早由美国加州大学教授 Lloyd Shapley 提出，用于经济活动中的利益合理分配等问题。在多人合作中，每个成员的贡献都不一样，所以对应的利益分配也应该有差异。一个特征的 Shapley value 是该特征在所有的特征序列中的平均边际贡献。基于 Shapley 值进行联盟成员的利益分配体现了各盟员对联盟总目标的贡献程度，避免了分配上的平均主义，更具合理性和公平性[29]，该方法具备完整理论基础（有效性、对称性、虚拟性和可加性），可同时用于模型的全局和局部解释。然而随着特征数的增加，其计算复杂度也会急剧增加。

独立于模型的特征相关解释方法能够告诉用户对预测结果的判断依据，但这些方法的适用数据类型和数据量大小是有限的，因此这些解释方法更适合针对表格数据的前期可视化探索解释。表 2 是对上述方法的优缺点比较。

表 2 特征相关解释方法比较

Table 2 Comparison of Feature-Related Explanation Methods

方法	全局/局部	优点	缺点
部分依赖图 PDP	全局	与预测之间具有因果关系；容易实现	最大特征数量为 2；特征之间的独立性假设不一定成立，适应性较差
个体条件期望 ICE	局部	与 PDP 相比，ICE 更直观	ICE 图只能显示一个特征，绘制很多曲线时，图过于拥挤
累计局部效应 ALE	全局	在特征相关时仍然有效;计算速度比 PDP 更快；用于分类特征	实现复杂且不直观；当特征强相关时，解释困难
特征交互	全局	统计信息无量纲，会检测各种类型的交互	计算量大
置换特征重要性	全局	重要性度量会自动考虑与其他特征的所有交互	数据实例的质量不同会产生偏差
Shapley Value	局部，全局	具有扎实的理论基础；适用于任何具有局部可视化但计算开销高的不可知 ML 模型	大量的计算时间

2.1.2 基于样本的解释

基于样本的解释通过选择数据集的特定实例解释模型的行为或者底层数据分布。当以人类可以理解的方式表示数据实例时，基于样本的解释是十分有意义的。下面将详细介绍几种基于样本解释的方法。

（1）对抗样本

对抗样本是指攻击者通过向原始样本做出微小的扰动而使得整个模型的注意力发生转移，从而做出一个错误的预测。在一些关键系统的应用中，如何提升模型的可靠性，不让其被攻击欺骗，是一个值得深入研究的问题。目前大多数可解释性问题通常关注正常样本的预测进行解释和分析，而忽视了模型在现实场景中可能遇到的对抗样本并没有解释模型发生错误的原因,从而对安全关键的深度学习应用造成威胁[30]。

孔祥维等人[31]提出了基于深度神经网络模型可解释性的对抗样本防御方法，其过程是首先通过投影梯度下降法生成与原始图像对应的对抗样本图像，然后将原始样本与对抗样本都作为模型的输入，通过计算特征图权重分布和激活图进行训练后，最后将测试样本输入到模型中进行分类，以此来剔除对抗样本，从而实现对抗样本的防御。董胤蓬等人为了实现可解释的深度神经网络，利用对抗样本检验深度神经网络的内部特征表示[32]。

（2）原型样本

原型是指数据中具有代表性的样本。在数据中找到原型的方法有很多，例如，K-means 算法找到的各个簇的中心点就可以作为一种原型。

使用原型进行图像识别是解释黑盒深度学习模型的方法之一。在 Nauta 等人[33]的研究中，使用分类模型认为最重要的视觉特征的信息来增强原型，从而提高可解释性。具体来说，通过量化色调、形状、纹理的影响来帮助用户理解原型的含义，并且可以生成相应的全局和局部解释。使用该方法可以提高图像识别方法的可解释性，还能检查视觉上相似的原型是否具有相似的解释。

目前在可解释性方法中考虑到时间效率，采用基于原型的可解释方法可以大幅提升解释效率。

（3）有影响力的样本

在模型训练过程中，删除某一个训练数据点时，可能会对模型的参数造成较大的影响。因此，有影响力的样本解释的思路是从构建模型的数据出发来进行修改和测试,与在已有模型上修改特征值的方法不同，既可以考察单个数据点对全局模型参数的影响程度，也可以给定一个预测实例，来寻找哪个训练数据点对它影响最大。然而，该方法的缺点是为了找到有影响力的实例，需要依次移除数据点，重新训练模型来评估，计算量巨大。在 Koh 等人[34]的研究中，作者利用模型的梯度来建立影响函数，可以在不重新训练模型的情况下近似计算数据点对模型参数造成的影响。在最近的 Guo 等人[35]的研究中，通过 s-test，KNN 加并行计算实现了高效率的影响函数（Influence Functions）计算。通过影响函数可以理解模型行为并检测数据集中的错误。

在可解释人工智能研究中，对于开发人员来说，识别和分析有影响力的实例有助于发现数据问题，从而更好地调试模型以了解模型的行为。

2.1.3 代理模型解释

代理模型使用可解释的模型来模仿黑盒的行为，与原始模型相比，降低了复杂性，更容易实现。主要有两种代理模型解释方法，一种是全局代理模型，另一种是局部代理模型。

全局代理模型[19]通过训练一个可解释的模型来近似黑盒模型的预测。首先，使用经过训练的黑盒模型对数据集进行预测，然后针对该数据集和预测训练

可解释的模型，训练好的可解释模型成为原始模型的代理。但由于代理模型仅根据黑盒模型的预测而不是真实结果进行训练，因此全局代理模型只能解释黑盒模型，而不能解释数据。

LIME(Local Interpretable Model-Agnostic Explanations)[36]是一种与模型无关的局部解释方法，与全局代理方法相比，它训练可解释的模型来近似单个预测，而不对整个模型进行解释。该方法的主要思想是通过扰动输入观察模型的预测变化，根据这种变化在原始输入中训练一个线性模型来局部近似黑盒模型的预测，线性模型中系数较大的特征被认为对输入的预测很重要，从而实现由白盒模型去解释黑盒模型的局部。

LIME 方法的优点是它可以解释表格数据、图像和文本数据，但其缺陷是解释不稳定，在 Zhang 等人[37]的研究中证明 LIME 方法存在三种不确定性的来源，即抽样过程的随机性、随抽样接近性的变化以及不同数据点的解释模型可信度的变化。

针对 LIME 只能对单个样本的预测提供解释的问题，Ribeiro 等[36]人又提出了 SP-LIME，该方法挑选尽可能少的样本但让这些样本能够尽可能多的覆盖一些更重要的特征，通过这些样本用户能够全面的了解模型做出的预测更依赖于哪些特征。

为了改进LIME方法精度较低的缺点，Ribeiro于2018年又提出了基于LIME的改进方法Anchor[38]，它将特征和输出简化成 IF-Then 形式，此外它可以找到输入中最重要的部分，由分类器进行预测，输出更容易理解。

针对 LIME 方法的不稳定性，Zhou 等人[39]提出了 S-LIME，它基于中心极限定理的假设检验框架来保证结果稳定性所需要的扰动点的数量，因此 S-LIME 方法可以产生稳定的解释。Shankaranarayana 等人[40]通过使用自动编码器对 LIME 进行修改，赋予局部模型更好的权重函数，称作 ALIME，该方法不仅能提高稳定性，还能提高局部保真度。

除了以上方法，ElShawi 等人[41]提出了 ILIME，通过依赖一种解释机制来解释任何基于监督学习的模型的预测，该机制选择要解释的实例中最具影响的实例进行预测，该方法与 LIME 方法相比，准确性更高。

代理模型解释方法在模型的可解释性研究中是具有启发性的，尤其是针对 LIME 的拓展研究，任何模型得到的预测结果都能使用 LIME 及其变种方法来解释，但是其缺点也是明显的，在大规模应用中必须解决稳定性。表 3 总结了 LIME 相关解释方法的特点。

表 3 LIME 相关解释方法特点总结

Table 3 Summary of Characteristics of Interpretation Methods Relevant to LIME

方法	优点	缺点
LIME	可以解释表格、图像和文本数据	只能对单个样本的预测提供解释，且解释不稳定
SP-LIME	对多个样本进行解释，且所选择的样本是尽可能少的，还能覆盖重要特征	算法精度较低，用户需要全面了解模型
Anchor	特征和输出简化成 IF-Then 形式，是人们最容易理解的表示方法	参数较多
S-LIME	可以产生稳定的解释	不适用于时序数据
ALIME	不仅能提高稳定性，还能提高局部保真度	不适用于文本和图像数据
ILIME	选择最具影响力的样本进行预测，准确性更高	不适用于文本和图像数据

2.2 依赖于模型的解释方法

依赖于模型的解释方法适用于特定的模型，因为每个解释方法都基于模型的内部结构，包括传统的自解释模型和特定模型的解释方法。

2.2.1 自解释模型

自解释模型指模型本身是可解释的，即对于一个已经训练好的学习模型，无需额外信息就可以理解模型的决策过程或决策依据。然而，由于人类认知的局限性，自解释模型的内置可解释性受到模型复杂度的制约，这要求自解释模型结构不能过于复杂，通常采用结构简单、容易理解的模型。这也使得与其他机器学习模型相比，自解释模型的预测效果通常较差。传统的自解释模型包括线性模型、逻辑回归、决策树等。表 4 中总结了常见的自解释模型及其方法对比。

表 4 自解释模型方法总结

Table 4 Summary of Self-Explaining Model Approaches

模型	优点	缺点	任务
线性回归	结构简单，易于实现	难以建模复杂的非线性数据	回归

逻辑回归	模型训练速度快，解决共线性问题	准确率不高，无法处理非线性数据	分类
决策树	易理解和实现，可以捕获特征间的交互	节点越多，树越深，越难理解决策规则	分类、回归
朴素贝叶斯	有稳定的分类效率	在实际中特征独立性假设不成立	分类
RuleFit	考虑了特征之间的交互作用	可解释性随着特征数量的增加而降低	分类、回归

一般来说，自解释模型是人类最容易理解的模型，但由于其简单的结构会对模型的拟合能力造成限制，而且大多数场景下的问题存在特征交互等问题，并不能用简单的线性模型来表示。五类常见的自解释模型具体描述如下。

（1）线性回归

线性回归[42]是研究变量之间定量关系的表达式，其形式如下：

$$y=\omega_0+\omega_x x_1+\omega_2 x_2+\ldots+\omega_n x_n+\varepsilon \quad (1)$$

其中，ω 和 ε 分别表示权重与偏差，通过权重可以看出每个特征对预测的影响有多大，以及是正相关还是负相关，目标变量 y 是 n 个特征变量的权重和。

线性回归模型结构简单，对模型输出的结果具有可解释性。然而该方法对线性关系有很强的依赖性，线性关系越强，得到的结果才会越好，无法处理更复杂的关系，因此它的解释能力是有限的。

（2）逻辑回归

线性回归的输出是一个数值，用来解决回归问题，而对于分类问题，最直观的一个解决方法是通过设定一个阈值，去预测标签。结合 sigmoid 函数，线性回归函数把线性回归模型的输出作为 sigmoid 函数的输入，于是就变成了逻辑回归[43]模型：

$$P(Y=1)=\frac{1}{1+e^{-\left(\omega_0+\omega_1 x_1+\omega_2 x_2+\ldots+\omega_n x_n\right)}} \quad (2)$$

假设已经训练好一组权值，代入上述公式，就可以得到预测标签为 1 的概率，从而判断输入数据的类别。逻辑回归在训练过程中，训练速度快，解决共线性问题，提供模型的可解释性，但其缺点是准确率不高，无法处理非线性数据。

（3）决策树

决策树[44]是根据数据特征划分的一棵树，每个分支是由不同的特征划分得到的。使用决策树算法可以解决多分类问题。它的解释原理比较简单，首先从树的根结点开始，依次对样本的某一个特征进行判断，然后再根据边的判断，转到下一子集，直到走到叶节点，最后将样本分配到叶节点的类中。每一条从根节点到达叶节点的路径代表一条决策规则。

决策树具有天然的可解释性，但其缺点是随着深度增加，叶子结点的数量会急剧增加，解释性大大降低。

（4）朴素贝叶斯

朴素贝叶斯算法[45]是应用最为广泛的分类算法之一，它基于贝叶斯定义和特征条件独立性假设，先通过给定的训练集，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y。

该算法具有很好的可解释性，而且分类效果比较稳定。但其缺点是特征独立性假设在实际应用中往往是不成立的，当特征个数较多或者特征之间相互关联时，分类效果并不好。

（5）RuleFit

RuleFit 算法[46]是由 Friedman 等人于 2008 年提出，该算法巧妙的将决策树和线性模型结合在一起，可以将特征交互添加到线性回归模型中。该算法分为两步：首先通过决策树模型产生规则，再将原有特征和这些新产生的规则作线性组合，拟合出一个稀疏的线性模型。

由于这个算法的规则由决策树产生，因此也具有可解释性，且可以处理分类和回归任务。但其缺点是可解释性会随着特征数量的增加而降低。

2.2.2. 特定模型的解释

特定模型的解释主要基于模型的内部结构。如 2.2.1 所述，线性模型中的权重属于特定模型的一类解释。近年来，此类方法的研究主要集中于深度神经网络的可解释性。当模型成为知识的来源，模型能够提取到怎样的知识在很大程度上依赖于模型的组织架构、对数据的表征方式，对模型的解释可以捕获这些知识。这类解释方法主要包括深度学习重要特征(Deep

Learning Important Features, DeepLIFT)、层方向的关联传播(Layer-wise Relevance Propagation, LRP)、激活最大化(Activation Maximization, AM)、基于梯度的方法以及类激活映射(Class Activation Mapping, CAM)。本文将特定模型的解释方法总结在如表 5 所示。

表 5 特定模型的解释方法总结

Table 5 Summary of Interpretation Methods for Specific Models

解释方法	全局/局部	优点	缺点	适用场景
DeepLIFT	--	在梯度为零的情况下，仍可以进行传播	无法应用在 RNN 模型	解释神经网络
LRP	--	可用于多种数据类型，具有较高的解释质量	无法解决梯度饱和问题	解释神经网络
激活最大化	全局	深入理解 DNN 内部特征	只能用于优化连续性数据	解释连续数据模型
基于梯度的方法	局部	有效定位输入图像的关键决策特征	计算过程复杂，会产生梯度消失	解释神经网络
类激活映射	局部	有效定位物体	可视化结果中存在噪声	图像识别

（1）深度学习重要特征

DeepLIFT[47]解释方法通过将神经网络中的所有神经元对每个输入变量的贡献反向传播来分解对特定输入的神经网络的预测。其工作原理是将实际输入上的神经元激活与“参考”输入上的神经元激活进行比较，并反向传播重要性信号，所有输入特征的贡献总和等于输出激活与其参考值的差值。使用参考差异可以使信息即使在梯度为零的情况下，也可以进行传播。

目前该方法主要应用在解释深度神经网络 DNN，在循环神经网络 RNN 上的应用有待进一步研究。

（2）层方向的关联传播

LRP [48]算法假设神经网络各层的神经元与网络的输出之间存在一定的相关性，相关性越大的神经元对网络输出的影响越大，重要性也就越高。该方法的主要目的是找出输入图像的单个输入像素对分类器预测的贡献。该方法利用深度泰勒分解技术，通过预先训练的网络，将输出的相关性向后分配，并确定节点对分类的贡献。根据激活度和网络权值，通过在下一层传播相关性来获得每层的相关性。解释器给出了与输入图像具有相同维数的像素级热图，从而可视化了输入图像中对所选类别有贡献的重要区域[49]。

该方法为模型预测提供内在的解释，可以更好地理解深度神经网络的推理过程，目前可用于多种数据类型，如图像、文本、信号和音频等[50]。

（3）激活最大化

激活最大化（Activation Maximization）[51]用来可视化每个神经元的输入偏好，即找到怎样的输入能够最大程度地激活在特定层的特定神经元。它可以帮助我们深入理解深度神经网络的内部特征，该方法的核心思想是通过迭代地调整输入，使得神经元的激活最大，从而可视化神经元学到的特征。激活最大化方法表述如下：

$$x^{*}=arg\ \max_{I\ s.t.}\alpha_{i,j}(\theta,x) \qquad (3)$$

其中$\alpha_{i,j}$，表示第 i 层第 j 个卷积的激活，是关于输入 x 的函数，θ 表示一个训练过的深度神经网络（Deep neural network，DNN）的参数，包括权重和偏差。通过梯度上升法可以找到一个局部最优解x^{*}。

通过激活最大化我们可以知道，模型是通过哪些特征作出预测的。Erhan 等人[52]提出的激活最大化方法是通过生成一个输入图像可以最大化一个特定滤波器激活，并通过生成的图像显示滤波器学习到的特征。

激活最大化方法可以帮助人们理解 DNN 内部的工作逻辑，其解释结果更准确，能反映待解释模型的真实行为。然而该方法无法直接应用于文本、图等离散型数据模型，因此难以直接用于解释自然语言处理模型和图神经网络模型。

（4）基于梯度的方法

基于梯度的方法通过将目标类别上的决策分数往原始图像求梯度得到。梯度是一种局部信息，可以反映局部位置的扰动对于最终预测的影响，通过计算输出对输入的梯度，得到特征重要性，解释影响预测结果的关键特征依据。卷积神经网络（Convolutional Neural Network, CNN）模型的可解释性方法主要有四种：反卷积(Deconvolution)、导向反向传播(Guided Backpropagation)、平滑梯度(Smooth Gradients)和积分梯度(Integrated Gradients)。

解释卷积神经网络的工作原理，就需要剖析 CNN 的每一层究竟学习到了什么。反卷积方法对 CNN 的

中间层进行了可视化。具体来说，通过 CNN 可以得到有关输入的各层的特征表示，而为了验证某层一个具体的特征值，将该层的特征值之外的所有值设置为零之后，将其作为反卷积网络的输入，经过反卷积每一层的操作后，特征值被映射回了输入图片的像素空间，然后说明图片中的哪些像素参与激活了该特征值[53]。

反卷积（Deconvolution）和导向反向传播（Guided Backpropagation）都是从 CNN 模型中较深的卷积层中学习特征，其基础都是反向传播，唯一的区别在于反向传播过程中经过 ReLU 层时对梯度的处理策略不同。反卷积中，保留了输出大于 0 的位置的信息。导向反向传播[54]是在普通的梯度反向传播的基础上，对卷积层的负梯度进行过滤操作，使得梯度仅保留与网络输出正相关的，并将最终每个输入样本得到的梯度作为其重要性的解释结果。使用反卷积可以大致看清目标的轮廓，但是有大量噪声在目标以外的位置，使用导向反向传播噪声较少，特征较为明显的集中在目标上，但是会出现梯度消失现象。

平滑梯度（Smooth Gradients）[55]的核心思想是选择输入目标图像，通过将噪声添加到图像中来对相似的图像进行采样，然后对这些图像生成的梯度图进行平均，最终实现图像的平滑和去噪，得到更好的可视化效果。

积分梯度（Integrated Gradients）[56]的提出是为了解决梯度饱和问题。由于神经网络是非线性的，当像素或特征增强到一定程度后可能会对网络决策的贡献达到饱和。积分梯度对各输入变量的梯度值进行积分，可以得到非饱和区非零梯度对决策重要性的贡献。

平滑梯度和积分梯度方法都是通过梯度图的平均生成显著图，但这两种方法的缺点是过程复杂，需要进行多次迭代，是一个很耗时的过程。4 种方法的优缺点对比如表 6 所示。

表 6　基于梯度的解释方法的优缺点比较

Table 6　Comparison of advantages and disadvantages of gradient-based interpretation methods

方法	优点	缺点
反卷积	可视化 CNN 内部学习到的特征	目标以外有噪声，不能解释分类结果
导向反向传播	基本没有噪声	梯度消失现象，不能解释分类结果
平滑梯度	可视化效果好	过程复杂，多次迭代很耗时
积分梯度	可以解决梯度饱和的问题	

基于梯度的解释方法可以有效定位输入图像的关键决策特征，但是大多数方法直接使用输出对输入的梯度作为特征重要性会遇到梯度消失的问题，而且该方法不能解释分类结果，其可视化效果有待进一步提升。

（5）类激活映射

类激活映射利用卷积单元的定位能力对卷积神经网络进行解释，通过生成类激活图来可视化 CNN 的关注区域。传统卷积神经网络通过全连接层后，由 SoftMax 分类。类激活映射[57]采用全局平均池化（Global Average Pooling, GAP）替换全连接层。通过 GAP 将最后一个卷积层的特征图转换为特征向量，最终将多个特征图按照权重叠加在一起，生成加权特征图，并叠加在原始图片上，通过深浅不同的像素点表示对分类的贡献，从而让我们知道模型是通过哪些像素点得到图片属于某个类别。

虽然该方法使用 GAP 可以减少参数，防止过拟合，但存在的缺陷是对原模型的结构进行了修改，导致需要重新训练模型，在实际应用中会很耗时，限制了使用场景，同时重训练模型做出的解释结果与待解释模型的真实行为之间存在一定的不一致性。

Grad-CAM[58]方法解决了 CAM 的缺陷，与其他方法不同，它使用最后一层的卷积层的梯度信息来理解每个神经元对于目标决策的重要性，即用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，最后通过显著性图的方式进行可视化，突出显示图像的重要区域。Grad-CAM 不需要修改模型结构或重新训练模型，而且它的可视化可以进一步区别分类，提高了分类器的可信赖性，可应用于不同的 CNN 解释，但存在梯度不稳定的问题。

为了应对某一分类的物体在图像中不止一个的情况，Chattopadhyay 等人[59]又进一步提出了 Grad-CAM++，其基本形式与 Grad-CAM 相同，仅使用的通道权重不同，但它的定位更准确，解释性更好，改善了同类多目标图像的可视化效果。

为了不修改模型结构，类激活映射的另一类研究方法是不依赖梯度的方式去计算权重。在 Wang 等人[60]的研究中发现，基于梯度的 CAM 方法除了标记目标物体之外，对于背景信息也会标记，而这可能与梯度有关，因此提出了无梯度方法 Score-CAM。该方法首先提取特征图，然后对特征图进行上采样作为掩码信息，重新得到模型对于图片在目标类别上的响应值，最后将特征图与得到的响应值线性加权求和，得到最终可视化结果。Score-CAM 可视化结果更为聚焦，背景中的噪声大量减少，在定位多个同类对象时也表现出更好的性能。4 种方法的优缺点对比如表 7 所示。

表 7 基于类激活映射方法的优缺点对比

Table 7 Comparison of advantages and disadvantages of CAM methods

方法	优点	缺点
CAM	有效减少参数，防止过拟合	需要修改原模型结构
Grad-CAM	可以应用于不同的卷积神经网络	梯度不稳定
Grad-CAM++	适用于同类多目标检测	有大量背景信息被标记
Score-CAM	无梯度方法，可视化效果好	-

与基于梯度的解释方法相比，类激活映射可以解释分类结果，尤其在同类多目标检测任务中定位更准确，解释性更好，但该方法的可视化效果中依然存在噪声。

2.3 因果解释方法

因果解释方法解释用不同的输入、参数训练模型时，模型会做出什么决定，因此对用户来说是更友好的。Pearl 等人[61]将可解释性分为三个层次，以及在每个层次上可以回答的特征问题，分别是统计相关的解释、因果干预的解释和基于反事实的解释，并指出反事实解释是实现最高层次可解释性的方法，因为它包含了介入和关联问题。

反事实解释是一种基于实例的因果解释方法，它并没有明确回答模型为什么会作出这样的决策，主要侧重的是“改变什么条件，模型的决策结果会发生改变”。例如，在一个信用贷款案例中，如果一个人的贷款申请被拒绝了，我们想知道的不仅是为什么会被拒绝，还想了解改变什么条件，申请会被批准，这对申请人来说是更有效的。因此反事实解释可以作为最终用户的推荐，实现他们想要的输出。

如何定义一个好的反事实解释是诸多学者正想解决的关键问题。Wachter[62]提出解释应该是接近真实世界和原始数据样本的，因此需要利用距离来度量最小扰动，并将反事实解释的优化问题定义为：

$$\underset{x'}{arg\ min}\ \underset{\lambda}{max}\ \lambda\left(f(x')-y'\right)^2+d(x,x') \quad (4)$$

在 Watcher 提出的反事实解释方法中将所有特征视为同等重要，从而导致生成的反事实会改变大量特征，而且生成的反事实不易理解，因此 Grath 等人[63]提出每个特征改变的能力和幅度是不同的，提出利用 k 近邻和全局特征重要性两种加权策略以生成反事实解释。

为实现对分类特征的有效处理，在 Mothial 等人[64]的研究中，对连续特征和分类特征分别进行了定义，将连续特征定义为 CF 示例与原始输入之间的 L1 距离的平均值，分类特征则通过判断特征是否相同，分配距离为 1 或 0。在 Russell 等人[65]的研究中提出使用混合整数规划来为复杂数据类型生成反事实，对于分类特征采用 One-Hot 编码，然后利用编码数据点训练分类模型。

反事实解释除了应用在表格数据上，在图像数据上也表现出了良好的性能。Liu 等人[66]利用生成模型对图像进行干预操作，反事实解释确定图像的哪些方面可以作出改动，以便分类器输出不同的类别，或者为修改后的图像提供更精确的分类。

考虑到优化损失函数需要大量的调参工作，Looveren 等人[67]提出了使用原型来指导反事实生成，该方法不仅加快了反事实的搜索过程，而且在表格和图像数据上都表现出了良好的性能。

表 8 反事实解释方法总结

Table 8 Summary of Counterfactual Explanation Methods

方法	应用	接近性	稀疏性	多样性	因果约束
[62]	表格	L1-norm	✓	×	×
[63]	表格	L1-norm	✓	×	×
[64]	表格	L1-norm	✓	✓	×
[65]	表格	L1-norm	✓	✓	×
[66]	图像	-	✓	×	×
[67]	表格/图像	L1/L2-norm	✓	×	×
[68]	表格	L2-norm	✓	×	✓
[69]	表格	L2-norm	✓	×	✓

在保留因果约束的反事实研究中，学者们提出的方法主要分为两类：一类是通过因果结构模型来建立

因果损失函数[68,69]。然而在实际应用场景中，变量的确切函数因果机制通常是未知的；另一类是先通过模型生成反事实，然后再通过用户反馈对模型进行微调或通过因果知识、领域知识对反事实进行过滤操作[69]。

在表 8 中对上述反事实解释方法进行了细粒度的比较，从中可以得出，反事实研究虽然在不断深入，但大多数研究方法都未充分考虑特征取值范围的领域约束、特征之间的因果关系以及反事实的易用性等问题，如何提出一种稳定、有效、易用的反事实解释方法，仍有待进一步研究。

3 可解释人工智能方法的评估

目前可解释性研究领域的大部分工作集中在开发新的方法或技术去提高预测任务的可解释性，很少有对可解释性方法的评估工作，导致不同类型的解释方法之间的可解释性难以比较。在本研究中我们将可解释性方法的评估分为定性评估和定量评估两类。

3.1 定性评估

定性评估方法依靠人的视觉感观来评价解释结果是否符合人的认知，该方法简单清晰，易于理解，对用户来说是更友好的。对于 XAI 方法的定性评估，主要从解释的基本单位和可视化效果两个方面进行评估。

解释的基本单位[70]指解释是由什么组成的，可以是特征重要性值、训练集的实例，规则列表和像素等。因此对于解释的基本单位我们想要知道其构成形式与数量。如果解释是由特征组成的，那么它是包含所有特征还是仅包含少数特征。如果提供的是反事实解释，那么提供的解释与决策结果之间是否具有因果关系以及反事实结果对用户来说是否可行。

目前对于深度神经网络，更多采用可视化的方式进行解释，即构建显著性图来突出显示最相关的信息，从而提供可解释性。因此对于可解释性的定性评估方法，主要从可视化效果方面考虑，包括两方面的问题：一方面是观察目标区域的集中度和覆盖面，显著性图需要重点关注感兴趣的目标区域，而忽略其它不相关区域。显著性图中突出的区域对感兴趣目标的覆盖越全面，表明可视化效果越好。另一方面是考查多目标的可视化效果。当多个同一类别的目标同时出现在图像中时，可视化方法能够同时定位多个目标，而不会遗漏其中的某个目标。

3.2 定量评估

依靠人的主观度量评估模型的可解释性方法是不够的，定量评估为比较不同的解释提供了一种客观的方法。本文将可解释性定量评估方法分为两类：

一类是无特定方法的一般评估指标。

常用的指标包括[19]：①保真度，衡量解释对黑盒模型预测的近似程度；②一致性，衡量相同任务产生的相似模型的解释结果是否会有差异；③稳定性，衡量特征的微小变化是否会改变解释；④完整性，衡量解释所包含的实例数所覆盖的解释范围。

另一类是特定方法的评估指标。

对于特征交互解释方法，Molnar 等人[71]提出了三种度量指标：特征的数量、特征之间的交互强度和特征的主效应复杂性来衡量用于解释的事后模型的复杂性。对于扰动的解释方法，Yeh 等人[72]提出了失真度和敏感性两个度量指标去评估模型的可解释性。Srinivas 等人[73]提出了像素扰动定量评价指标，通过删除 k 个最不突出的像素后衡量分类器输出的变化。对于因果解释方法的评估，目前的因果解释方法大多是基于反事实的解释，因此这类方法的评估是通过衡量生成反事实解释的好坏来衡量。常用的评估指标包括[74]：①稀疏性，衡量反事实的大小；②数据流性相近性，衡量反事实解释是否接近模型训练数据分布；③接近度，衡量对反事实样本的模型预测是否接近于预定义的输出；④多样性，为数据实例生成的反事实解释应该彼此不同。Mothilal 等人[64]通过在反事实样本和原始输入样本上训练一个模型来预测新输入的类别，并比较该模型与原始模型的模拟程度，模拟程度越高越高则反事实解释的效果越好。在表 8 中总结了现有的反事实解释方法应用的评价数据集，可以看出反事实方法主要应用在图像和表格数据集上。因此对于不同的数据类型，评估方法应该是有差异的，对于表格数据较小的反事实是更容易解释的，对于图像数据，对像素的扰动应该是尽可能小的。

对于可解释性的评估方法而言，未来我们一方面要从从应用场景、人类认知等角度来设计定性的评估指标，另一方面需要从解释结果的一致性、稳定性以及不同解释方法的差异性等角度设计评价指标，对解释方法进行综合评估。总之，模型的解释是否可靠真实并且能否提高用户的决策能力始终是评估方法需要聚焦的问题。

4 可解释人工智能的应用

可解释人工智能发展的目的是落地到具体的 AI 应用场景。目前可解释性已经逐渐应用于医疗诊断、欺诈检测、自动驾驶和智能化运维等领域。

4.1 医疗诊断

随着 AI 的智能化，人们开始研究将 AI 应用在医

疗领域。尤其在医学影像分析方面可以借助 AI 精确的搜索能力和定位能力，从而实现提升临床诊断效率与准确率的目的。然而，医患关系建立在信任的基础上，没有可解释的人工智能，医生无法信任模型，也很难与患者进行沟通[75]。

在 Sappagh 等人[76]的研究中，开发了一个利用随机森林作为分类器的两层模型，该模型可以将阿尔兹海默症症状诊断和进展检测两个过程结合在一起，并使用 Shapley 特征归因框架对分类器的每一层提供全局和基于实例的解释，帮助临医生增强了对疾病的临床决策。Lamy 等人[77]在乳腺癌数据集研究中采用基于案例推理的方法提供视觉解释或视觉推理，临床医生通过可视化界面可以查询到不同患者的病例为什么是相似的，增强临床验证。

4.2 欺诈检测

伴随着金融服务规模的快速增长，也带来了各种各样的问题，金融欺诈案件层出不穷。人工智能模型在检测金融欺诈方面表现出良好的性能，然而这些模型仍然缺乏可解释性。

Lin[78]等人研究了风险检测模型的可解释性，利用 Shapley 值评价每个金融特征对预测结果的边际贡献，并提出 group Shap 方法来评估公司的不同能力，从而为风险管控部门提供风险提示，对投资管理和客户监控具有重要价值。Wang[79]等人认为用户丰富的交互数据形成了一个大型的图网络，而数据集中的数据少部分有标签，大部分无标签，因此提出半监督注意力图神经网络对图中标记和未标记的数据进行检测，同时注意力机制使模型具有可解释性，告诉检测部门哪些是欺诈的重要因素以及用户被预测为欺诈的原因。

4.3 自动驾驶

近几年，自动驾驶依靠人工智能、视觉计算、雷达和全球定位系统等技术的发展在研发方面取得了较大突破。然而，在目前的技术水平的限制下，自动驾驶汽车还不能真正投产使用。一方面，已有的城市交通基础设施以及交通法规还不能适应自动驾驶汽车的驾驶需求。另一方面，自动驾驶汽车的安全和实时决策还有待进一步验证，人类还不能理解自动驾驶汽车所作出的智能决策。

Kim 等人[80]提出使用视觉注意力模型来突出显示可能影响网络输出的图像区域，然后采用因果过滤步骤来确定哪些输入区域会真正影响输出，该方法产生了更简洁的视觉解释。为了自动驾驶汽车具有可解释性的高精度模型，Scores[81]提出了一种新的自组织神经模糊方法，该方法可以对不同自动驾驶条件下发生的不同动作状态进行分类，所提出的方法能够以 IF...THEN 的方式向人类提供解释。

4.4 智能化运维

对于很多企业来说，随着业务体量的增加，现有的业务逻辑也越来越复杂，一些小概率的故障出现得越来越频繁，从而带来了极大的人力消耗。虽然现在提出了自动化运维，但也需要专业人员的大量的排错经验来作支撑，而且避免不了再次发生同样的故障。然而运用智能化运维，可以通过 AI 技术作出充分判断，对故障信息进行直接的溯源，并第一时间向运维人员展示故障的根本原因及定位，极大提升工作效率，也大大节省了处理故障的时间。

Ahmed 等人[82]研究了针对蜂窝服务提供商端到端性能下降的检测与定位，当模型检测到每个端到端的实例性能下降后，使用关联规则挖掘技术来定位性能下降的原因，实现了根因定位。Li 等人[83]提出了一种可操作且可解释的方法 DejaVu 用于在线服务系统中的重复故障定位检测。该方法以系统中的历史故障和依赖关系作为输入，离线训练模型，对于即将发生的故障，使用模型在线检测故障发生的位置和发生的故障类型，并使用全局和局部解释方法进行解释。

5 研究局限与展望

尽管可解释人工智能已取得一定的研究成果，但仍处于初级阶段，许多关键问题亟待解决。基于对上述研究的梳理和分析，从学术研究和业务应用角度对可解释人工智能面临的挑战进行讨论分析与展望。

从学术研究的角度，总结为以下问题：

（1）解释方法结合。现有的研究较少关注如何将不同的解释方法结合起来，未来可以考虑将不同的解释方法结合在一起，如正反结合，事实解释侧重于“为什么”，反事实解释侧重于“怎么做”，构建更为强大的模型揭示方法。

（2）可靠性与稳定性。现有的一些解释方法是不可靠、不稳定的。如 LIME，两个非常接近的点可能会导致两种截然不同的解释。因此，解释性算法还有赖于可靠的理论基础。未来解释方法需要经过 AI 专家的认可，确保算法的内在可靠性。

（3）知识驱动。随着深度学习与知识图谱等技术的深度融合，利用数据中的因果与逻辑关系，助力人工智能朝着认知智能的方面发展，例如，反事

实解释研究目前缺乏因果约束，导致生成的反事实对用户来说是不可行的。因此，未来可以考虑在生成反事实之前利用领域知识增强特征之间的因果约束。

（4）评价体系。对 XAI 方法还没有一个统一的评价体系。虽然目前有研究从定性和定量两个角度进行评估，但由于定性评估带有主观性和不可控性，而定量分析也没有达到相应的预期。究其原因，是由于决策者对于不同的决策任务有不同的理解和要求。因此，本文认为未来需要根据不同的应用领域进行定义。

从业务应用的角度，总结为以下问题：

（1）个性化解释。目前在可解释性研究中，往往忽略了向谁解释、为谁解释、何时解释，不同用户、不同应用场景对可解释性的需求是不同的。正如在关键应用系统中，如果不能理解不同场景下 AI 的决策行为，一旦发生事故，就会极大损坏用户的利益。对于系统开发者来说，AI 的可解释性可以帮助他们在系统出现问题时更加及时精确的找到问题的根源。对于企业来说，深入了解 AI 作出决策的原理有利于保证决策的公平性，维护公司和品牌的利益。对于个人来说，了解系统作出决策的原因可以帮助他们改进某方面的劣势，促进决策成功的进程。

（2）工程落地。目前的可解释人工智能方法的设计主要集中在对模型的研究，缺乏在实际 AI 应用产品或应用场景中如何融合和适应的经验。如何融合可解释人工智能与实际 AI 复杂应用系统仍需要深入探讨。

6 总结

可解释人工智能是一个重要的研究领域，本文对可解释人工智能从基本概述、方法分类、方法评估、应用分析几个方面进行了深入阐述，并分析了当前研究存在的局限性，展望其未来的研究方向。构建可信、可靠、可解释的人工智能是各行各业讨论的焦点，在技术同质化的情况下，系统性创新可以有效推动复杂场景落地，并有效提升人工智能走向可信人工智能的能力。

参考文献：

[1] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)[J]. IEEE access, 2018, 6: 52138-52160.

[2] Zhang Q, Zhu S C. Visual interpretability for deep learning: a survey[J]. Frontiers of Information Technology & Electronic Engineering, 2018, 19(1): 27-39.

[3] Zhang Y, Tiňo P, Leonardis A, et al. A survey on neural network interpretability[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2021.

[4] 苏炯铭,刘鸿福,项凤涛,吴建宅,袁兴生.深度神经网络解释方法综述[J].计算机工程,2020,46(09):1-15.
SU Jiongming, LIU Hongfu, XIANG Fengtao, WU Jianzhai, YUAN Xingsheng. Survey of Interpretation Methods for Deep Neural Networks [J]. Computer Engineering, 2020, 46 (09): 1-15.

[5] 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(03):1-12.
HUA Yingying, ZHANG Daichi, GE Shiming. Research progress in the interpretability of deep learning models [J]. Journal of Cyber Security, 2020,5(3):1-12.

[6] 曾春艳,严康,王志锋,余琰,纪纯妹.深度学习模型可解释性研究综述[J].计算机工程与应用,2021,57(08):1-9.
ZENG Chunyan, YAN Kang, WANG Zhifeng, YU Yan, JI Chunmei. Survey of Interpretability Research on Deep Learning Models[J].Computer Engineering and Applications,2021,57(08):1-9.

[7] Speith T. A review of taxonomies of explainable artificial intelligence (XAI) methods[C]//2022 ACM Conference on Fairness, Accountability, and Transparency. 2022: 2239- 2250.

[8] Minh D, Wang H X, Li Y F, et al. Explainable artificial intelligence: a comprehensive review[J]. Artificial Intelligence Review, 2022, 55(5): 3503-3568.

[9] 李瑶,左兴权,王春露,黄海,张修建.人工智能可解释性评估研究综述[J].导航定位与授时,2022,9(06):13-24.
LI Yao, ZUO Xingquan, WANG Chunlu, HUANG Hai, ZHANG Jianjian. Research Progress of Artificial Intelligence Interpretability Evaluation[J]. Navigation positioning and timing, 2022,9 (06): 13-24.

[10] Islam S R, Eberle W, Ghafoor S K. Towards quantification of explainability in explainable artificial intelligence methods[C]//The thirty-third international flairs conference. 2020.

[11] Carvalho D V, Pereira E M, Cardoso J S. Machine learning interpretability: A survey on methods and metrics[J]. Electronics, 2019, 8(8): 832.

[12] Mohseni S, Block J E, Ragan E. Quantitative evaluation of machine learning explanations: a human-grounded benchmark[C]//26th International Conference on Intelligent User Interfaces. 2021: 22-31.

[13] DW, Gunning D. Aha. "DARPA’s explainable artificial intelligence program." AI Mag, 40.2 (2019): 44.

[14] 可解释、可通用的下一代人工智能方法重大研究计划 2022 年度项目指南[J].模式识别与人工智能,2022,35(05):481-482.
Major Research Program on Interpretable and Generalizable Next-Generation Artificial Intelligence Methods 2022 Annual Project Guide [J]. Pattern Recognition and Artificial Intelligence,2022,35(05):481-482.

[15] Gunning D, Stefik M, Choi J, et al. XAI—Explainable artificial intelligence[J]. Science Robotics, 2019, 4(37): eaay7120.

[16] Das S, Agarwal N, Venugopal D, et al. Taxonomy and survey

of interpretable machine learning method[C]//2020 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2020: 670-677.
[17] Duval A. Explainable artificial intelligence (XAI)[J]. MA4K9 Scholarly Report, Mathematics Institute, The University of Warwick, 2019: 1-53.
[18] Tan S , Caruana R , Hooker G , et al. Learning Global Additive Explanations for Neural Nets Using Model Distillation:, 10.48550/arXiv.1801.08640[P]. 2018.
[19] Molnar C. Interpretable machine learning[M]. Lulu. com, 2020.
[20] Fan F L, Xiong J, Li M, et al. On interpretability of artificial neural networks: A survey[J]. IEEE Transactions on Radiation and Plasma Medical Sciences, 2021, 5(6): 741-760.
[21] Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (xai): A survey[J]. arXiv preprint arXiv: 2006.11371, 2020.
[22] Ribeiro M T, Singh S, Guestrin C. Model-agnostic interpretability of machine learning[J]. arXiv preprint arXiv: 1606.05386, 2016.
[23] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.
[24] Goldstein A, Kapelner A, Bleich J, et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation[J]. journal of Computational and Graphical Statistics, 2015, 24(1): 44-65.
[25] Galkin F, Aliper A, Putin E, et al. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects[J]. BioRxiv, 2018: 507780.
[26] Friedman J H, Popescu B E. Predictive learning via rule ensembles[J]. The annals of applied statistics, 2008, 2(3): 916-954.
[27] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.
[28] Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultane- ously[J]. J. Mach. Learn. Res., 2019, 20(177): 1-81.
[29] Sundararajan M, Najmi A.The many Shapley values for model explanation[C]//International conference on machine learning. PMLR, 2020: 9269-9278.
[30] Yuan X, He P, Zhu Q, et al. Adversarial examples: Attacks and defenses for deep learning[J]. IEEE transactions on neural networks and learning systems, 2019, 30(9): 2805- 2824.
[31] 孔祥维, 杨浩. 一种基于深度神经网络模型可解释性的对抗样本防御方法: CN112364885A[P]. 2021.
KONG, Xiangwei, YANG, Hao. An adversarial sample defense method based on deep neural network model interpretability:,CN112364885A [P]. 2021.
[32] 董胤蓬, 苏航, 朱军. 面向对抗样本的深度神经网络可解释性分析. 自动化学报, 2022, 48(1): 75−86.
Dong Yin-Peng, Su Hang, Zhu Jun. Interpretability analysis of deep neural networks with adversarial examples. Acta Automatica Sinica, 2022, 48(1): 75−86.
[33] Nauta M, Jutte A, Provoost J, et al. This looks like that, because... explaining prototypes for interpretable image recognition[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2021: 441-456.
[34] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]//International conference on machine learning. PMLR, 2017: 1885-1894.
[35] Guo H, Rajani N F, Hase P, et al. Fastif: Scalable influence functions for efficient model interpretation and debugging[J]. arXiv preprint arXiv:2012.15781, 2020.
[36] Ribeiro M T, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier[C]// Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 1135-1144.
[37] Zhang Y, Song K, Sun Y, et al. " Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations[J]. arXiv preprint arXiv:1904.12991, 2019.
[38] Ribeiro M T, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
[39] Zhou Z, Hooker G, Wang F. S-lime: Stabilized-lime for model explanation[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 2429-2438.
[40] Shankaranarayana S M, Runje D. ALIME: Autoencoder based approach for local interpretability[C]//International conference on intelligent data engineering and automated learning. Springer, Cham, 2019: 454-463.
[41] ElShawi R, Sherif Y, Al-Mallah M, et al. ILIME: Local and global interpretable model-agnostic explainer of black-box decision[C]//European Conference on Advances in Data- bases and Information Systems. Springer, Cham, 2019: 53-68.
[42] Haufe S, Meinecke F, Görgen K, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging[J]. Neuroimage, 2014, 87: 96-110.
[43] Bertsimas D, King A. Logistic regression: From art to science[J]. Statistical Science, 2017: 367-384.
[44] Quinlan J R. Learning decision tree classifiers[J]. ACM Computing Surveys (CSUR), 1996, 28(1): 71-72.
[45] Webb G I, Keogh E, Miikkulainen R. Naïve Bayes[J]. Encyclopedia of machine learning, 2010, 15: 713-714.
[46] Friedman J H, Popescu B E. Predictive learning via rule ensembles[J]. The annals of applied statistics, 2008: 916-954.
[47] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences[C]// International conference on machine learning. PMLR, 2017: 3145-3153.

[48] 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): e0130140.

[49] Wang Z, Huang X, Yang J, et al. Universal Adversarial Perturbation Generated by Attacking Layer-wise Relevance Propagation[C]//2020 IEEE 10th International Conference on Intelligent Systems (IS). IEEE, 2020: 431-436.

[50] Montavon G, Binder A, Lapuschkin S, et al. Layer-wise relevance propagation: an overview[J]. Explainable AI: interpreting, explaining and visualizing deep learning, 2019: 193-209.

[51] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[J]. arXiv preprint arXiv:1312. 6034, 2013.

[52] Erhan D, Bengio Y, Courville A, et al. Visualizing higher-layer features of a deep network[J]. University of Montreal, 2009, 1341(3): 1.

[53] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer, Cham, 2014: 818-833.

[54] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net[J]. arXiv preprint arXiv:1412.6806, 2014.

[55] Smilkov D, Thorat N, Kim B, et al. Smoothgrad: removing noise by adding noise[J]. arXiv preprint arXiv:1706.03825, 2017.

[56] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks[C]//International conference on machine learning. PMLR, 2017: 3319-3328.

[57] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

[58] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.

[59] Chattopadhay A, Sarkar A, Howlader P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks[C]//2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018: 839-847.

[60] Wang H, Wang Z, Du M, et al. Score-CAM: Score- weighted visual explanations for convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 24-25.

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

[62] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR[J]. Harv. JL & Tech., 2017, 31: 841.

[63] Grath R M, Costabello L, Van C L, et al. Interpretable credit application predictions with counterfactual explana- tions[J]. arXiv preprint arXiv:1811.05245, 2018.

[64] Mothilal R K, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations[C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020: 607-617.

[65] Russell C. Efficient search for diverse coherent explanations[C]//Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019: 20-28.

[66] Liu S, Kailkhura B, Loveland D, et al. Generative counterfactual introspection for explainable deep learning[C]// 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, 2019: 1-5.

[67] Looveren A V, Klaise J. Interpretable counterfactual explanations guided by prototypes[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2021: 650-665.

[68] Duong T D, Li Q, Xu G. Prototype-based counterfactual explanation for causal classification[J]. arXiv preprint arXiv: 2105.00703, 2021.

[69] Mahajan D, Tan C, Sharma A. Preserving causal constraints in counterfactual explanations for machine learning classifiers[J]. arXiv preprint arXiv:1912.03277, 2019.

[70] Doshi-Velez F, Kim B. Considerations for evaluation and generalization in interpretable machine learning[M]// Explainable and interpretable models in computer vision and machine learning. Springer, Cham, 2018: 3-17.

[71] Molnar C, Casalicchio G, Bischl B. Quantifying interpretability of arbitrary machine learning models through functional decomposition[J]. Ulmer Informatik-Berichte, 2019:41.

[72] Yeh C K, Hsieh C Y, Suggala A, et al. On the (in) fidelity and sensitivity of explanations[J]. Advances in Neural Information Processing Systems, 2019, 32.

[73] Srinivas S, Fleuret F. Full-gradient representation for neural network visualization[J]. Advances in neural information processing systems, 2019, 32.

[74] Moraffah R, Karami M, Guo R, et al. Causal interpretability for machine learning-problems, methods and evaluation[J]. ACM SIGKDD Explorations Newsletter, 2020, 22(1): 18-33.

[75] Zhang Y, Weng Y, Lund J. Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. *Diagnostics*. 2022; 12(2):237.

[76] El-Sappagh S, Alonso J M, Islam S M, et al. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease[J]. Scientific reports, 2021, 11(1): 1-26.

[77] Lamy J B, Sekar B, Guezennec G, et al. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach[J]. Artificial intelligence in medicine, 2019, 94: 42-53.

[78] Lin K, Gao Y. Model interpretability of financial fraud detection by group SHAP[J]. Expert Systems with Applications, 2022, 210: 118354.

[79] Wang D, Lin J, Cui P, et al. A semi-supervised graph attentive

network for financial fraud detection[C]//2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019: 598-607.

[80] Kim J, Canny J. Interpretable learning for self-driving cars by visualizing causal attention[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2942-2950.

[81] Soares E, Angelov P, Filev D, et al. Explainable density-based approach for self-driving actions classification[C]// 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE, 2019: 469-474.

[82] Ahmed F, Erman J, Ge Z, et al. Detecting and localizing end-to-end performance degradation for cellular data services based on tcp loss ratio and round trip time[J]. IEEE/ACM Transactions on Networking, 2017, 25(6): 3709-3722.

[83] Li Z, Zhao N, Li M, et al. Actionable and interpretable fault localization for recurring failures in online service systems[C]// Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2022: 996-1008.