计算机科学
COMPUTER SCIENCE

人工智能可解释性:发展与应用

王冬丽, 杨珊, 欧阳万里, 李抱朴, 周彦

引用本文

王冬丽, 杨珊, 欧阳万里, 李抱朴, 周彦人工智能可解释性:发展与应用[J]. 计算机科学, 2023, 50(6A): 220600212-7.

WANG Dongli, YANG Shan, OUYANG Wanli, LI Baopu, ZHOU Yan. Explainability of Artificial Intelligence:Development and Application [J]. Computer Science, 2023, 50(6A): 220600212-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600212

人工智能可解释性:发展与应用

王冬丽[1]　杨　珊[1]　欧阳万里[2]　李抱朴[3]　周　彦[1]
1 湘潭大学自动化与电子信息学院　湖南 湘潭 411105
2 悉尼大学电气与信息工程学院　悉尼 2006
3 百度美国研究院　森尼韦尔 94086
(wangdl@xtu.edu.cn)

摘　要　近年来人工智能在诸多领域和学科中的广泛应用展现出了其卓越的性能,这种性能的提升通常需要牺牲模型的透明度来获取。然而,人工智能模型的复杂性和黑盒性质已成为其应用于高风险领域最主要的瓶颈,这严重阻碍了人工智能在特定领域的进一步应用。因此,亟需提高模型的可解释性,以证明其可靠性。为此,从机器学习模型可解释性、深度学习模型可解释性、混合模型可解释性3个方面对人工智能可解释性研究的典型模型和方法进行了介绍,进一步讲述了可解释人工智能在教学分析、司法判案、医疗诊断3个领域的应用情况,并对现有可解释方法存在的不足进行总结与分析,提出人工智能可解释性未来的发展趋势,希望进一步推动可解释性研究的发展与应用。

关键词:人工智能;机器学习;深度学习;可解释性
中图法分类号　TP391

Explainability of Artificial Intelligence:Development and Application

WANG Dongli[1],YANG Shan[1],OUYANG Wanli[2],LI Baopu[3] and ZHOU Yan[1]
1 School of Automation and Electronics Information,Xiangtan University,Xiangtan,Hunan 411105,China
2 School of Electrical and Information Engineering,The University of Sydney,Sydney 2006,Australia
3 Baidu Research(USA),Sunnyvale,CA 94086,USA

Abstract　In recent years,the extensive application of artificial intelligence in many fields and disciplines has shown its excellent performance. The improvement of this performance usually needs to sacrifice the transparency of the model. However,the complexity and black box nature of artificial intelligence models have become the main bottleneck in its application in high-risk fields,which seriously hinders the further application of artificial intelligence in specific fields. Therefore,it is urgent to improve the interpretability of the model to prove its reliability. Therefore,this paper introduces the typical models and methods of AI interpretability research from three aspects:machine learning model interpretability,deep learning model interpretability,and hybrid model interpretability,further describes the application of interpretable AI in teaching analysis,judicial judgment,and medical diagnosis,and summarizes and analyzes the shortcomings of existing interpretable methods,puts forward the development trend of the future research direction of AI interpretability,and hope to further promote the development and application of interpretability research.

Keywords　Artificial intelligence,Machine learning,Deep learning,Interpretability

1　引言

随着神经网络(Neural Network,NN)和深度学习(Deep Learning,DL)的发展,人工智能(Artificial Intelligence,AI)、机器学习、脑机接口及其相关的子领域在智慧医疗和图像处理等各个领域的应用中都表现出了令人鼓舞的性能。但当前人工智能存在着不可解释的黑盒问题,这意味着用户只能看到结果,无法了解其做出决策的原因和过程,因而难以分辨人工智能背后的逻辑。这样的人工智能系统难以得到人类的信任和理解,阻碍了人们对智能系统的进一步应用,因此人工智能的可解释性已成为亟待研究和解决的问题[1]。

为了提高人工智能模型的可解释性和透明性,增加用户与决策模型之间的信任度,减少模型在实际应用中的潜在威胁,学术界和工业界近年来对其进行了广泛和深入的研究并且提出了一系列的模型可解释性方法。然而,由于不同的研究者解决问题的角度不同,所提出的可解释性方法也各有侧重。因此,亟需对现有工作进行系统的整理和科学的总结、归类,以促进该领域的研究。

基金项目:国家重点研发计划项目(2020YFA0713503);国家自然科学基金项目(61773330);国家航空科学基金项目(20200020114004);湖南省科技创新计划项目(2020GK2036)
This work was supported by the National Key Research and Development Program of China(2020YFA0713503),National Natural Science Foundation of China(61773330),Aeronautical Science Foundation of China(20200020114004) and Science and Technology Innovation Program of Hunan Province,China(2020GK2036).
通信作者:周彦(yanzhou@xtu.edu.cn)

本文通过对国内外可解释性相关的研究工作进行总体调研并对其进行分类，简单地介绍模型可解释性相关技术的实际应用场景，并从双驱动决策推理模型、跨模态可解释人工智能框架、基于微分方程的机器学习建模、基于原型的模型 4 个方面给出可解释性未来的发展方向。

2 可解释类型分类

关于人工智能可解释类型的分类，有许多不同的观点。比如，Hua 等[2]根据可解释性原理对现有可解释性方法进行分类；Kong 等[3]从系统应用视角和决策收益者视角出发，对模型解释方法进行分类；Zeng 等[4]从自解释模型、特定模型解释、不可知模型解释、因果可解释性 4 个方面对主要可解释性方法进行总结和分析。然而，目前人工智能可解释性研究综述缺乏从混合模型的角度对可解释性方法进行分析和总结，而混合模型比以往解释方法更能在透明度和预测性能之间取得平衡。基于此，本文加入混合模型可解释性研究的最新进展，并对目前的可解释性方法重新进行分类，将其分为以下 3 个类别：机器学习模型可解释性、深度学习模型可解释性和混合模型可解释性。具体如图 1 所示。

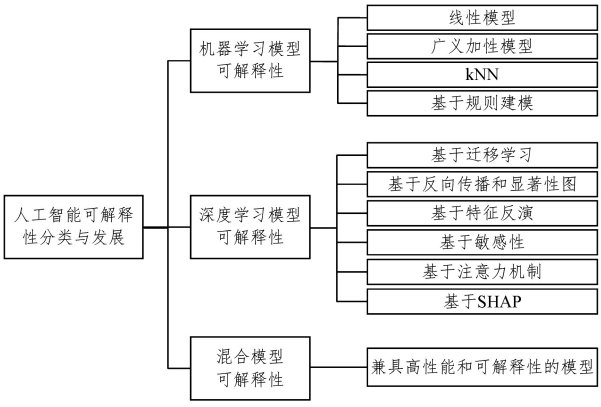


图 1 人工智能可解释性方法的分类与发展

Fig. 1 Taxonomy and development of artificial intelligence interpretability methods

2.1 机器学习模型可解释性

机器学习模型算法和结构相对简单，容易被人类认知和理解。下面列举几类典型的机器学习模型。

2.1.1 线性模型

线性模型中的权重直接反映了样本特征的重要性。权重绝对值越大，该特征对最终预测结果的贡献越大，反之则越小。如果权重值为正，则该特征与最终的预测类别正相关，反之则负相关。但是，由于人类认知的局限性，线性模型结构不能太过复杂，高维线性模型的可解释性未必优于深度神经网络。在模型是高度非线性的情况下，可以通过线性探针[5]来提高可解释性。研究还表明，神经网络层数越深，探针的表现越好。

2.1.2 广义加性模型

线性模型因为准确率低而无法满足需要，而复杂模型的高准确率通常是以牺牲自身可解释性为代价。作为一种折中，广义加性模型既能提高简单线性模型的准确率，又能保留线性模型良好的内置可解释性。广义加性模型的一般形式为 $g(y)=f_1(x_1)+f_2(x_2)+\cdots+f_n(x_n)$，其中 $g(y)$ 为连接函数，$f_j(x_j)$ 为单特征模型，也称为特征 x_j 对应的形函数。在广义加性模型中，形函数本身可能是非线性的，每一个单特征模型可能采用一个非常复杂的形函数 $f_j(x_j)$ 来量化每一个特征 x_j 与最终决策目标之间的关系，从而捕获到每一个特征与最终决策目标之间的非线性关系，因此广义加性模型的准确率高于简单线性模型。又因为广义加性模型通过简单的线性函数组合每一个单特征模型得到最终的决策形式，消除了特征之间的相互作用，因此其可以保留简单线性模型良好的可解释性，从而解决了复杂模型因为特征之间复杂的相关关系而削弱自身可解释性的问题。

2.1.3 KNN

KNN(K-Nearest Neighbors)是解决分类问题的机器学习模型典型方法。KNN 模型的输出取决于所计算样本之间的相似性和距离，其机制类似于人类基于经验的决策，是一种监督算法。由于模型透明，可解释性好，该方法被广泛应用于解释复杂模型的应用中。Wang 等[6]结合 δ-neighborhood 和 KNN 形成的 KNN 粗糙集模型，引入了属性约简方法，有效地处理了异构数据，且性能比 δ-neighborhood 和 KNN 更好。Zheng[7]等人提出了一种由原始 KNN 算法驱动的新型分类器，大大提高了模型分类精度和可解释性。

2.1.4 基于规则建模

由于规则能够更加直观地描述和解释系统机理，因此基于规则的建模方法(Rule-based Modeling Approach，RBM)[8]在理论及应用中得到了广泛的认可与关注。传统产生式规则的形式通常写成：IF X THEN Y。其中，X 为规则前提，Y 表示结论，它对确定性信息具有较为直观的描述能力。为了处理不确定性信息，研究者提出了能够有效利用定量信息和定性知识处理模糊不确定性的基于模糊规则的建模方法。随后，在此基础上引入置信框架，提出了基于证据推理算法的置信规则库推理方法。

RBM 的优势在于其内置的可解释性，现有研究主要从知识库、推理机、模型优化这 3 个方面对其可解释性进行探讨。可解释的知识库需要具备清晰明确的语义，完备、简洁、一致的规则以及具有物理意义的结构和参数，知识库的可解释性保证其能以可理解的方式准确描述实际系统中的不确定信息。但是为了产生可靠的结果，推理机对不确定信息的处理能力就显得至关重要。常用的不确定性推理方法主要包括贝叶斯概率推理方法、信任函数、可能性推理方法以及似然推理方法。实际工程系统由于结构复杂，专家很难完全掌握其机理信息并对其进行解释，因此先利用有限的专家知识构建初始模型，再利用优化学习方法对初始模型的结构和参数进行调整和解释是常用的办法。

2.2 深度学习模型可解释性

2.2.1 基于迁移学习

当模型的结构过于复杂时，从整体上来理解模型是非常困难的，因此需要降低模型的复杂度。随着迁移学习的发展，不仅能够实现模型结构的迁移，还可以将模型的可解释性进行迁移。利用具有可解释性的模型，比如线性模型、决策树模型，将黑盒的深度学习模型迁移到这些可解释的模型中，从而解释这些难以解释的模型。代理模型和模型蒸馏是典型的降低模型复杂度的方法。

局部代理模型是一种适用于所有模型的事后分析方法。其基本思想是用一些可解释的模型去局部拟合不可解释的黑盒模型，使得模型结果具有一定的可解释性。Ribeiro 等[9]提出的局部可理解的解释技术(Local Interpretable Model-agnostic Explanations，LIME)即为一种代理模型方法。该方法

首先通过向输入样本中添加扰动，获得模型的响应反馈数据，然后凭此数据构建局部线性模型，并将该模型用作特定输入值深度模型的简化代理。Ribeiro 表示该方法可作用于识别对各种类型的模型和问题域的决策影响最大的输入区域。但是 LIME 无法准确解释包含序列数据依赖关系（如 Recurrent Neural Networks，RNN）的神经网络，因此 Guo 等[10]提出了一种适用于网络安全方面的非线性近似局部解释方法 LEMNA（Local Explanation Method using Nonlinear Approximation）。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，然后引入 Fused Lasso 正则化以处理 RNN 模型中特征间的依赖问题，有效地弥补了 LIME 方法的不足，提高了解释的保真度。Setzu 等[11]提出的 GLocalX 可作为规则提取的替代方法，它将逻辑规则应用于黑箱模型。GLocalX 分层次地聚合从局部决策规则中提取的局部解释，以此来创建全局解释。实验结果表明，GLocalX 取得了比原始模型更高的性能。

模型蒸馏是利用结构相对简单的学生模型（Student Model）来模拟结构复杂的教师模型（Teacher Model），从而完成从教师模型到学生模型的知识迁移过程。Hinton 等[12]通过训练单一相对较小的网络模拟原始复杂网络或集成网络模型的预测概率来提炼复杂网络的知识，以模拟原始复杂网络的决策过程，并且证明单一网络能达到与复杂网络几乎同样的性能。Zhao 等[13]提出的跨模态知识泛化方法，用于在教师模型不可用的目标数据集中训练学生模型。通过将知识建模作为学生参数的先验，使用教师表示作为监督信号来训练学生学习目标任务，将从源数据集中学到的提炼的跨模态知识推广到不同域的目标数据集，实验表明此方法在标准基准数据集上进行 3D 手部姿态估计识别表现得十分有竞争力。模型蒸馏是学生模型对原始模型的一种近似模拟，只是将一部分有效的知识迁移到学生模型中的一种方法，因此不能完全解释原始模型的决策行为。基于迁移学习的解释方法的优缺点如表 1 所列。

表 1　基于迁移学习的解释方法

Table 1　Interpretation methods based on transfer learning

方法	优点	缺点
代理模型	操作简单，易于理解	只适用于局部近似解释
模型蒸馏	有效压缩模型大小	适用条件苛刻

2.2.2　基于反向传播和显著性图

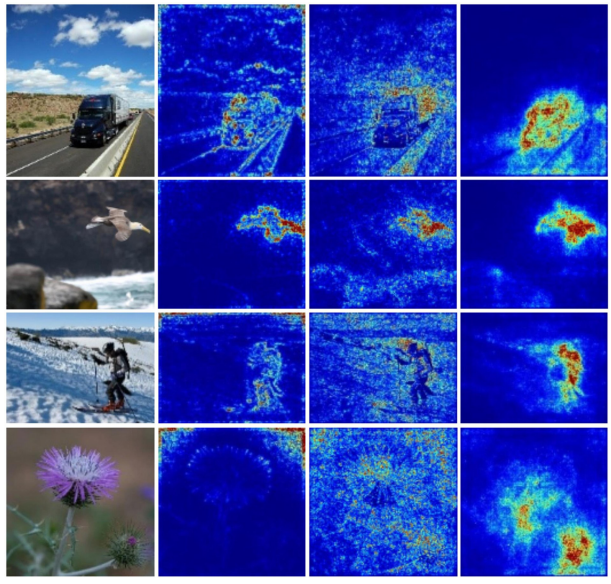
基于反向传播（Back Propagation）的解释方法的核心思想是利用深度神经网络的反向传播机制将模型中影响决策的重要信号从模型的输出层传播到模型的输入层，从而推导输入样本中的不同特征的重要性的值，这些值可以采用超像素（如热图）的形式来表示。

（1）基于梯度的显著图。Springenberg 等[14]提出的通过 ReLU 非线性反向传播时操纵梯度的视觉解释方法（Guided-BP），反向传播时保留了梯度与激活值均为正的部分。由于大部分深度神经网络的非线性映射采用 ReLU 函数，其负半轴为饱和区且梯度为零，因而无法揭示任何有效信息。因此 Sundararajan 等[15]提出了积分梯度方法（Integrated Gradients），其有效地解决了深度神经网络中神经元饱和问题导致无法利用梯度信息反映特征重要性的问题。但是这些方法得到的显著性图都存在视觉可见的噪音，为此 Smilkon 等[16]提出了平滑梯度法（SmoothGrad）。通过向待解释样本中添加噪声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。表 2 列出了基于梯度的解释方法的特点。图 2 所示为基于梯度解释方法的显著图。

表 2　基于梯度的解释方法

Table 2　Gradient-based interpretation methods

方法	特点分析
GuideBP	目标特征较为集中
Integrated gradients	解释 CNN 内部，存在较少噪声
SmoothGrad	定位图像决策特征，无法量化贡献



（a）input　（b）GuidedBP　（c）Integrated　（d）SmoohGrad

图 2　基于梯度的解释方法

Fig. 2　Gradient-based interpretation methods

上述像素空间梯度可视化方法具有高分辨率，但不是类区分性的。类激活映射方法（CAM）[17]是高度类区分的。例如，“猫”解释只突出了包含猫和狗的图像中的“猫”区域，而不是“狗”区域。但是 CAM 不能应用于预先训练的网络和特定类型的神经网络架构，而且在全连接层会丢失掉空间信息。因此，Selvaraju 等[18]设计了具有高分辨率，又具有区分性，且可应用于任意神经网络架构的 Grad-CAM 方法，来解决这一问题。但是 Grad-CAM 仅生成粗粒度可视化，无法解释图像中同一对象的多个实例。此外，Grad-CAM 生成的热图的定位相对于覆盖图像中的类区域而言不是很准确，因此 Chattopadhay 等[19]提出了 Grad-CAM＋＋来弥补这些缺点。Grad-CAM＋＋考虑了梯度的加权平均值，可在一个类的所有位置生成热图，其中特定的类在图像中以分散或附着的方式定位。因此，当图像中存在单个类的多个实例时，生成的热图能更好地对模型进行解释，而且能提供更细粒度的解释结果。

（2）基于激活值的显著图。分层相关性传播方法 LRP[20]和 DeepLIFT[21]基于激活值的反向传播来提高显著图的视觉质量。LRP 递归地计算层中每个神经元的相关性分数，让它等于神经元本身的输出（即神经元的激活值），以了解图像单个像素在图像分类任务中做出的贡献。DeepLIFT 方法强调了在目标输入之外引入一个参考输入进行解释的重要性，它将每个神经元的激活与其参考激活进行比较，并根据实际输出和参考输出之间的差异为每个输入分配一个重要性评分，解决了分段联系函数饱和时梯度总为零的问题。其中，在

输出层，每个单元的相关性是指在初始网络输入处激活的单元与在参考输入处激活的单元的相对影响。

2.2.3 基于特征反演

基于反向传播的解释方法重点关注模型的输入层和输出层，忽略了带有大量信息的中间层。特征反演(Feature Inversion)作为一种可视化和理解 DNN 中间特征表征的技术，可以充分利用模型的中间层信息，以提供对模型整体行为及模型决策结果的解释。Du 等[22]提出的一个代表性的特征反演解释框架通过在执行导向特征反演过程中加入类别依赖约束，不仅可以准确地定位待输入实例中用于模型决策的重要特征，还可以提供对 DNN 模型决策过程的深入理解。

2.2.4 基于敏感性

敏感性分析方法是通过逐一改变自变量的值来解释因变量受自变量变化影响大小的一种技术。根据是否需要利用模型的梯度信息，敏感性分析方法可分为模型相关方法和模型无关方法。

(1)模型相关方法利用模型的局部梯度信息评估特征与决策结果的相关性。常见的相关性定义为：

$$R_i(x)=\left(\frac{\partial f}{\partial x_i}\right)^2 \tag{1}$$

其中，$f(x)$为模型的决策函数，x_i为待解释样本 x 的第 i 维特征。相关性分数 $R_i(x)$可通过梯度反向传播来求解，最后以热力图的形式可视化相关性分数，可以直观地理解输入的每一维特征对决策结果的影响程度。

(2)模型无关方法无需利用局部梯度信息，只关注待解释样本特征值变化对模型最终决策结果的影响。LIME[9]是一种与模型无关的局部近似方法，首先在关注的样本点附近进行轻微扰动，然后探测模型输出发生的变化，根据这种变化在兴趣点附近拟合出一个可解释的简单模型(如线性模型)。LIME 的主要限制之一是它提供的解释很大程度上取决于分配给扰动样本的权重，为此 ILIME[23]通过对扰动实例的影响以及与待解释实例的距离来加权扰动实例解决了这个问题。

敏感性分析方法只能捕获到单个特征或局部变化对最终结果的影响程度，不一定关注实际的结果相关特征，因此敏感性分析方法提供的解释结果通常相对粗糙且难以理解。此外，敏感性分析方法无法解释特征之间的相关关系对最终结果的影响。

2.2.5 基于注意力机制

深度学习神经网络模型由于结构复杂，可解释性差，因此必须引入外部解释模块对其进行解释。一种有效的方法就是引入注意力机制。注意力机制源于对人类认知神经学的研究，人类并不倾向于一次性处理事件的全部信息，而是选择性地专注于部分有用的信息，同时忽略其他可感知的信息，这就是人脑的注意力机制，本质是对信息进行加权。注意力机制可以作为资源分配方案，是解决信息过载问题的主要手段。此外，注意力机制具有良好的可解释性，注意力权重热图直接体现了模型在决策过程中感兴趣的区域。近年来，在自然语言处理和计算机视觉等领域，基于注意力机制的可解释性方法已成为一大研究热点。

在自然语言处理领域，Lin 等[24]引入注意力机制提出了一个获取可解释的句子嵌入模型，为了将可变长度的句子编码为固定大小的嵌入，用二维嵌入矩阵而非向量代表句子嵌入来修剪权重连接，矩阵的每行表示句子的不同部分的特征，多行代表从不同的角度对句子进行嵌入。由于引入了注意力机制，嵌入矩阵的每行都有其对应的标记权重向量，因此可直接可视化句子中每个元素的贡献程度(如图 3 所示，颜色越红，注意力越重)。此外，研究人员受 Transformer 模型[25]的启发，根据编码器部分提出了 BERT[26]双向编码器。Transformer 是一种基于自注意力机制的序列模型，它通过多个注意力矩阵来表示不同位置的注意力强度。BERT 是一个双向语言模型，它随机屏蔽部分输入标记，然后预测那些被屏蔽的标记，并使用线性二元分类器来辨别两个句子是否连接。注意力头是 BERT 的基本组成模块，通过可视化注意力头的权重分布可以清楚地知道模型权重更偏向哪种词性。实验表明，双向编码器在屏蔽语言模型和下一句预测任务中的精度得到了显著提高。

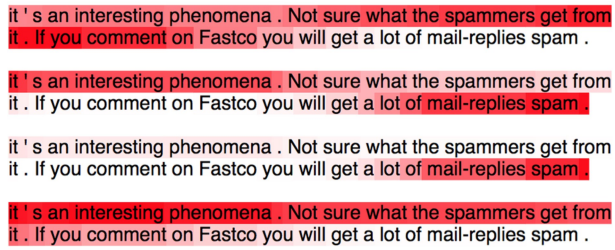


图 3 注意力权重热图

Fig. 3 Heatmap of attention weight

在计算机视觉领域，Hu 等[27]提出的挤压和激励网络(Squeeze-and-Excitation，SE)开创了通道注意力的先河。挤压模块通过全局平均池化层压缩通道特征图中的全局空间信息。激励模块通过全连接层以及非线性层，利用全局信息学习特征通道之间的相关性，筛选出针对特征通道的注意力权重，选择性地增强有益的特征通道并抑制无用的特征通道，从而实现特征通道的自适应校准。此外，Transformer 架构已经是自然语言处理领域的首选模型，但其在视觉领域的应用仍然非常有限。Dosovitskiy 等[28]将 Transformer 成功用于图像处理任务。首先将图片划分为若干块，每个块相当于一个单词，从而可以使用与词向量相同的编码模型，然后对块添加位置编码，使得注意力权重更容易观察到图像中的物体，帮助模型正确评估注意力权重，再使用全连接网络对块进行线性变换得到线性变换序列并输入进 Transformer，获得了与 CNN 相媲美甚至更出色的结果(注意力可视化图如图 4 所示)，成功解决了计算机视觉领域过度依赖 CNN 的问题。

图 4 注意力可视化图

Fig. 4 Maps of attention visualization

2.2.6 基于 SHAP

SHAP[29](SHapley Additive exPlanations)是一种受博弈论启发的方法，它使用加性特征归因方法计算每个特征的重要性值来确定特征对模型的影响程度。与一般解释模型不同的是，SHAP 可以识别每个输入特征的重要性值是正数还是负数，而且特征的每个观测值都可以获取其 SHAP 值，因此

它既可用于局部解释，也可用于全局解释。SHAP 的另一个优点是可对任何模型和任何类型的数据进行解释。KernelSHAP[29] 和 TreeSHAP[30] 是最常用的 SHAP 解释方法。KernelSHAP 是受 LIME[9] 启发的精确 Shapley 值的加权线性回归近似，可用于为任何黑盒模型提供局部解释。TreeSHAP可为任何基于树的模型提供解释，它利用决策树结构来分解决策树或决策树集成模型中每个输入的贡献。但是，KernelSHAP 与大多数基于排列的方法一样，没有考虑特征相关性，经常过度加权不太可能的数据点，因此 TreeSHAP 通过显式建模条件期望预测解决了这个问题。在解释黑盒模型方面，SHAP 是可视化特征交互和特征重要性方法中最全面和占主导地位的方法之一。

2.3 混合模型可解释性

机器学习模型结构简单，可解释性好但泛化性差。深度学习黑盒模型结构复杂，准确性高，但可解释性差。因此，可将两种模型结合起来形成高性能且可解释性好的混合可解释模型。近年来有少许研究团队在这方面做过一些探索，例如 Hu 等[31] 提出了一种迭代蒸馏方法，提出的通用框架将深度神经网络(CNN 或 RNN 等)与结构化逻辑规则相结合，将逻辑规则的结构化信息转换为神经网络的权重，使神经网络能够通过迭代规则知识蒸馏过程的同时从标记实例和逻辑规则中学习。研究者将框架应用在情感分析和命名实体识别任务上，通过高度直观的规则，提高了神经模型的可解释性，获得了比当时最佳系统更好或相当的结果。这是第一个将逻辑规则与一般主流深度神经网络集成到一个框架中的方法，结果表明，此种方法有助于整合更丰富的人类知识，且在视觉任务方面具有巨大的潜力。

最近，Wang 等[32] 提出了一个可构建混合预测模型的框架。该框架可将本身可解释性好的模型与任何预训练的深度学习黑盒模型的优势相结合而形成混合模型，即用在解释性最好的数据集上的可解释模型替换黑盒模型，以此来获得混合模型一定的透明度。混合模型使用帕累托边界来衡量透明度和准确性，在透明度和预测性能之间取得了有效的平衡。在这个框架下选取关联规则和线性模型作为易解释模型，与最先进的黑盒模型结合，在结构化数据和文本数据上训练测试出两个混合模型，将训练出来的模型应用于心血管疾病预测，模型可为患者是否会患上心血管疾病提供专业建议。但此模型仅适用于结构化数据和文本数据，不适用于原始图像处理和分类问题。Yeganejou 等[33] 结合卷积神经网络和模糊系统架构形成一种混合卷积模糊分类器，使用卷积神经网络作为特征提取器，然后对提取的特征进行聚类，通过识别每个集群的中心点并评估输入数据中每个像素的重要性，为模型创建了一个可解释机制。实验结果表明，CNN 特征提取器显著提高了模糊分类器的性能和解释能力。

混合模型可解释性的主要优点是它为黑盒模型提供了鲁棒性和可解释性，有效缓解了在高精度黑盒模型和易解释模型之间进行选择的两难境地。

3 人工智能可解释性应用

3.1 教学分析

大数据教育的可用性和人工智能技术的增强，促进了人工智能在教育领域的应用。但近几年新冠疫情的爆发导致学生不能在校上课，传统学习方法的使用受到限制，此时为了让教育系统根据不同学生的表现做出区别性响应，并对响应做出合理的解释，适时地介入学生学习或者进行有效的干预，让系统辅助老师对学生进行教导来提高教学质量，可解释人工智能教育系统便显得尤为重要[34]。

在线教学时需要了解哪些因素和特征导致学生表现不佳，并采取措施提高他们的成绩表现。Hasib 等[35] 使用模型预测中学生的成绩表现，然后用 LIME 方法对模型进行解释，通过更改特征值来更改数据样本，并记录每次修改对预测的影响。模型表明影响学生表现的特征之间存在显著关联。这是人工智能应用于教育领域的典型案例，可解释性人工智能赋予了人工智能教育新的意义，是知识价值的重要体现。

3.2 司法判案

人工智能技术在辅助法官判案等诸多方面起着重要作用，但是判决罪名这种极为谨慎的决定有着很高的风险，而且面对海量的司法判决文书，将法官判案与人工智能技术相结合，提高审判工作效率，不仅能解决法官人数不足而出现的案多人少的难题，而且给人工智能可解释性的应用提供了一个极具挑战性的应用场景。在真实场景中，用户不仅关心判决结果，还想知道与模型决策结果相关的法律依据，或者知晓是什么原因导致这样的决策结果，这就极需模型对决策结果给出相应的解释。下面介绍人工智能辅助判案在司法判决领域的典型应用。

司法判决预测的任务是分析法律文章和文件，从大量历史案例中提取法律因素及其关系，通过分析其文本事实描述来确定未决案件的司法结果。但法律词汇在汉语中没有全球标准，语义歧义普遍存在，而且预测结果不易为公众所理解。Li 等[36] 结合规则与深度学习的方法提出了一个认知计算框架来提取法律因素，并对概念和关系进行扩充，生成预测模型再进行训练。当将事实描述引入框架时，每种预测结果的概率将自动给出。除了给出预测结果之外，还对预测结果以提供归纳规则的方式进行解释，这有助于外行人士或非专业人士理解预测结果。Bao 等[37] 提出了一种利用相关法条来预测多个罪名的注意力神经网络 LegalAtt。该模型使用事实描述查找相关法条并生成法条表示序列，再将表示序列输入到注意力层，得到具有法条注意力的事实表示来进行罪名预测。

3.3 医疗诊断

人工智能技术的发展让其在医疗诊断领域的应用越来越广泛，但随着医学检验项目的不断增多，以及不同病症和病程的检验结果越来越复杂，即使是颇有经验的医学专家也难免会给出经验性的误诊。身患疾病不仅是一项指标异常所决定的，大多是多项指标不合格所引起，正确分析指标之间的关联及其内在关系也是一项极其艰巨的任务。而人工智能技术具有对复杂文本及图像的强大处理能力，故能及时辅助医生进行疾病诊断和病情分析。

医疗诊断对人工智能辅助诊疗的需求不断增加，尤其是近几年新冠疫情流行期间。但是人工智能诊断模型不仅需要高精确度，还需要令人信赖的可解释性，特别是在高风险的临床领域。Soares 等[38] 研究了一种可解释的深度学习方法，通过观察病患的计算机断层扫描的胸部成像图片，识别病人是否为 COVID-19 患者。胸部放射学评估通常是评估疑似 COVID-19 患者的关键步骤，分析显示新冠患者胸部扫描成像中肺双侧浑浊，尤其是小叶和亚节段区域。该方法在准确性、F1 分数方面超过了 ResNet 等主流深度学习方法。同时，

可解释深度学习分类器提供了医学图像的高度可解释的 IF THEN 规则，可供医学专家用于新冠病毒感染早期诊断的决策过程。此外，Couteaux 等[39]提出了一种可解释的 DeepDream 方法，可用于肝脏扫描来识别是否患有肿瘤。该方法通过对给定肝脏图像进行梯度上升来最大化激活神经元，其输出曲线显示的最大化过程中的特征演变有利于神经网络的可视化和可解释性。图 3 所示为肝脏肿瘤分割可视化图，颜色越接近白色代表患肿瘤概率越高。

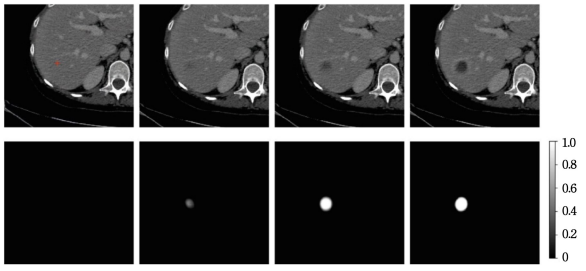


图 5 肝脏肿瘤分割可视化图

Fig. 5 Visualization of liver tumor segmentation

4 总结与展望

4.1 总结

本文从机器学习模型可解释性、深度学习模型可解释性和混合模型可解释性 3 个方面对人工智能可解释性研究进行了综述。现有的可解释方法还存在很多问题，比如很难把握模型准确性和可解释性之间的平衡，文献[32]中的混合模型虽然在达到准确度的同时兼具了可解释性，但不适用于图像处理问题，未来可以尝试将专门处理图像的可解释模型与高准确率的深度学习模型相结合来处理图像问题；其次，模型开发者容易忽略终端用户的真实体验，设计出来的模型不能为来自各个领域的用户提供可理解的解释，根据不同用户背景做出不同的解释可增强人类对人工智能的信任度。现已有可解释交互式人工智能技术来提高人工智能对人类的解释性，未来可设计允许用户灵活调整解释技术的解释界面，或者根据 AI 可解释性的不同需求，允许用户能够单独提取特定解释，以此增强人机交互程度来提高解释质量。

4.2 展望

4.2.1 双驱动决策推理模型

深度神经网络提供了一种强大的机制从大量数据中进行学习，尽管已经取得了令人瞩目的成就，但其高准确预测性严重依赖大量标记数据，纯粹的数据驱动学习可能会导致无法解释，有时甚至产生违反直觉的结果。人类的认知过程表明，人们不仅从具体示例中学习，还从不同形式的一般知识和丰富经验中学习。混合模型中将深度神经网络和逻辑规则相结合的方法提供了一种新的思路，如果可以更好地将人类意图和领域知识整合到神经模型中，建立数据驱动的机器学习与知识驱动的符号计算相融合的双驱动决策推理模型，将有望突破神经网络模型不可解释的这一瓶颈。

4.2.2 跨模态可解释人工智能框架

随着互联网上多源数据的爆炸式增长，单一模态模型已经无法满足用户需求，跨模态模型应运而生。但多模态数据的异质性和不能在线交互的特性导致模型无法满足一些细致的需求。比如百度搜索黑色上衣白色裤子，尽管大多数结果都是正确的，但是仍有少部分错误结果，说明模型对多模态数据中包含的细粒度信息的分析不够。另一方面，现有证据表明，人类视觉系统能够在小数据条件下有效运行，人类可以利用先验知识和上下文从极少数样本中学习新概念，在庞大的数据量的情况下，小样本学习的作用就不可忽视了。模型蒸馏中的跨模态知识泛化就是代表性的跨模态方法。现有方法大多都是单模态或双模态处理模型，人其实是一个多模态学习的总和。研究适用于同时理解和处理视频、语音、自然语言、点云、地理数据等多模态复杂数据和小样本学习的跨模态可解释人工智能框架，是人工智能的一个重要发展方向。

4.2.3 基于微分方程的机器学习建模

现有机器学习方法一般运用概率论、线性代数、优化理论等基础数学，模型的设计缺少系统指导，大多数模型都缺少可解释性，这也限制了它的应用。如果将机器学习与微分方程融合，从数学角度出发设计网络架构，并分析它们的泛化性能力和可解释性，那么神经网络架构就是数值微分方程，网络训练的实质就是最优控制，神经网络的设计也获得了强有力的数学理论指导，从而有望设计出更有鲁棒性和可解释性的机器学习模型。

4.2.4 基于原型的模型

未来一个具有很大研究空间的方向是基于原型的模型解释。人类通常将复杂实体与先前已知的原型进行比较，得出它们之间的相似性，来描述复杂项目，而不是直接对其提供详细的描述，这就是基于原型的模型解释。早在多年前就有研究人员尝试了基于原型的模型的研究[40]，但到目前为止，这类模型还没有在深度学习的背景下进行大量开发，这为模型可解释性研究开辟了一个有潜力的发展空间。

参 考 文 献

[1] CHAO L M, YIN X L. AI Governance and System: Current Situation and Trend[J]. Computer Science, 2021, 48(9): 1-8.

[2] HUA Y Y, ZHANG D C, GE S M. Research Progress in the Interpretability of Deep Learning Models[J]. Journal of Cyber Security, 2020, 5(3): 1-12.

[3] KONG X W, TANG X Z, WANG Z M. A Survey of Explainable Artificial Intelligence Decision[J]. Systems Engineering-Theory & Practice, 2021, 41(2): 524-53.

[4] ZENG C Y, YANK, WANG Z F, et al. Survey of Interpretability Research on Deep Learning Models[J]. Computer Engineering and Applications, 2021, 57(8): 1-9.

[5] ALAIN G, BENGIO Y. Understanding intermediate layers using linear classifier probes[J]. arXiv: 1610.01644, 2016.

[6] WANG C, SHI Y, FAN X, et al. Attribute Reduction Based on K-nearest Neighborhood Rough Sets[J]. International Journal of Approximate Reasoning, 2019, 106: 18-31.

[7] ZHENG S, DING C. A Group Lasso Based Sparse KNN Classifier [J]. Pattern Recognition Letters, 2020, 131: 227-233.

[8] ZHOU Z J, CAO Y, HU C H, et al. The Interpretability of Rule-based Modeling Approach and Its Development[J]. Acta Automatica Sinica, 2021, 47(6): 1201-1216.

[9] 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.

[10] GUO W B, XU J. Using Lemna to Explain the Application of Deep Learning in Network Security(Part I)[J]. China Education Network, 2019(z1): 40-43.

[11] SETZU M, GUIDOTTI R, MONREALE A, et al. Glocalx-from

Local to Global Explanations of Black Box AI Models[J]. Artificial Intelligence，2021，294：103457.

[12] HINTON G，VINYALS O，DEAN J. Distilling the Knowledge in A Neural Network[J]. Computer Science，2015，14(7)：38-39.

[13] ZHAO L，PENG X，CHEN Y，et al. Knowledge as Priors：Cross-modal Knowledge Generalization for Datasets without Superior Knowledge[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：6528-6537.

[14] SPRINGENBERG J T，DOSOVITSKIY A，BROX T，et al. Striving for simplicity：The all convolutional net[C]// Proceedings of 3rd ICLR(Workshop Track). 2015.

[15] SUNDARARAJAN M，TALY A，YAN Q. Axiomatic Attribution for Deep Networks[C]// International Conference on Machine Learning. PMLR，2017：3319-3328.

[16] SMILKOV D，THORAT N，KIM B，et al. Smoothgrad：removing noise by adding noise[C]// ICML Workshop on Visualization for Deep Learning. 2017.

[17] 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.

[18] SELVARAJU R R，COGSWELL M，DAS A，et al. Grad-CAM：Visual Explanations from Deep Networks via Gradient-based Localization[J]. International Journal of Computer Vision，2020，128(2)：336-359.

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

[20] BINDER A，MONTAVON G，LAPUSCHKIN S，et al. Layer-wise relevance propagation for neural networks with local renormalization layers[C]// International Conference on Artificial Neural Networks. 2016：63-71.

[21] SHRIKUMAR A，GREENSIDE P，KUNDAJE A. Learning important features through propagating activation differences [C]// International Conference on Machine Learning. PMLR，2017：3145-3153.

[22] DU M，LIU N，SONG Q，et al. Towards explanation of dnn-based prediction with guided feature inversion[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018：1358-1367.

[23] ELSHAWI R，SHERIFY，AL-MALLAH M，et al. ILIME：Local and global interpretable model-agnostic explainer of black-box decision[C]// European Conference on Advances in Databases and Information Systems. Cham：Springer，2019：53-68.

[24] LIN Z，FENG M，SANTOS C N D，et al. A structured self-attentive sentence embedding[C]// Proceedings of the 5th International Conference on Learning Representations. Toulon，France，2017：1-15.

[25] GALASSI A，LIPPI M，TORRONI P. Attention，please! a critical review of neural attention models in natural language processing[J]. arXiv：1902. 02181，2019.

[26] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. 2019：4171-4186.

[27] HU J，SHEN L，ALBANIE S，et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2019，42(8)：2011-2023.

[28] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：Transformers for image recognition at scale[C]// Proceedings of the 9th International Conference on Learning Representations. 2021.

[29] LUNDBERG S M，LEE S I. A unified approach to interpreting model predictions[J]. Advances in Neural Information Processing Systems，2017，30.

[30] LUNDBERG S M，ERION G，CHEN H，et al. From local explanations to global understanding with explainable AI for trees [J]. Nature Machine Intelligence，2020，2(1)：56-67.

[31] HU Z T，MA X Z，LIU Z Z，et al. Harnessing deep neural networks with logic rules[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(Volume 1：Long Papers). 2016：2410-2420.

[32] WANG T，LIN Q. Hybrid predictive models：when an interpretable model collaborates with a black-box model[J]. Journal of Machine Learning Research，2021，22(137)：1-38.

[33] YEGANEJOU M，DICK S，MILLER J. Interpretable deep convolutional fuzzy classifier[J]. IEEE Transactions on Fuzzy Systems，2019，28(7)：1407-1419.

[34] WANG S X. AI Empowerment Education[J] China education network，2021(1)：15.

[35] HASIB K M，RAHMAN F，HASNAT R，et al. A machine learning and explainable AI approach for predicting secondary school student performance[C]// 2022 IEEE 12thAnnual Computing and Communication Workshop and Conference(CCWC). 2022：399-405.

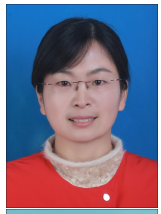
[36] LI J，ZHANG G，YU L，et al. Research and design on cognitive computing framework for predicting judicial decisions[J]. Journal of Signal Processing Systems，2019，91(10)：1159-1167.

[37] BAO Q，ZAN H，GONG P，et al. Charge prediction with legal attention[C]// CCF International Conference on Natural Language Processing and Chinese Computing. Cham：Springer，2019：447-458.

[38] SOARES E，ANGELOV P，BIASO S，et al. SARS-CoV-2 CT-scan dataset：A large dataset of real patients CT scans for SARS-CoV-2 identification[J]. MedRxiv，2020.

[39] COUTEAUX V，NEMPONT O，PIZAINE G，et al. Towards interpretability of segmentation networks by analyzing DeepDreams[M]// Interpretability of machine intelligencein medical image computing and multimodal learning for clinical decision support. Cham：Springer，2019：56-63.

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



WANG Dongli，Ph. D，associate professor. Her main research interests include pattern recognition and distributed learning，intelligent decision-making and information processing.

ZHOU Yan，Ph. D，professor. His main research interests include machine vision and cluster robotics，signal processing and information fusion.