深度学习模型可解释性的研究进展

化盈盈[1,2], 张岱墀[1,2], 葛仕明[1]

[1] 中国科学院信息工程研究所 北京 中国 100093
[2] 中国科学院大学网络空间安全学院 北京 中国 100049

摘要 深度学习在很多人工智能应用领域中取得成功的关键原因在于, 通过复杂的深层网络模型从海量数据中学习丰富的知识。然而, 深度学习模型内部高度的复杂性常导致人们难以理解模型的决策结果, 造成深度学习模型的不可解释性, 从而限制了模型的实际部署。因此, 亟需提高深度学习模型的可解释性, 使模型透明化, 以推动人工智能领域研究的发展。本文旨在对深度学习模型可解释性的研究进展进行系统性的调研, 从可解释性原理的角度对现有方法进行分类, 并且结合可解释性方法在人工智能领域的实际应用, 分析目前可解释性研究存在的问题, 以及深度学习模型可解释性的发展趋势。为全面掌握模型可解释性的研究进展以及未来的研究方向提供新的思路。

关键词 深度学习模型; 可解释性; 人工智能
中图法分类号 TP181 DOI 号 10.19363/J.cnki.cn10-1380/tn.2020.05.01

Research Progress in the Interpretability of Deep Learning Models

HUA Yingying[1,2], ZHANG Daichi[1,2], GE Shiming[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Deep learning has succeeded in many areas of artificial intelligence, and the key reason for this is to learn a wealth of knowledge from massive data through complex deep networks. However, the high degree of complexity in deep learning models often makes it difficult for people to understand the decision-making results, which makes deep learning models unexplainable and limits their practical deployment. Therefore, there is an urgent need to improve the interpretability of deep learning models and make the models transparent to promote the development of artificial intelligence. This paper aims to systematically study the research progress in the interpretability of deep learning models. And we make a new division of these interpretable methods from the perspective of interpretability principles. According to the practical application of interpretability, we analyze and summarize the problems existing in the current interpretable research and the future development trend of explainable artificial intelligence. It provides new ideas to comprehensively understand the current progress and the further direction of interpretability.

Key words deep learning models; interpretability; artificial intelligence

1 引言

深度学习模型[1]在许多领域都具有非常好的性能, 比如人脸识别、图片分类、自然语言处理等, 但是这种表现更多的依赖于模型高度的非线性和调参技术。人们无法探知深度模型究竟从数据中学到了哪些知识, 以及如何进行最终决策的。这种“端到端”的决策模式导致深度学习模型的解释性极弱。站在人的角度分析, 模型的决策过程是无法理解的, 即模型是不可解释的[2]。

深度学习模型的不可解释性存在很多的潜在危险[3], 尤其在安全攻防领域。一方面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, 比如[4]对抗样本攻击模型时, 很难说明是哪些原因导致结果出现了如此大的偏差, 从而无法对模型的攻击进行追踪和溯源。此外, 一个不可解释的模型由于无法给予用户更多可靠的信息, 在很多领域的实际部署会受到极大地

通讯作者: 葛仕明, 博士, 副研究员, 博士生导师, Email: geshiming@iie.ac.cn。
本课题得到国家自然科学基金(No.61772513)资助。
收稿日期: 2020-02-07; 修改日期: 2020-04-22; 定稿日期: 2020-04-29

限制。模型的不可解释性所带来的一连串问题，也在不断驱动我们深入地探究如何提高深度学习模型的可解释性。因此，人们一直致力于更透彻地去理解深度学习模型内部复杂的过程，从而达到进一步优化模型的目的。

为了提高深度学习模型的可解释性，已经提出了很多可解释性方法[5-6]，比如利用特征重要性[7]衡量不同特征对决策结果的影响，或者用可解释的决策树模拟深度学习模型的预测输出[8]等。但是目前的研究成果依然存在很多的不足，尤其是缺乏对可解释性研究现状的总结与分析，现有的工作更多的侧重于对可解释性方法的罗列介绍，没有对可解释性的研究成果进行全面了解和深入分析，不利于可解释性研究的进一步推进。尽管深度学习模型的可解释性早已成为研究的热点，并且在人工智能领域取得了一定的关注，但是目前的可解释性研究成果相对分散，这些可解释性方法缺乏系统的分析总结。

基于此，本文对深度学习模型的可解释性进行了深入的调研，以促进可解释性的进一步发展。为了全面掌握可解释性研究的进展，并为可解释性中存在的开放问题提供新的研究视角，本文将从新的角度切入对目前的可解释性工作进行系统地分析总结，探索可解释研究的内在规律，预测可解释性未来的发展态势和研究方向。基于可解释性研究的原理，从模型结构、特征分析、可解释性迁移三个角度分析目前的研究现状。通过对可解释性方法的系统性介绍，再结合当前可解释性的一些典型应用，分析可解释性方法存在的问题和未来的发展趋势。

本文将从以下方面展开：首先介绍可解释性的研究现状，主要从可解释性的相关概念和可解释性方法两个方面展开。然后结合可解释性的实际应用，分析可解释性研究取得的进展和存在的不足。最后对可解释性的发展趋势进行总结与展望。

2 可解释性的研究现状

广义上来说我们对可解释性的需求主要来源于对问题和任务了解得还不够充分。而机器学习的目的是从数据中发现知识或解决问题[9]，那么在这个过程中只要提供给用户关于数据或模型的可以理解的信息，就可以更充分地发现知识、理解和解决问题。基于此，本文将详细分析可解释性的研究现状。围绕现有的可解释性成果，介绍可解释性的相关概念和解开深度学习模型的可解释性方法，为提高模型的可解释性提供进一步的研究思路。

2.1 什么是可解释性

可解释性[10]是指我们具有足够的可以理解的信息，来解决某个问题。具体到人工智能领域，可解释的深度模型能够给出每一个预测结果的决策依据，比如银行的金融系统[10]决定一个人是否应该得到贷款，并给出相应的判决依据。如图 1，分类器不仅要识别图片中的猫，而且要给出分类依据。比如决策树模型利用信息理论的筛选变量标准帮助理解不同变量对决策结果的影响程度，所以决策树模型是一个用户友好的可解释性模型。而用户最不友好的深度神经网络则属于黑盒模型，模型高度的非线性让人难以理解模型内部的决策过程，不能用人类可以理解的方式解释模型的具体含义和行为，所以深度学习模型不具有很好的解释性。

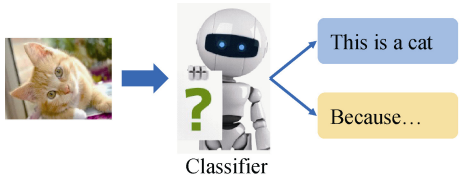

图 1 可解释的人工智能

Figure 1 Explainable Artificial Intelligence

根据可解释性的解释范围，我们可以分为全局可解释和局部可解释[11]。全局可解释是基于整个数据集中的因变量和预测变量之间的关系来理解模型的决策，即建立模型的输出和输入之间的关系。局部可解释是对单个数据点的决策进行解释，通常只需关注该数据点和该点周围特征空间中的局部子区域，并尝试基于该局部区域理解该点的模型决策。局部可解释和全局可解释通常结合使用，共同解释深度模型的决策结果。

全局可解释性：(1)整体模型的可解释性是指同时理解整个模型，并解释全局模型输出。比如输入特征对预测结果的重要性程度，以及不同特征之间的相互作用等等。这种可解释性是基于对模型、特征和习得的知识(如权重、参数、结构等)的整体看法来理解模型的决策结果，需要利用训练的模型、算法知识和数据，但是全局模型的可解释性在实践中很难实现。(2)模块化层面上的全局可解释性是在模块层面上理解模型，将全局模型的可解释性进行模块化。考虑到全局模型的可解释性通常无法实现，可以在模块层面上解释。比如从模型中解构出部分权重进行理解，尽管权重仅在模型中其他特征的上下文中有意义，但是这些权重仍然要比深度模型中所有的参数更好理解。

局部可解释性: (1)单一预测的局部可解释性是对模型的一个预测结果进行解释。局部可解释的预测结果可能只依赖于线性或某些单调的特征, 而非对它们有复杂的依赖性, 所以局部可解释性通常比全局可解释性更容易和准确。因此可以通过扰动输入, 观察输出的变化, 确定模型是基于哪些特征进行决策的, 以解释预测的原因。(2)一组预测的局部可解释性是对多个样本的预测结果进行解释。多个样本的模型预测可以用单一预测的局部可解释性方法来分别进行解释, 然后聚合为一组。此外, 也可以使用模块化的全局可解释性方法, 将获取的样本组视为完整的数据集, 然后使用包含此子集的全局方法来解释。

基于可解释性的相关概念, 目前主要从以下三个方面研究深度学习模型的可解释性:

(1) 使深度学习模型内部的组件尽可能变得透明、可理解, 这是基于模型结构的可解释性。比如可以通过可视化技术[12-15]来实现模型的透明化, 或者重新训练具有可解释性的模型。

(2) 从模型中解构出尽可能多的知识进行理解[16-21], 属于建模中的可解释性。比如从特征层面进行解释, 可以从深度神经网络中学习可理解的特征语义图, 以一种可解释性的方式对模型的知识进行表征。或者根据特征扰动对模型预测的影响, 判断特征的重要性程度。

(3) 生成人为可以理解的解释, 属于建模后的可解释性[22-25]。比如可以借助一些本质上可解释的模型对黑盒模型的预测结果进行事后解释。

2.2 如何解释深度学习模型

为了解构人工智能中的黑盒模型, 更好地理解模型的预测结果, 人们提出了很多可解释性方法。根据不同的标准, 这些可解释性方法可以分为不同的类别。例如, 建模中的可解释性是训练可解释的模型(如决策树、线性模型等); 建模后的可解释性是对模型的预测进行解释, 不依赖于模型的训练。基于解释黑盒模型的原理, 本文将这些可解释性方法大致分为:

模型内部可视化: 对模型内部学习的权重参数、神经网络的神经元或者特征检测器等进行可视化[26-29]。由于权重直接反映特征对模型最终预测的贡献, 所以可以非常粗暴地可视化出模型内部的权重。同理, 也可以对神经元或特征检测器可视化, 展示出输入特征在模型内部的变化。尽管这类可解释性方法可以直观地观察到模型内部输入的运算过程, 但是缺乏普适性, 很难得出通用的可解释性, 而且解释的效果也有待提升。

特征统计分析: 对不同的特征进行汇总统计或者显著性可视化, 以此建立特征和预测之间的因果关系[30-35]。许多可解释性方法根据决策结果对每个特征进行汇总统计, 并返回一个定量的指标, 比如特征重要性衡量不同特征对预测结果的重要性程度, 或者特征之间的交互强度。此外, 还可以对特征显著性统计信息进行可视化, 比如直观地展示出重要性特征的特征显著图, 或者显示特征和平均预测结果关系的部分相关图。特征统计分析方法主要是从特征层面上解释深度模型, 特征作为可解释性和模型之间的桥梁。

本质上可解释模型: 利用本质上可解释的模型近似模拟黑盒模型, 然后通过查看可解释性模型内部的参数或者特征统计信息来解释该黑盒模型[36-37]。比如借助可解释的决策模型或稀疏性的线性模型来近似黑盒模型, 可以通过蒸馏等方法, 在可解释的模型上建立输入和输出之间的关系, 实现可解释性的迁移。这种可解释模型近似的方法通常不考虑黑盒模型内部的参数, 直接对模型进行“端到端”的近似。下面将对这些可解释性方法进行详细的介绍, 见表 1。

表 1 可解释性方法

Table 1 Explanation Methods

如何解释深度学习模型	典型方法
模型内部可视化	可视化系统 CNNVis[38]; 可视化工具 Lucid[39];
特征统计分析	LIME[40]; CAM[41]; Grad-CAM[42]; 解释图表征[45];
本质上可解释模型	DLIME[47]; 决策树量化解释[48]; 决策树正则化[49]

2.2.1 模型内部可视化

由于人类对于世界的认知和感受主要来自于视觉, 良好的可视化可以有效地帮助人们理解深度网络模型的内部组件, 并进行有效的优化和调节。下面将介绍基于可视化技术来解释深度神经网络的典型方法。

可视化系统 CNNVis: 为了更好的理解神经网络内部的工作机制, 朱军等[38]提出了可视化系统 CNNVis。该系统首先提取神经元的衍生特征和神经元之间的连接, 然后对提取的这些层、神经元以及连接关系进行聚合, 最后实现对模型内部的可视化。该可视化系统包括有三个模块: (1)数据预处理模块: 把神经网络转换为一个有向无环图, 其中每一个神经元都是根据由一个节点和神经元之间的连接进行表征, 然后该模块可以计算神经元的派生特征和它们之间的联系; (2)聚合模块: 首先对特征图进行聚类,

从每个特征图集群中选择代表图层, 然后将神经元聚类在代表图层中, 并从每个神经元集群中选择代表性神经元; (3)可视化模块: 可视化出每一个神经元集群, 可以从中分析网络学习的特征、激活特征以及对结果的贡献等。CNNVis 还具有交互功能, 可以人为改变数据聚合过程, 从而更好的观察模型内部的运作过程。

可视化工具 Lucid: 是一个建立在 Deep Dream 上的神经网络可视化库, Deep Dream 是进行可视化神经网络理解图像方式的早期尝试, 而 Lucid[39]是改进后的用于研究神经网络可解释性的一套基础架构和工具。它提供顶尖的特征可视化技术实现和灵活的抽象, 使探索新的研究方向变得非常简单。Lucid 利用可视化技术研究神经网络自身的运行方式, 结合特征可视化和其他可解释性技术来理解神经网络如何决策。这种结合允许我们稍微“站在神经网络内部”, 看到神经网络在某一具体时刻如何决策, 及其如何影响最终输出。该技术通过可视化每个神经元, 能够看到哪个部分的检测器被激活。我们不仅可以看到检测结果, 而且能看到神经网络依据哪些特征来识别图片的。

可视化技术是深度学习模型可解释性研究的重要途径, 同时也是最直观的解释性方法。但是可视化方法也存在一定的局限, 一方面可视化神经网络得到的结果绝大部分依然是人类难以理解的, 而且也缺少对这种解释方法的评价标准, 从而会降低可视化解释结果的可信度; 另一方面, 可视化方法通常是和其他的解释性方法相结合, 而可视化技术作为最终解释结果的表征。可视化技术更多的用于局部可解释性方法, 以特征图的形式来解释深度神经网络的决策机制[10]。

2.2.2 特征统计分析

深度学习模型由于其内部复杂的结构, 会导致模型的特征和预测结果之间的因果关系难以理解, 所以模型是不可解释的。基于此, 可以通过对特征进行统计分析, 以建立特征和输出之间的因果关系, 从而实现模型内部的透明化。因此特征统计分析方法是指对深度模型的特征进行汇总分析或者显著性可视化, 对混乱的特征进行统计分析, 计算不同特征对模型输出的贡献, 并对显著性特征进行可视化。该方法是基于特征对模型进行解释, 以下是几种典型的基于特征统计分析的可解释方法。

(1) LIME

由于深度模型内部的特征经过复杂的变化, 并且不同特征之间也可能相互影响, 所以无法直接建立起某一特征和输出之间的关系。为了衡量特征对输出的贡献, 可以改变该特征值, 然后通过输出结果的变化判断该特征的重要性程度。

Marco[40]等人提出了 LIME(Local Interpretable Model-Agnostic Explanation)方法, 用人类可以理解的表征方式来解释分类模型, 该方法的核心思想是在预测结果的附近学习一个可解释性的模型, 实现对模型预测结果的局部可解释。LIME 方法通过向输入样本中添加扰动, 根据模型输出的变化来判断不同特征对预测结果的影响程度, 从而实现对黑盒模型决策过程的可解释。然后根据这些扰动的数据点距离原始数据的距离分配权重, 基于扰动后的样本学习一个可解释的模型。由于深度学习模型的决策边界是非线性的, 所以 LIME 方法是通过学习一个局部线性模型来解释样本的分类结果。输入样本中加入的必须是人类可以理解的扰动, 比如遮挡输入图片的某部分, 从而确保模型的预测结果一定会发生变化。由于该方法只是在输入值的周围做微小的扰动, 并没有深入模型内部, 所以 LIME 是和模型无关的可解释性方法。并且 LIME 方法在文本和图像分类领域都取得了很好的解释性效果, 极大地提高了人类对人工智能的信任。

LIME 和其他一些类似的方法[41-42]通过向输入中加入随机扰动, 或者选择输入中的某些特征, 以此生成对单个预测结果的解释。这些可解释性方法由于其简单易操作性而备受欢迎, 但是会导致解释结果的不稳定性, 即对于相同的预测结果, 模型给出的解释却有所不同。可解释性方法缺乏稳定性会影响解释结果的准确性, 从而降低人们对人工智能模型的信任。

(2) CAM & Grad-CAM

卷积神经网络的最后一个卷积层包含有丰富的语义和结构知识, 而全连接层的特征都是人类难以理解的。所以可以充分利用最后一个卷积层的特征来解释神经网络, 然后借助可视化技术理解模型内部的特征, 以实现对神经网络输出结果的解释。基于此, 我们将会介绍 CAM 和 Grad-CAM 两种方法。

CAM: 周博磊[43]提出了类别激活方法 CAM(Class Activation Mapping)来解释深层神经网络, 该方法的核心思想是在不使用任何边界框的前提下, 实现目标定位。CAM 方法引入了全局平均池化层(GAP)替换掉卷积神经网络中的全连接层, 然后得到最后一个卷积层中每个特征图的均值, 经过加权和之后就能得到实际的输出结果, 如图 2。此外, 该方法会强制最后一个卷积层生成和目标类别数量一致的特征图, 使经过 GAP 和 Softmax 层之后得到

分类结果, 从而可以为 GAP 输出的每个特征图赋予实际的含义。对模型输出结果进行解释时, 可以直接对 GAP 的输出进行可视化, 即以热力图的形式可视化出对应的特征图的加权和, 从而可以判断出对分类结果具有显著作用的特征。

由于 CAM 方法中没有全连接层, 所以模型对输入的尺寸没有要求, GAP 可以更加充分的利用空间信息。而且没有全连接层的参数也增加了模型的鲁棒性, 不易产生过拟合。但是 CAM 方法是通过修改模型的结构, 然后重新训练新的模型以实现对黑盒模型的解释。所以该方法会增加模型训练的成本, 而且极大地限制了模型的应用场景。

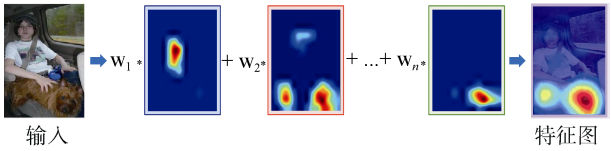

图 2 CAM 方法
Figure 2 CAM Method

Grad-CAM: 为了解决 CAM 方法存在的问题, Ramprasaath 等人提出了 Grad-CAM[44]方法生成对卷积神经网络的可视化解释, 该方法利用加权梯度类激活映射, 使任何目标特征的梯度经过最后一个卷积层后产生大致的局部特征图, 显示出图像中对目标预测分类重要的区域。Grad-CAM 对最终的加权和加了一个 ReLU 层, 原因在于我们只关心对类别有正影响的那些像素点, 如果不加 ReLU 层, 最终可能会带入一些属于其他类别的像素, 从而影响解释的效果。该方法将现有的细粒度可视化方法与 Grad-CAM 结合产生高分辨率的分类可视化特征, 并将其运用到图像分类, 图像文字描述以及视觉问答, 如图 3。Grad-CAM 适用于各种各样的 CNN 网络模型且不会改变网络结构, 也不需要重新训练。

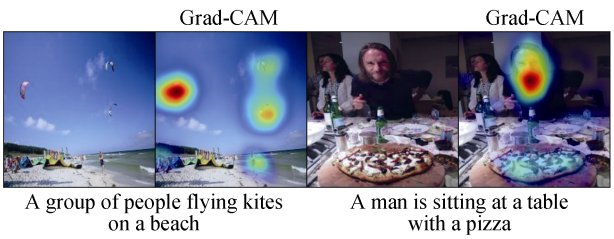

图 3 Grad-CAM 的图像文字描述解释
Figure 3 Image Captioning Explanations via Grad-CAM

(3) 特征图表征

特征图表征是指从深度神经网络中学习一个可解释性的语义图, 实现对模型内部知识的解构。利用学习到的语义图来表征深度学习模型的知识, 由于该语义图是可以人为理解的, 所以该语义图可以实现对深度模型的解释。

张拳石等人[45]提出了解释图的概念, 它揭示了隐藏在预训练的神经网络内部的知识层次, 即通过在深度神经网络内部学习一张解释图来实现对深度模型的解释。这种简单而有效的方法是以无监督的方式自动从过滤器的特征图中发现目标部分, 而无需标注信息。学习到的解释图有多层, 并且每一层对应于神经网络中的卷积层。解释图中的每个节点代表一个特定的部分, 从而可以从输入中解构出不同的目标部分, 如图 4。我们将目标部分与每个过滤器的特征分开。因此, 我们可以从单个过滤器中学习多个节点。解释图中的边对节点之间的共同激活关系和变形的空间关系进行建模。较高层中的节点代表较大的部分, 而较低层中的节点描述该部分的子区域。我们可以将解释图视为神经网络中间层特征的压缩, 通过使用数千个图节点来表示数百万个神经单位编码的信息, 从而实现对深度模型内部可解释性知识的解构。

图 4 解释图表征
Figure 4 Explanatory Graph Representation

由于神经网络中有非常复杂的语义和结构知识, 很难以人类可理解的方式完全对神经网络进行解释。而特征统计分析的方法则是通过对模型中的知识进行解构表征, 实现模型的可解释, 这种方法是对模型的隐藏知识进行逐步解释。

2.2.3 可解释模型

随着迁移学习的发展, 不仅能够实现模型结构的迁移, 我们也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, 通过将黑盒的深度学习模型迁移到这些可解释的模型中, 从而可以解构这些不可解释的模型。

(1) 线性模型

理解预测结果背后的原因对于评估模型的可信度很重要, 直接影响人工智能模型的实际应用, 而且有助于把不可信的模型或者决策结果变成可信的当人工智能模型用于社会中的实际问题时, 决策结果的可信度就非常重要。比如人工智能用于医疗诊

断[46], 人们要求模型给出诊断结果, 并且要能对结果进行解释, 一方面便于医生对整个决策过程进行监督, 另一方面确保人们可以基于模型的诊断结果采取行动。线性模型由于其结构的简单而具有良好的可解释性, 我们可以借助线性模型的可解释性来解开神经网络的黑盒特性。该方法的核心是训练一个线性模型来学习黑盒模型的输出结果, 从而可以在神经网络的输入和输出之间建立线性关系, 即可认为实现了对模型预测结果的可解释, 如图 5。

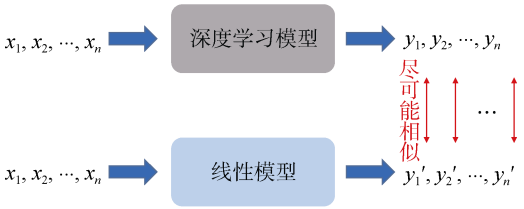

图 5 利用线性模型解释黑盒模型

Figure 5 Explaining Black-box Model via Linear Model

DLIME: LIME 方法中也使用了线性模型来解释深度模型的预测结果, 为了解决LIME中存在的问题, Muhammad 提出了 DLIME[47]方法(Deterministic LIME)。考虑到层次聚类的确定性和实施简单性, 而且层次聚类不需要聚类的先验知识, 所以 DLIME 方法首先使用层次聚类对数据集进行聚类分组, 然后生成一系列样本和相应的预测结果。不同于 LIME 方法直接加入随机扰动, 该方法使用 K 近邻选择出和测试样本最相似的近邻数据点, 然后利用选择出的样本重新训练一个线性回归模型, 以生成解释性结果。DLIME 方法中使用的是自下而上的聚合聚类, 根据最近的数据点和聚类之间的欧式距离计算相邻聚类之间的相似性, 其中很重要的一步的确定层次聚类的集群数目, 因为数据集的聚类数目可能会影响线性模型的解释效果。LIME 和其他基于随机扰动的可解释性方法在每一次迭代时生成的解释结果都在变化, 而实验结果表明 DLIME 方法生成的解释结果始终都是稳定的。但是该方法存在一个问题, 数据集中样本的数目会影响聚类的效果, 从而影响局部预测结果的准确性。

(2) 决策树模型

运用可解释性模型来解开深度神经网络的黑盒特性通常是基于模型预测结果进行的全局可解释, 该方法利用可解释性的模型来模拟黑盒模型的输出结果, 可以将复杂的深度模型迁移到可解释的模型中, 从而实现对模型决策结果的解释, 如图 6。

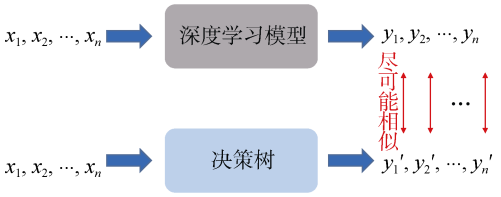

图 6 利用决策树模型解释黑盒模型

Figure 6 Explaining Black-box Model via Decision Tree

决策树量化解释: 张拳石等研究者使用决策树在语义层次上来量化解释深度神经网络预测的逻辑[48], 即对每个输入图像, 确定哪些物体部位被用于预测, 并量化测量每个物体部位对预测的贡献度。研究者通过略微修正神经网络而解开模型内部的知识表征, 并学习一种决策树来解释神经网络的预测结果。我们学习到一种分类物体的神经网络, 带有顶部卷积层的解开表征, 其中每个过滤层表征一个特定的物体部位。以一种由粗到精的方式, 决策树解码神经网络全连接层中隐藏的各种决策模式。给定一张输入图像, 我们来量化分析模型预测结果的基本原理。为了对 CNN 做出量化解释, 该方法学习 CNN 高层卷积层中物体部位的明确表征, 并挖掘存储在全连接层之中的潜在决策模式。然后决策树按照由粗到细的方式组织这些潜在的决策模式, 从而实现对模型的解释。

决策树正则化: 斯坦福大学的 Mike Wu 等[49]利用决策树模型的模仿性, 构建一个模拟决策树来逼近训练后的神经网络的预测结果, 从而实现对深度模型的可解释。但是训练深度神经网络时会出现很多局部极小值, 其中只有部分极小值容易模仿。因此, 用这种方法可能最后会陷于一个难以模仿的极小值(生成一个巨型决策树, 无法在合理时间内走完)。如果我们想在优化过程中提高模仿性, 则可以尝试找到更具可解释性的极小值。完美情况是, 我们训练一个行为非常像决策树的神经网络, 因为我们仍然想利用神经网络的非线性。另一种方式是使用简单决策树正则化深度神经网络, 我们称之为树正则化。给定决策树与数据集, 我们能计算平均路径长度以作为模拟、解释平均样本的成本。通过把这一项加入到目标函数, 我们就能鼓励神经网络生成简单的决策树并惩罚复杂而巨大的决策树。

3 可解释性的现状分析

为了提高黑盒模型的可解释性, 提出了很多可解释性方法。一方面可以在模型训练后应用模型分

析的方法使机器学习模型可解释, 即建立事后解释性。另一方面, 我们可以将可解释与模型学习分开, 即我们专注于与模型无关的解释方法。

对于深度学习模型, 可解释性主要体现在三个方面: 第一, 对于使用者来说, 如果人工智能的技术只是提一些建议或者帮助我们做决定, 那么做决定的人必须要能够理解这个决策, 为什么人工智能系统给他们提了这个建议。比如, 医生借助人工智能诊断疾病时, 要能理解为什么医疗诊断系统做这样的建议; 第二, 对于受到 AI(人工智能)影响的人, 如果 AI 自己做了决定, 那些受到决定影响的人要能够理解这个决定; 第三, 对于开发者来说, 理解了深度学习的黑盒子, 可以通过提供更好的学习数据, 改善方法和模型, 提高系统能力。而且提高深度学习模型的可解释性和透明度, 将有助于模型的除错、引导未来的数据收集方向、为特征构建和人类决策提供真正可靠的信息, 最终在人与模型之间建立信任[10]。

3.1 可解释性的实际应用

尽管人工智能已经广泛应用于许多领域, 但是具有可解释性的模型依然极度缺乏, 从而会严重影响深度学习的可信度。比如在安全领域, 人工智能在安全攻防方面展现了巨大的应用潜能, 但是安全从业者不理解深度模型的决策依据, 一方面无法信任模型的判别结果, 另一方面不能很好的诊断和追踪模型的错误, 这会极大地限制人工智能在该领域的实际应用。随着可解释人工智能的不断发展, 已经有很多相对成熟的可解释性方法, 并且已经成功应用于具体的领域, 比如可解释的医疗诊断系统、可解释的推荐系统, 以及可解释的金融算法模型。

3.1.1 医疗诊断系统

在医疗领域, 人工智能可以基于大量的病理数据进行疾病的诊断和预防, 医疗诊断系统借助深度学习方法训练计算机进行医学图像诊断, 比如根据病变图像识别早期癌症, 预测心脏疾病等。大量的实验数据证明, 人工智能用于医疗诊断的准确性甚至会超过医生, 而且深度学习完全是基于大数据进行诊断, 可以避免一些可能发生的主观错误, 所以人工智能医疗诊断系统在辅助医生诊断方面, 具有广泛的应用前景[50-54]。

医疗诊断系统在投入临床实践时, 受限于人工智能的不可解释性。理论上, 深度模型应该以医生可以理解的方式给出诊断结果, 即模型是基于哪些医疗图像特征和诊断标准进行推理的, 最终得出了什么样的诊断报告[55]。但是由于深度模型缺乏可解释性, 导致医生无法理解模型的诊断结果, 所以就极大地限制了医疗诊断系统的临床应用。此外, 当医生之间的诊断结果不一致时, 由人工智能医疗诊断系统提供参考性的意见则是非常重要的, 所以为了医疗诊断系统的临床应用, 人工智能的可解释性就显得尤为重要。下面我们将会介绍目前可解释性在医疗诊断系统的典型应用, 这是医疗诊断可解释性的重要进展。

Google 团队[56]研究了一种具有可解释性的心血管疾病医疗诊断系统, 主要利用视网膜图像作为参考, 来预测和心血管疾病相关的各种风险因素, 比如年龄、血压、吸烟史等的。然后基于这些风险指标直接预测心血管疾病发作的可能性, 这个医疗诊断系统能够以 70%的准确率识别出患有心血管疾病的图像。为了保证该诊断系统同时具有可解释性, 可以利用注意力机制自动生成一幅热力图, 用来显示对诊断结果具有显著影响的像素。此外, 加州大学圣地亚哥分校的张康教授[57]研发出了一种可以精确诊断致盲性视网膜疾病与肺炎的医疗诊断系统, 该系统可以在 30 秒内确定患者是否需要接受治疗, 并且具有 95%的诊断准确性。更重要的一点是, 这种医疗诊断系统具备可解释性。系统在给出诊断报告的同时, 会向人们显示决策过程中神经网络所激活的区域, 即可以提供诊断的依据。这种可解释的医疗诊断系统不仅可以为医生提供具有参考价值的诊断报告, 而且可以实现对病变部位的定位, 有助于为后续进一步的治疗提供帮助。

3.1.2 推荐系统

推荐系统是给用户推荐其感兴趣的内容, 并给出个性化的建议, 比如各式各样的购物消费平台。但是大多数的推荐系统只是给出最终的结果, 而缺少对推荐结果的解释。基于这些推荐系统的解释结果, 可以有依据的选择更明智、更准确的推荐结果, 从而提高用户对该推荐系统的信任程度[58-61]。

目前可解释性的推荐系统主要有嵌入式和后处理两种。嵌入式的可解释是向推荐系统中融入可解释的模块, 属于建模中的可解释, 嵌入式的方法适用于开发人员。可解释的模块通过选择输入对象的特征, 确定对推荐结果有显著影响的特征作为解释结果。比如向用户推荐物品时, 用作解释的物品特征可能是一些词语、语句等等。嵌入式的可解释方法通常具有很高的模型解释性, 但是受限于建模的困难, 解释的质量和连贯性难以保证。而后处理的可解释是对在推荐结果给出后进行解释, 属于事后解释, 后处理的方法则更适合向普通用户进行解释。可解释的模块只用于处理推荐结果, 和推荐系统无关,

所以解释内容不受推荐系统影响。这种解释方法适用于不同的推荐系统, 而且易于实现, 但是模型的解释性较差。

可解释的推荐系统通常以特征、用户和物品作为解释的参考依据, 推荐系统在对推荐对象进行序列建模时, 可以基于其中的细粒度特征对推荐结果进行解释[62], 这种以特征为媒介的解释需要判定用户对不同特征的感兴趣程度, 从中找出最适合用于解释的特征。基于用户进行解释的推荐系统需要对用户的喜好进行分析, 然后利用相似的用户作为解释的依据。以物品为媒介进行解释的推荐系统是根据用户的购买历史、浏览列表等进行推荐, 对这类推荐结果的解释会增加用户对系统的接受程度。目前具有可解释的推荐系统通常是将三者相结合, 比如可以借助知识图谱建立特征、用户和物品之间的关系, 然后根据被推荐对象和推荐系统选择合适的媒介进行可解释性的推荐。

Nan Wang 等人[63]开发了一种多任务的可解释性推荐系统(MTER), 以提高用户对推荐结果的满意度。该系统将用于推荐的用户偏好建模和用于解释的用户评论建模整合在一起, 不仅可以对用户的偏好进行推荐, 而且可以给出用户对特定商品有所偏好的依据, 即对推荐结果进行解释。MTER 系统可以利用用户的评论来给出被推荐物品的总体评估结果, 基于此可解释的推荐结果, 可以增强用户对该推荐系统的信任度。

3.1.3 金融风控

算法可解释性和透明性是一个重要的人工智能问题, 对算法的安全感、信赖感、认同度取决于算法的透明性和可理解性。在智能金融领域[64-65], 算法的透明性尤为迫切和重要。

深度学习模型的不可解释性严重影响人工智能在金融风控领域的应用, 比如一家银行使用人工智能产品推荐系统, 旨在帮助理财产品的交叉销售。但是由于管理人员无法解释模型建议背后的基本原理, 因此无法采纳这些建议。此外, 如果依据模型的不透明建议直接采取行动, 可能会带来严重的后果。金融风控模型所需的可解释性程度是银行根据风险偏好做出政策的关键, 比如将所有深度学习模型保持在相同的高标准可解释性或根据模型的风险进行区分。因此, 模型必须能够为决策提供明确的原因解释[66]。

索信达和香港大学团队已开发出一种新型基于网络结构约束的可解释性神经网络模型[67], 该模型在保持较高预测精确度的同时, 大幅度提升了模型的可解释性。该模型使用三种网络结构化约束: 稀疏可加子网络、正交投影和光滑函数, 其中稀疏可加子网络保证了子网络中岭函数的稀疏性, 即使得模型尽量简洁、紧致, 用最少的岭函数来构建模型。正交投影为数据旋转提供了正交基, 使得模型可辨识性增强。光滑函数使得岭函数更加光滑。与其他模型(如多层感知支持向量机、随机森林等)相比, 可解释性神经网络模型的预测精度并不低, 所以这是一种更简化的高精度新型可解释神经网络模型。索信达期待能将这种新型的可解释性机器学习模型大规模应用于银行业务中去, 帮助客户创造更大的价值。

2019 年 2 月, 波兰政府对银行法进行了修订, 赋予客户在做出信贷拒绝时获得解释的权利[68]。因此用户对决策过程享受知情权, 即如果决策过程是自动的, 银行需要解释做出决策的依据。如果银行使用基于机器学习的智能金融系统, 那么系统的可解释性就是至关重要的。因为我们不仅需要快速的决策结果, 而且要能对结果的准确性进行验证, 同时保证用户对整个系统的可理解性。具备可解释性的金融系统不仅能获得较高准确性的预测结果, 而且可以取得用户的信任, 减少深度学习模型在金融行业的应用局限。

3.2 可解释性的问题分析

人工智能在许多领域已经投入使用, 但是依然缺乏模型的可解释性研究, 具备可解释性的人工智能应用很少, 从而会导致模型的可信度和安全性降低[69]。神经网络的发展为机器学习和人工智能领域带来了显著的突破。复杂的网络结构层出不穷, 在计算机视觉和自然语言处理领域获得了极大的成功。除了模型的预测表现, 透明度和可解释性也是机器学习模型是否值得信赖的重要考核标准。然而, 大部分神经网络都是黑盒模型, 其内部的决策过程很难被人们了解。如果没有充分的可解释性, 这些模型在医疗、金融等领域的应用将受到很多限制。

随着深度学习模型的实际应用不断推广, 人们对模型的要求也在不断增加。在保证模型准确性的前提下, 如何提高模型的可解释性已经成为了研究的热点[70]。在涉及建模预测时, 高风险环境中使用的模型需要解释性, 因为我们可能会为预测的错误而付出巨大的代价[71], 而具备可解释性的模型在面对这些问题的时候可以对异常产生的原因进行追踪和定位。而低风险环境中, 可解释性同样也是很有价值的。比如电影推荐系统的错误并不会产生严重的后果, 人们更关心推荐结果。但是在产品的研发和部署阶段之后, 解释性可以为系统的调试和维护提供方向, 有利于理解报错的原因。基于此, 我们不仅要知

道模型预测的结果, 而且需要知道模型为什么会做出预测, 了解更多关于问题、数据以及模型可能失败的信息, 以规避模型预测的风险[72]。

为了增强神经网络模型的透明性, 研究者们已经探索出许多可解释方法来解读神经网络的决策结果, 但是目前可解释性的研究成果仍然无法满足对深度模型的要求, 主要还存在以下问题:

(1) 对深度模型进行解释的效果不理想。尽管已经提出了很多的可解释性方法, 有的方法也取得了不错的解释结果。但是目前所能实现的可解释性仍然达不到人们对神经网络的要求, 深度模型内部的运作机制依旧不是人为可以理解的方式。目前的可解释性研究更多的还是停留于初级探索阶段, 尤其在安全领域, 从而限制了人工智能的进一步应用。

(2) 缺乏统一的可解释性评价指标。由于可解释的概念偏向于抽象, 所以导致缺乏可信的评价指标, 更多的是定性评价可解释性方法, 缺少统一的定量指标。可解释性衡量指标的缺乏归根到底还是人们对人工智能的可解释性理解不够, 从而会影响可解释性的研究。

(3) 可解释性的应用领域有限。已有的方法主要是用于解释深度学习在图像分析领域的应用, 而在安全应用方面, 比如逆向工程和恶意软件分析领域, 缺乏可解释的研究。而且现有的方法通常有较低的解释精度。对于拥有模糊边界的应用而言, 比如图像识别, 相对较低的解释精度是可以接受的。但是对于安全应用, 比如二进制分析而言, 即使对于一个字节的解释偏差也会导致严重的误解或者错误。

此外, 目前这些可解释性方法极大地受限于算法、模型结构、应用场景等因素, 尽管可以用来解释深度学习模型的行为决策和预测结果, 但是在以下几种情况下, 可解释方法可能无法正常工作: (1)如果模型为互动建模, 比如随机森林。由于目前的可解释性方法仍然达不到实时、可互动的解释, 所以解释互动建模的模型有待研究; (2)特征是否相互关联, 特征之间的相互作用会极大地增加模型解释的难度, 不仅要考虑特征的显著性, 而且要评估不同特征之间的关联性对模型决策的影响。目前的可解释方法并没有考虑特征之间的相互作用; (3)如果模型没有正确地建模因果关系, 由于可解释方法直接对模型进行解释, 而缺乏前期对模型建模正确性的测试; (4)如果解释方法的参数设置不正确, 有些可解释方法很大程度上取决于超参数的设置, 比如LIME方法中的参数设置会影响解释结果, 参数的稳定性直接会影响可解释性的可信度。

4 总结与展望

目前的深度学习技术仍不完美, 有待于进一步提升,尤其是模型的可解释性问题。由于模型内部的参数共享和复杂的特征处理, 很难解释模型到底学习到了什么知识, 以及如何做出最终的决策。此外, 很难辨别通过深度学习训练出来的数学模型是如何获得特定的预测、推荐或决策的。因此深度学习模型即使能够完成目标任务, 获得的效用也可能有限, 特别是当预测或决策可能对个人、社会等产生不良影响时。在这种情况下, 用户有时需要知道运作背后的原理,例如为什么算法可以从具有法律影响的事实调查结果到具有监管影响的商业决策中给出推荐建议, 以及为什么某些因素在特定情况下如此重要。但是出于安全性考虑以及伦理和法律的需要, 算法的可解释性又是十分必要的。

尽管深度学习模型的可解释性已经取得了不错的研究成果, 但是如何生成可解释性结果是一个非常复杂的过程, 目前依然存在很多的挑战。首先是研究者对模型可解释性的重视程度仍然不够, 开发者关注更多的还是精确度, 而忽视了长期的用户体验; 其次是目前的可解释性算法过于复杂, 就极大地限制了可解释性模型的实际部署; 最后是可解释在实时性和普适性方面仍然需要改进, 而且缺乏一套通用的可解释性系统, 目前的可解释性对算法、模型和场景等有很强的依赖。由此可见, 深度模型的可解释性仍然有很长的路要走。

如何保持模型性能且具备可解释性将是未来一个重要研究方向。人工智能系统的可解释性并不是一个新问题, 随着深度学习的成功和采用, 它也在不断发展, 带来了更多样化、更先进的应用, 也带来了更多的不透明性。更大及更复杂的模型使我们很难用人类的语言来解释为什么会做出某种决定, 这是人工智能系统在应用领域的使用率仍然很低的原因之一。因此可解释性将会是未来研究的热点, 并且仍然有许多值得研究的方向, 比如可解释性的量化, 实现对模型可解释性的统一度量, 将可解释性指标纳入模型的评估体系; 利用更先进的认知理论模仿人脑的运作模式, 从而设计出可解释的深度学习模型; 研究实时可交互的智能人机系统[2], 在满足可解释性的前提下, 实现人机交互。

参考文献

[1] LeCun Y, Bengio Y, Hinton G. Deep Learning[J]. *Nature*, 2015, 521(7553): 436-444.

[2] Wu F, Liao B B, Han Y H. Interpretability for Deep Learning[J]. *Aero Weaponry*, 2019(1):39-46.
(吴飞, 廖彬兵, 韩亚洪. 深度学习的可解释性[J]. *航空兵器*, 2019(1):39-46.)

[3] Ras G, van Gerven M, Haselager P. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges[M]. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing, 2018: 19-36.

[4] Dong Y , Su H , Zhu J , et al, Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples[EB/OL]. 2017: ArXiv Preprint ArXiv:1901.09035.

[5] R. Guidotti, A. Monreale, S. Ruggieri, et al. A survey of methods for explaining black box models[J]. *ACM Comput. Surv*. 2018, 51(5):18-36.

[6] Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)[J]. *IEEE Access*, 2018, 6:52138-52160.

[7] Schwab P, Miladinovic D, Karlen W. Granger-Causal Attentive Mixtures of Experts: Learning Important Features with Neural Networks[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33:4846-4853.

[8] Humbird K D, Peterson J L, McClarren R G. Deep Neural Network Initialization with Decision Trees[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(5):1286-1295.

[9] Mitchell T M, Machine learning, https://www.springer.com/journal/10994, 1997.

[10] Bao W, Yue J, Rao Y L. A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-short Term Memory[J]. *PLoS One*, 2017, 12(7):e0180944.

[11] C. Molnar, Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://www.cornell.edu/video/kilian-weinberger-interpretable-machine-learning. 2019.

[12] Q. shi Zhang , S. chun Zhu. Visual interpretability for deep learning: a survey[J]. *Front. Inf. Technol. Electron. Eng*. 2018, 19(1): 27-39.

[13] J. Wang, L. Gou, W. Zhang, et al. Deepvid: deep visual interpretation and diagnosis for image classifiers via knowledge distillation[J]. *IEEE Trans. Vis. Comput. Graph*. 2019, 25(6): 2168-2180.

[14] B. Zhou. Interpretable representation learning for visual intelligence[C]. *MIT EECS*, 2018:256-263.

[15] J. Shi, H. Zhang, J. Li. Explainable and Explicit Visual Reasoning over Scene Graphs[EB/OL]. 2018: arXiv:1812.01855.

[16] B. Zhou, D. Bau, A. Oliva, et al. Interpreting Deep Visual Representations via Network Dissection[C]. *IEEE Trans. Pattern Anal. Mach. Intell*. 2018, 11(2):26-35.

[17] M. Aubry, B. C. Russell. Understanding deep features with computer-generated imagery[C]. *IEEE Int. Conf. Comput*. 2015: 2875-2883.

[18] Y. Geng, J. Chen, E. Jimenez-Ruiz, et al. Human-centric Transfer Learning Explanation via Knowledge Graph [EB/OL]. 2019: arXiv:1901.08547.

[19] X. Wang, D. Wang, C. Xu, et al. Explainable Reasoning over Knowledge Graphs for Recommendation[C]. *AAAI Conf*. 2019: 5329-5336.

[20] W. Zhang, B. Paudel, W. Zhang, et al. Interaction embeddings for prediction and explanation in knowledge graphs[C]. *WSDM*. 2019: 96-104.

[21] Y. Bai, H. Ding, S. Bian, et al. SimgNN: A neural network approach to fast graph similarity computation[C]. *WSDM*. 2019: 384-392.

[22] R. Chen, H. Chen, G. Huang, et al. Explaining Neural Networks Semantically and Quantitatively[EB/OL]. 2018: arXiv:1812.07169.

[23] Hohman F, Park H, Robinson C, et al. Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2020, 26(1):1096-1106.

[24] B. A. Plummer, M. I. Vasileva, V. Petsiuk, K. Saenko, et al. Why do These Match? Explaining the Behavior of Image Similarity Models[EB/OL]. 2019: arXiv:1905.10797.

[25] S. M. Lundberg. Explainable AI for Trees: From Local Explanations to Global Understanding[J]. *Nature Machine Intelligence*, 2019: 1-72.

[26] G. Castañón , J. Byrne. Visualizing and quantifying discriminative features for face recognition[C]. *13th IEEE Int. Conf. Autom. Face Gesture Recognition*, 2018: 16-23.

[27] RichardWebster B, Kwon S Y, Clarizio C, et al. Visual Psychophysics for Making Face Recognition Algorithms more Explainable[M]. Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 263-281.

[28] W. Nie, Y. Zhang, A. B. Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations[C]. *35th Int. Conf. Mach. Learn. ICML*. 2018: 6105-6114.

[29] Zhou B L, Sun Y Y, Bau D, et al. Interpretable Basis Decomposition for Visual Explanation[M]. Computer Vision–ECCV 2018. Cham: Springer International Publishing, 2018: 122-138.

[30] S. Hooker, D. Erhan, P.-J. Kindermans,et al. Evaluating Feature Importance Estimates[EB/OL]. 2018: arXiv:1806.10758.

[31] Ventura F, Cerquitelli T, Giacalone F. Black-Box Model Explained through an Assessment of Its Interpretable Features[J]. *New Trends in Databases and Information Systems*, 2018: 138-149.

[32] R. C. Fong , A. Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation[C]. *IEEE Int. Conf. Comput*. 2017: 3449-3457.

[33] V. Petsiuk, A. Das, K. Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models[EB/OL]. 2018: arXiv:1806.07421.

[34] Q. Zhang, W. Wang, S. C. Zhu. Examining CNN representations with respect to dataset bias[J]. *AAAI Conf*. 2018: 4464-4473.

[35] C. F. Baumgartner, L. M. Koch. Supplementary Material for : Visual Feature Attribution using Wasserstein GANs[EB/OL]. 2018: arXiv:1711.08998.

[36] B.-J. Hou , Z.-H. Zhou. Learning with Interpretable Structure from RNN[EB/OL]. 2018: arXiv:1810.10708.

[37] A. Warnecke, D. Arp, C. Wressnegger, et al. Don't Paint It Black: White-Box Explanations for Deep Learning in Computer Security[EB/OL]. 2019: arXiv:1906.02108.

[38] Liu M C, Shi J X, Li Z, et al. Towards Better Analysis of Deep Convolutional Neural Networks[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2017, 23(1):91-100.

[39] Olah, Christopher. The Building Blocks of Interpretability[J]. *Distill*, 2018, 3(3):18-36.

[40] M. T. Ribeiro, S. Singh, C. Guestrin. 'Why should i trust you?' Explaining the predictions of any classifier[C]. *ACM SIGKDD*. 2016: 1135-1144.

[41] Lengerich, Benjamin J. Visual Explanations for Convolutional Neural Networks via Input Resampling[EB/OL]. 2017: ArXiv Preprint ArXiv:1707.09641.

[42] Wagner, Jorg. Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Network[C]. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019: 9097 - 9107.

[43] B. Zhou, A. Khosla, A. Lapedriza et al. Learning Deep Features for Discriminative Localization[EB/OL]. 2018: arXiv:1512.04150.

[44] R. R. Selvaraju, A. Das, R. Vedantam, et al. Grad-CAM: Why did you say that?[EB/OL]. 2016: arXiv:1611.07450.

[45] Q. Zhang, R. Cao, F. Shi, et al. Interpreting CNN knowledge via an explanatory graph[C]. *AAAI Conf*. 2018: 4454-4463.

[46] Lekas H M, Alfandre D, Gordon P, et al. The Role of Patient-provider Interactions: Using an Accounts Framework to Explain Hospital Discharges Against Medical Advice[J]. *Social Science & Medicine*, 2016, 156:106-113.

[47] M. R. Zafar, N. M. Khan. DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems[EB/OL]. 2019: arXiv:1906.10263.

[48] Q. Zhang, Y. Yang, H. Ma, et al. Interpreting CNNs via Decision Trees[EB/OL]. 2018: arXiv:1802.00121.

[49] M. Wu, M. C. Hughes, S. Parbhoo, et al. Explaining Classifiers with Causal Concept Effect [EB/OL]. 2019: arXiv:2005.02817.

[50] Zhang, Zizhao. MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network[C]. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017: 3549 - 3557.

[51] Li, Jingyuan. Discovering Interpretable Medical Workflow Models[C]. *IEEE International Conference on Healthcare Informatics (ICHI)*, 2018: 437-439.

[52] Sriram, Aditya. Projectron - A Shallow and Interpretable Network for Classifying Medical Images[C]. *International Joint Conference on Neural Networks (IJCNN)*, 2019:1-9.

[53] Holzinger, Andreas. What Do We Need to Build Explainable AI Systems for the Medical Domain[EB/OL]. 2017: ArXiv Preprint ArXiv:1712.09923.

[54] Xie, Yao. Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis[C]. *IUI Workshops*, 2019:256-263.

[55] Kermany, Daniel S. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning[J]. *Cell*, 2018, 172(5): 1122-1131.

[56] Poplin, Ryan. Predicting Cardiovascular Risk Factors in Retinal Fundus Photographs Using Deep Learning[J]. *Nature Biomedical Engineering*, 2018, 2(3): 158-164.

[57] E. Pastor ,E. Baralis. Explaining black box models by means of local rules[C]. *ACM Symp. Appl. Comput*.,2019: 510-517.

[58] Zhang X J, Zhao Z Y, Li C, et al. An Interpretable and Scalable Recommendation Method Based on Network Embedding[J]. *IEEE Access*, 2019, 7:9384-9394.

[59] Yu S, Wang Y B, Yang M, et al. NAIRS: A Neural Attentive Interpretable Recommendation System[C]. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 2019: 790-793.

[60] Hu L, Jian S L, Cao L B, et al. Interpretable Recommendation via Attraction Modeling: Learning Multilevel Attractiveness over Multimodal Movie Contents[C]. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018: 3400-3406.

[61] J. Gao, X. Wang, Y. Wang et al. Explainable Recommendation through Attentive Multi-View Learning[C]. *AAAI Conf. Artif. Intell*. 2019: 3622-3629.

[62] Z. Sun, J. Yang, J. Zhang. Recurrent knowledge graph embedding for effective recommendation[C]. *RecSys*. 2018: 297-305.

[63] Wang X, He X N, Feng F L, et al. TEM: Tree-enhanced Embedding Model for Explainable Recommendation[C]. *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 2018: 1543-1552.

[64] Sohangir, Sahar. Big Data: Deep Learning for Financial Sentiment Analysis[J]. *Journal of Big Data*, 2018, 5(1): 3.

[65] Siami-Namini, Sima, Akbar Siami Namin. Forecasting Economics and Financial Time Series: ARIMA vs. LSTM[EB/OL]. 2018: ArXiv:1803.06386.

[66] Wang D, Quek C, Ng G S. Bank Failure Prediction Using an Ac-

curate and Interpretable Neural Fuzzy Inference System[J]. *AI Communications*, 2016, 29(4): 477-495.

[67] Suoxinda AI Lab, Team of Professor Aijun Zhang at the University of Hong Kong. The application of interpretable neural networks in fintech—a case study of a joint-stock bank. 2019.
(索信达 AI 实验室, 香港大学张爱军教授团队.可解释神经网络在金融科技的应用—某股份制银行应用案例研究[R]. 2019.)

[68] Please explain. Interpretability of black-box machine learning models, https://appsilon.com/please-explain-black-box/, April, 2019.

[69] Q. Zhang, Y. N. Wu, , S. C. Zhu. Interpretable Convolutional Neural Networks[C]. *IEEE Comput*. 2018: 8827-8836.

[70] M. Du, N. Liu, X. Hu. Techniques for Interpretable Machine Learning[EB/OL]. 2018: arXiv:1808.00033.

[71] L. Chen, C. Yagemann, E. Downing. To believe or not to believe: Validating explanation fidelity for dynamic malware analysis. Tracking botnets. http://www.honeynet.org/papers/bots, 2005.

[72] W. Guo, D. Mu, J. Xu et al. Lemna: Explaining deep learning based security applications[C]. *ACM Conf. Comput. Commun. Secur*. 2018: 364-379.

化盈盈 于 2018 年在南开大学通信工程专业获得学士学位。现在中国科学院信息工程研究所攻读博士学位。研究领域为深度学习和计算机视觉。研究兴趣包括: 可解释的人工智能、神经网络的对抗防御等。Email: huayingying@iie.ac.cn

葛仕明 于 2008 年在中国科学技术大学电子工程专业获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为计算机视觉、深度学习与人工智能安全。研究兴趣包括: 联邦学习、知识蒸馏、信息不充分条件下的机器学习问题及其现实应用等。Email: geshiming@iie.ac.cn

张岱墀 现在清华大学攻读学士学位。将于 2020 年在中国科学院信息研究工程所攻读博士学位, 研究领域为深度伪造与检测, 人工智能可解释性。研究兴趣包括: 人工智能可解释性的表征与量化、深度伪造检测深层次机理、基于时序特征的深度检测算法等。Email: zhang_daichi@163.com