计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000044

深度学习可解释性综述

陈 冲 陈 杰 张 慧 蔡 磊 薛亚茹
中国石油大学(北京)信息科学与工程学院 北京 102249

摘 要 随着数据量呈爆发式增长,深度学习理论与技术取得突破性进展,深度学习模型在众多分类与预测任务(图像、文本、语音和视频数据等)中表现出色,促进了深度学习的规模化与产业化应用。然而,深度学习模型的高度非线性导致其内部逻辑不明晰,并常常被视为"黑箱"模型,这也限制了其在关键领域(如医疗、金融和自动驾驶等)的应用。因此,研究深度学习的可解释性是非常必要的。首先对深度学习的现状进行简要概述,阐述深度学习可解释性的定义及必要性;其次对深度学习可解释性的研究现状进行分析,从内在可解释模型、基于归因的解释和基于非归因的解释3个角度对解释方法进行概述;然后介绍深度学习可解释性的定性和定量评估指标;最后讨论深度学习可解释性的应用以及未来发展方向。

关键词:深度学习;可解释性;归因解释;非归因解释;评估方法
中图法分类号 TP181

Review on Interpretability of Deep Learning

CHEN Chong,CHEN Jie,ZHANG Hui,CAI Lei and XUE Yaru
College of Information Science and Engineering,China University of Petroleum(Beijing),Beijing 102249,China

Abstract With the explosive growth of data volume and the breakthrough of deep learning theory and technology,deep learning models perform well enough in many classification and prediction tasks(image,text,voice and video data,etc.),which promotes the large-scale and industrialized application of deep learning. However,due to the high nonlinearity of the deep learning model with undefined internal logic,it is often regarded as a "black box" model which restricts further applications in key fields(such as medical treatment,finance,autonomous driving). Therefore,it is necessary to study the interpretability of deep learning. Firstly,recent studies on deep learning,the definition and necessity of explaining deep learning models are overviewed and described. Secondly,recent studies on interpretation methods of deep learning,and its classifications from the perspective of intrinsic interpretable model and attribution-based/non-attribution-based interpretation are analyzed and summarized. Then,the qualitative and quantitative performance criteria of the interpretability of deep learning are introduced. Finally,the applications of deep learning interpretability and future research directions are discussed and recommended.

Keywords Deep learning,Interpretability,Attribution-based interpretation,Non-attribution-based interpretation,Evaluation method

1 引言

近年来,深度学习模型[1]在医疗健康[2-3]、图像处理[4-5]、机器翻译[6-7]、自动驾驶[8]领域应用广泛,并取得了卓越的性能。然而,深度神经网络(Deep Neural Networks,DNN)是一个具有高度非线性的复杂结构,本质上可以将其视为一个"黑箱"模型。显然,这种不透明建模技术在许多领域无法得到信任,严重限制了深度学习模型在许多敏感或高风险领域的使用。例如,在医学领域,使用深度学习模型的目的是构建应用程序,帮助医生做出医疗决策。但是,深度学习模型中有数百万个参数,用户给模型一个输入,只是返回一个决策结果,并不知道模型内部具体的决策机制,可解释性较差,导致医学专家不信任模型的决策结果。在自动驾驶领域,深度学习模型的决策结果直接关乎人的生命安全。出于对安全的考虑,人们需要了解深度学习模型做出某个特定决策的原因,获得用户的信任。因此,深度学习模型可解释性亟待提高,研究深度学习模型的决策结果如何被人类理解以及如何使深度学习模型尽可能透明变得迫在眉睫。

DNN的解释已经受到了国内外研究人员的普遍关注,并且涌现出了大量的研究成果,如以"(Interpretable or Explanation or Explainability or Interpretability) and (Deep Learning or Deep Neural Networks)"为主题词在Web of science核心

到稿日期:2022-10-09 **返修日期**:2023-02-27
基金项目:国家自然科学基金(62006247);国家重点研发计划(2019YFC1510501,2022YFC2803704)
This work was supported by the National Natural Science Foundation of China(62006247) and National Key R&D Program of China(2019YFC1510501,2022YFC2803704).
通信作者:陈冲(chenchong@cup.edu.cn)

合集中检索2010年1月至2022年8月的英文文献，可以发现论文数量呈指数级增长，如图1所示。

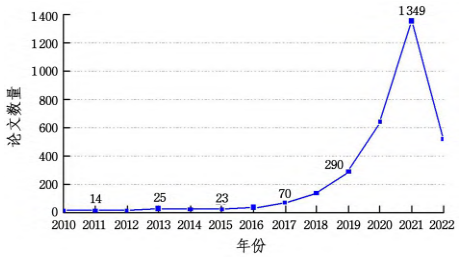

图1　不同年份可解释深度学习文献数量变化

Fig. 1　Changes in the number of explainable deep learning literature in different years

国内外学者围绕可解释人工智能（Explainable Artificial Intelligence，XAI）领域从不同的研究角度和侧重点进行概述，Arrieta 等[9]阐述了可解释的概念和相关术语，并从设计透明模型（Transparent Models）和事后解释（Post-hoc Explainability）两个角度对现有的解释技术进行归纳总结。为了审查和理解 XAI 方法，Arrieta 等[9]创建了一套可以统一类别和概念的框架；Murdoch 等[10]从机器学习可解释方法评估的角度引入预测（Predictive）、描述（Descriptive）和相关（Relevant）PDR 框架，并为评估可解释方法提出了 3 个要求，即预测准确性、描述准确性和相关性，并且从基于模型和事后分类的角度对解释方法进行归纳汇总；Samek 等[11]专注于研究事后解释，为深度学习可解释方法提供了理论基础，并在典型的应用场景（人脸识别、语音识别）中演示如何使用 XAI 方法。国内，Su 等[12]对卷积神经网络（Convolutional Neural Networks，CNN）、循环神经网络（Recurrent Neural Network，RNN）和生成对抗网络（Generative Adversarial Networks，GAN）的解释方法进行分析梳理；Zeng 等[13]从自解释模型、特定模型解释、不可知模型解释和因果可解释 4 个方面对可解释方法进行了总结分析。最近，Lei 等[14]从解释深度学习模型的逻辑规则、决策归因和内部结构表示 3 个方面，介绍了可解释研究的几种模型和算法；Li 等[15]利用 CiteSpace 工具对检索的文章进行可视化分析，从被动解释、主动干预解释和补充解释等方面对现有可解释研究工作进行了分析。

本文立足于深度学习的可解释性，从模型内在可解释性、基于归因的解释和非归因的解释 3 个方面对涉及不同领域和决策任务的可解释方法进行系统综述。首先，对可解释的定义及解释的原因进行阐述；其次，对可解释方法的研究现状进行分析，将深度学习可解释方法重新分类汇总，将其分为三大类，即内在可解释模型、基于归因的解释和基于非归因的解释，并对每种解释方法进行详细概述；然后，介绍深度学习可解释性的定量和定性评估指标；最后，概述了可解释性的应用以及未来发展方向。图 2 给出了本文的内容框架。

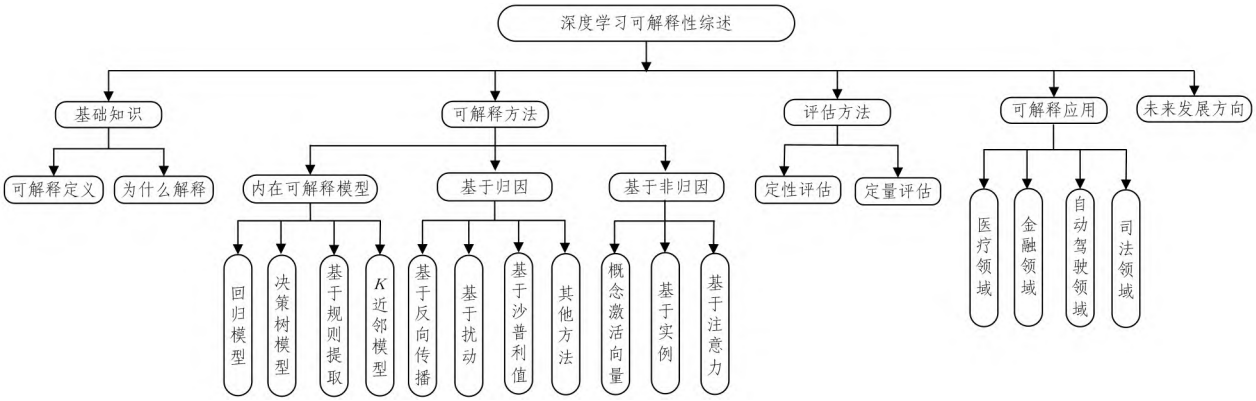

图2　综述内容框架图

Fig. 2　Overview content framework

2　可解释性定义及原因

2.1　可解释性定义

描述可解释性的单词主要有解释（Interpretability 或 Explainability）、理解（Understandability）。Arrieta 等[9]从模型的特征这方面给出了每个词的定义。Interpretability 表示向人类解释或提供意义的能力，即将抽象概念（如预测类别）映射到人类可以理解的领域。例如，图像（像素阵列）或文本（单词序列）是可解释的，人们可以查看、阅读它们，而抽象向量空间（单词嵌入）是不可解释的。Explainability 是连接人类和模型决策之间的接口。解释会产生输入特征对输出的贡献程度，其贡献由相关分数表示。例如，在图像中，解释可以是热力图的形式，代表输入图像的哪些像素支持分类决策；在自然语言处理（Natural Language Processing，NLP）中，解释可以用高亮文本表示。Understandability 是 XAI 最重要的概念，表示让人类理解模型如何工作，而无须解释其内部结构或模型内部处理数据的算法。

目前，对解释的研究横跨认知科学、心理学与哲学等领域。但是，各个学科对可解释性并没有一个明确的定义。较为流行的可解释性定义是 2017 年在 ICML 会议中提出的："Interpretation is the process of giving explanations to human"，意思是，"解释是向人类解释的过程"。Arrieta 等[9]将受众作为解释机器学习模型的关键因素，提出"Given a certain audience，explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand"。这表明模型必须为特定受众提供功能清晰或容易理解的细节和原因。Doshi-velez[16]等在 2021 年提出"Interpretability is the ability to explain or to present in understandable

terms to a human",也就是"以可理解的方式向人类解释或呈现的能力"。综上所述,解释的最终目的是让特定受众理解模型的决策原因和推理过程,建立模型和受众之间相互信任的关系。

2.2 解释模型的必要性

正如引言所述,可解释性是人工智能在解决实际问题中面临的主要障碍之一。大多数深度学习模型不透明原因是神经网络结构由连续的非线性变换层和可调整的权重和偏差组成,这些层与层之间的单个转换在数学上很容易理解,但是一旦神经网络经过训练,就会因网络结构的高度非线性而很难推断出它在权重和偏差之间以及从每个变换到下一个变换之间是如何组合信息、获得最终输出的。因此,深度学习模型应用在金融、司法和医疗等领域时,即使模型达到了期望的输出,但缺乏对模型的理解也会导致许多问题。例如,模型可能过度拟合数据,建立和输入噪声之间的映射关系,而不是捕获输入和输出之间有意义的映射。此外,在许多领域(地球物理、石油勘探等),样本量非常有限,输出和噪声的映射可能导致网络无法驱动输入和输出之间正确的建模关系。因此,解释深度学习模型的决策原理对不同的领域而言都至关重要。

3 深度学习可解释方法

本节首先介绍内在可解释性模型(回归、决策树、基于规则提取和 K 近邻),其次从归因和非归因两个角度介绍深度学习可解释方法。

3.1 内在可解释模型

内在可解释性指模型根据人类观察的决策边界或特征来解释决策的能力。Lipton[17]指出内在可解释方法分为 3 个层次的解释:可模拟性、可分解性和算法透明性。这 3 个层次分别表示:1)模型被人类模拟的能力;2)解释模型输入、参数和输出的能力;3)解释算法运行的能力。线性/逻辑回归、决策树和 K 近邻等简单模型往往符合这 3 个层次中的一个或多个。然而,对于 DNN 而言,添加到每个隐藏层中的非线性函数使得模型输出难以解释。

3.1.1 线性/逻辑回归模型

线性回归模型是利用一组自变量预测连续的因变量;而逻辑回归是一种用于预测二分类任务的模型。尽管线性/逻辑回归具有内在可解释性,但由于其可解释性与特定受众有关,因此向非专家受众解释模型决策时,可能会需要事后解释技术(主要是可视化)。Dingen 等[18]提出了一种可视化分析工具(RegressionExplorer),允许用户交互式地探索逻辑回归模型。该工具应用在临床生物统计学领域,有助于专家快速生成、评估和比较不同的模型,充分探索待选模型参数值的全局模式,以便协助专家提出新理论或开发新模型。

3.1.2 决策树模型

决策树模型是一种基于分层结构的机器学习算法,它满足透明模型的所有约束[17],具有很好的可解释性,因此衍生出许多基于决策树的深度学习模型。Zilke 等[19]介绍了 DNN 中的具体挑战,并且提出了一种基于决策树的 DNN 规则提取(Deep neural network Rule Extraction via Decision tree induction, DeepRED)方法,该算法将 CRED 算法[20](为浅层网络设计)扩展到 DNN 的多个隐藏层,通过分析权重值,在每个隐藏层神经元和输出神经元水平上提取规则。此外,Nguyen 等[21]提出了一种精确可转换决策树(Exact-Convertible Decision Tree, EC-DT)算法。该算法将具有校正线性单元激活函数的神经网络转换为提取多元规则的代理树,因此可以有效地从神经网络中提取重要规则。

3.1.3 基于规则的提取

基于规则的提取是解释决策的另一种方法,通过生成基本的、直接的规则来表示所学习的数据,如 if-then 规则或者表示知识的几个规则的组合。基于规则的算法旨在设计可解释和直观的模型,但是模型可解释性受到生成规则的长度和数量的影响,大量的规则可能不利于模型的解释。一种解决方法是将基本规则转化为模糊规则[22],使用模糊逻辑和模糊集表示各种形式的知识,并对变量之间的交互关系进行建模。而基于模糊的规则是透明模型,因为它将深度学习模型和模糊逻辑相结合来创建更易于理解的深层网络[23]。因此,基于规则的提取已被广泛应用于深度学习模型中。例如,Keneni 等[24]通过基于规则的 Sugeno 模糊推理,为无人机设计了一个可解释的转向控制框架。该框架的设计分为两个阶段:1)指导无人机按照指定任务飞行,并记录其在遭遇不同天气状况和不同飞行模式时采取的一系列动作,完成数据收集;2)在数据上使用减法聚类算法训练 Sugeno 型模糊推理模型,通过访问模型的性能和规则数量优化减法聚类算法中的参数,并使用自适应网络模糊推理系统(Adaptive-Network-based Fuzzy Inference System, ANFIS)对模型进行微调,以提供无人机决策的可解释特征。

3.1.4 K 近邻模型

K 近邻(KNN)算法是通过样本之间的距离,度量其邻域关系,从而选出某一样本附近的 K 个邻居样本,达到分类的效果。在可解释性方面,KNN 模型的输出取决于测量样本之间相似性的距离函数,因此可以清晰地知道当 K 改变时,输出是如何更新的。因此,KNN 被广泛应用在深度学习可解释中。Zheng 等[25]提出了一种基于原始 K 近邻算法的稀疏 KNN 分类器(Group Lasso Sparse KNN Classifier, GL-SKNN),该分类器利用稀疏组套索选择最相关的类,并提取最大信噪比的稀疏特征,最后将回归权重总和作为预测类指标,以此提高分类精度并使模型具有可解释性。但是,KNN 的可解释性严重依赖于特征数量、K 邻居数量和样本之间相关性的距离函数,高 K 值阻碍了模型的可模拟性,而大量特征或复杂的距离函数又阻碍了模型的可分解性。

综上所述,构建内在可解释模型只有在合理的情况下才能有效解释。当存在高维线性模型、树太深或者规则集太大时,即使可以捕获"黑盒"的内部逻辑,也无法提供人类可理解的解释。

3.2 基于归因的方法

基于归因的解释方法旨为网络的每个输入特征分配归因值,得到输入特征对模型决策结果的重要程度。例如,DNN 接受一个输入样本 $x=[x_1, x_2, \cdots, x_N]$ 产生输出 $S(x)=[S_1(x), S_2(x), \cdots, S_c(x)]$,其中 N 是特征总数,c 是输出神经元总数。给定一个目标神经元,归因方法的目标是确定每个

输入特征 x_i 对输出 $S(\boldsymbol{x})$ 的相关性或贡献。在下文中，我们将归因方法分为基于反向传播的方法、基于扰动的方法、基于 Shapley 值的方法和其他方法。

3.2.1　基于反向传播的方法

基于反向传播的方法并非无视要解释的模型，而是将模型的内部结构整合到解释过程中。基于反向传播的方法是利用反向传播识别输入图像中用于决策的特征，从而生成事后归因映射。

Simonyan 等[26]提出的显著性映射(Saliency maps)方法，使用反向传播计算网络输出分类得分函数 $S_c(\boldsymbol{x})$ 相对于输入 $\boldsymbol{x}$ 的梯度 Φ_{Sal}：

$$\Phi_{\text{Sal}}(S_c,\boldsymbol{x})=\nabla S_c(\boldsymbol{x}) \tag{1}$$

显著性映射方法是一种局部解释方法，其反映了 DNN 某一部分的输入图像区域对预测结果的重要性。Shrikumar 等[27]提出的输入 X 梯度方法(Input X Gradient)扩展了显著性方法，使用梯度和输入元素的乘积得到输入 $\boldsymbol{x}$ 对预测结果的最终贡献 $\Phi_{\text{Input X Grad}}$：

$$\Phi_{\text{Input X Grad}}(f_c,\boldsymbol{x})=\boldsymbol{x}\odot\nabla S_c(\boldsymbol{x}) \tag{2}$$

实际上，输入 X 梯度方法是每个特征对近似线性化模型输出的总贡献，它的优势在于允许利用输入信息进行更好的可视化。

反卷积(Deconvolution)主要用在 CNN 的解释中。ZFNet[28]在反向传播过程中将负梯度设置为零生成特征图，然后在 AlexNet[29]中的 5 个卷积层上进行反卷积和特征可视化，从而实现解释的目的。Springenberg 等[30]提出的引导反向传播(Guided Backpropagation)方法是对反卷积方法的改进，它的工作原理是计算输出 $f_c(\boldsymbol{x})$ 对输入 $\boldsymbol{x}$ 的梯度，并且为了找出图片中的最大激活特征，其在反向传播过程中将负梯度设置为零。Sundararajan 等[31]提出的积分梯度(Integrated Gradients，IG)方法定义为：从基线 $\boldsymbol{x}'=(x_1',\cdots,x_D')$ 到输入 $\boldsymbol{x}=(x_1,\cdots,x_D)$ 的直线路径上梯度的路径积分 Φ_{IG}^d 为：

$$\Phi_{\text{IG}}^d(f_c,\boldsymbol{x})=(\boldsymbol{x}_d-\boldsymbol{x}_d')\times\int_{\alpha=0}^{1}\frac{\partial F(\boldsymbol{x}'+\alpha\times(\boldsymbol{x}-\boldsymbol{x}'))}{\partial\boldsymbol{x}_d}\mathrm{d}\alpha,\quad\forall\, d\in\{1,\cdots,D\} \tag{3}$$

其中，基线 $\boldsymbol{x}'$ 是原始图像 $\boldsymbol{x}$ 中某个特征不出现时的输入图像，例如 $\boldsymbol{x}'$ 可以是黑色或者全零图像，甚至可以是随机噪声。IG 满足完整性公理[31]，也就是归因必须考虑基线 $\boldsymbol{x}'$ 和输入 $\boldsymbol{x}$ 的输出差异。这种方法需要一个参考基线，这个额外的输入使解释方法变得更加复杂。此外，由于 IG 方法需要的样本数较多，可能会非常耗时。平滑梯度(Smooth Grad，SG)[32]是一种向图像中添加噪声以生成新图像的技术，每次将随机高斯噪声 $\mathrm{N}(0,\sigma^2)$ 添加到给定的输入图像中，并计算相应的梯度。

$$\Phi_{\text{SG}}(\boldsymbol{x})=E[\Phi(\boldsymbol{x}+\mathrm{N}(0,\sigma^2))] \tag{4}$$

这种方法可以理解为一种平均化的过程，可以使初始解释 Φ_{SG} 更平滑。但是，由于样本数量增多，这种方法也会增加计算时间。

Zhou 等[33]提出的类激活映射(Class Activation Maps，CAM)是一种可视化方法，其以热力图的形式表示对分类结果影响较大、特定于某一输出类别的图像区域。为了生成 CAM 图，在最后一个卷积层之后添加一个全局平均池化层(Global Average Pooling，GAP)，将 GAP 的输出进行线性组合以生成类预测。然后，通过激活最后的卷积层并计算加权和，得到每个类的 CAM(记为 Φ_{CAM}^c)。

$$\Phi_{\text{CAM}}^c(f_c,\boldsymbol{x})=\sum_{k=1}^{n}w_k^cA^k \tag{5}$$

其中，w_k 表示第 k 个神经元的权重，$A^1,A^2,\cdots,A^n$ 表示 CNN 最后一层的 n 张特征图。但是，将 CNN 中的分类器替换之后需要重新训练模型才能得到 GAP 的权重。Selvaraju 等[34]在 CAM 的基础上提出梯度类激活映射(Gradient-Class Activation Maps，Grad-CAM)方法，该方法克服了上述缺点，根据特定层的特征图 A^k 计算输出 $f_c(\boldsymbol{x})$ 的梯度，此时，对每个通道 k(沿宽度 W 和高度 H)的梯度进行平均，以获得特定层的特征图对目标分类 c 的重要性权重 a_k^c：

$$a_k^c=\frac{1}{H\cdot W}\sum_{i}^{W}\sum_{j}^{H}\frac{\partial f_c(\boldsymbol{x})}{\partial A_{ij}^k} \tag{6}$$

将每个通道的重要性权重 a_k^c 与特征图 A^k 相乘，并将所有通道的结果相加，通过 ReLU 激活，得到 Grad-CAM 的输出 $L_{\text{Grad-CAM}}=\mathrm{ReLU}(\sum_k a_k^cA^k)$，然后对 $L_{\text{Grad-CAM}}$ 进行上采样，以匹配输入图像的分辨率：

$$\Phi_{\text{Grad-CAM}}(f_c,\boldsymbol{x})=UPSAMPLE(\mathrm{ReLU}(\sum_k a_k^cA^k)) \tag{7}$$

事实上，Grad-CAM 方法会突出显示图像中对 $f_c(\boldsymbol{x})$ 有正贡献的区域，但是没有突出细节上的表示。Selvaraju 等[34]结合 Guided Backpropagation 的细粒度优势和 Grad-CAM 的定位优势，提出了引导梯度类激活映射(Guided Grad-CAM)。

$$\Phi_{\text{GuidedGrad-CAM}}(f_c,\boldsymbol{x})=\Phi_{\text{Grad-CAM}}(f_c\boldsymbol{x})\odot\Phi_{\text{GuidedBackprop}}(f_c,\boldsymbol{x}) \tag{8}$$

Chattopadhyay 等[35]提出的 Grad-CAM＋＋方法是对 Grad-CAM 的扩展，该方法提供了更好的视觉可解释性，并且能够检测多个目标对象，弥补了 Grad-CAM 方法的缺陷，即在处理出现多个相同类别的图像时，可能会导致目标对象定位不准确的问题。

Bach 等[36]提出的分层相关传播(Layerwise Relevance Propagation，LRP)是一种基于非线性分类器像素分解的方法，该方法计算层之间的像素相关性分数，并根据结果生成热力图。从最后一层开始，以反向传播的方式根据相关性得分逐层确定每个神经元的贡献大小。第 l 层第 j 个神经元的相关性得分 $R_j^{(l)}$ 是由第 $(l+1)$ 层第 k 个神经元的相关性 $R_k^{(l+1)}$ 计算得到的。

$$R_j^{(l)}=\sum_k\left(\alpha\frac{a_j^{(l)}\omega_{jk}^{+}}{\sum_j a_j^{(l)}\omega_{jk}^{+}}-\beta\frac{a_j^{(l)}\omega_{jk}^{-}}{\sum_j a_j^{(l)}\omega_{jk}^{-}}\right)\cdot R_k^{(l+1)} \tag{9}$$

式(9)是 LRP 方法最常用的 $\alpha\beta$ 规则，只要满足条件 $\alpha-\beta=1$ 并且 $\beta\geqslant 0$，就可以通过赋予 α 更多或更少的权重来产生积极或消极的影响。其中，$a_j^{(l)}$ 表示第 l 层神经元 j 的激活，ω_{jk}^{+} 和 ω_{jk}^{-} 表示神经元 j 和 k 之间的最大、最小权重。Kindermans 等[37]提出的 PatterNet 和 PatternAttribution 技术也是一种逐层反向传播方法，但是与 LRP 方法不同的是，这两种方法需要计算数据中重要信息的方向 $\boldsymbol{\alpha}^{+}$(见式(10))：PatterNet 方法用 $\boldsymbol{\alpha}^{+}$ 替换权重 $\boldsymbol{\omega}$；PatternAttribution 方法用 $\boldsymbol{\omega}\odot\boldsymbol{\alpha}^{+}$ 替换权重 $\boldsymbol{\omega}$。

$$\boldsymbol{\alpha}^{+}=\frac{E[\boldsymbol{x}\cdot\boldsymbol{y}]-E[\boldsymbol{x}]\cdot E[\boldsymbol{y}]}{\boldsymbol{\omega}^{\mathrm{T}}\cdot E[\boldsymbol{x}\cdot\boldsymbol{y}]-\boldsymbol{\omega}^{\mathrm{T}}\cdot E[\boldsymbol{x}]\cdot E[\boldsymbol{y}]} \tag{10}$$

其中，x 表示层输入，y 表示层输出，ω^{T} 表示当前层的权重，$E[x]$表示对 x 求期望。

3.2.2 基于扰动的方法

近年来，基于扰动的方法被广泛应用于解释深层图像模型[38]。该方法的基本思想是通过修改模型的输入来监测模型输出结果的变化。对模型输出的变化表明了输入的哪些部分对模型的决策结果是重要的。比如，在图像分类的情况下，对图像的扰动较大，可能会使分类器的输出结果发生变化，通过将原始图像的输出和扰动元素的输出进行比较，可以估计扰动元素的重要性。

Zeiler 等[28]提出的遮挡(Occlusion)方法系统性地用给定基线 x'(全零补丁)替换不同的连续矩形补丁，并计算相应的输出 $f_c(x)$的变化。

$$\Phi_{\text{Occlusion}}(f_c, x) = f_c(x) - f_c(x') \tag{11}$$

根据这些得分可以生成一个特征归因图，突出显示遮挡对预测结果产生的影响。这种方法不需要访问模型的内部，因此可用于解释任何模型。但是，遮挡分析方法需要额外的基线输入。

Ribeiro 等[39]提出的局部可解释模型(Local Interpretable Model Explanations，LIME)是一种与模型内部结构无关的方法，可以使用简单的、更易解释的模型 g(如线性回归，$g(z)=w\cdot z$)局部逼近复杂模型。从数据角度看，LIME 的目的是观察输入数据变化时，对模型输出结果的影响。因此，LIME 本质上是一种基于扰动的方法。

Wang 等[40]提出的分类加权类激活(Score-Weighted Class Activation，Score-CAM)方法是一种基于 CAM 的梯度无关的解释方法。该方法首先提取特征图，然后将每个激活作为原始图像上的遮掩，并获得其在目标类上的前向传递分数。最后，可以通过重要性得分 α_k^c 与特征图 A^k 的线性组合生成视觉解释结果 $\Phi^c_{\text{Score-CAM}}$。

$$\Phi^c_{\text{Score-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{12}$$

其中，重要性得分 $\alpha_k^c = C(A_l^k)$，$C(\cdot)$表示特征图 A^k 的 CIC(Channel-wise Increase of Confidence)得分。Petsiuk 等[41]提出的随机输入抽样解释(Randomized Input Sampling for Explanation，RISE)方法通过随机屏蔽输入计算特征的重要性来测试模型的输出。它会生成一个显著性图，表明每个像素对模型预测的重要程度。

Fong 等[42]提出的极值扰动(Extremal Perturbations)方法会对神经网络中特定神经元的激活产生非常大的影响。当“删除”输入图像 x 的不同区域 R 时，观察输出 $f(x)$的值如何变化。并且使用 3 种方法访问图像的生成过程：用常量替换区域 R、注入噪声、模糊图像。定义 $m:\Lambda\to[0,1]$是一个掩码，将每个像素 $u\in\Lambda$ 和标量值 $m(u)$联系起来。扰动算子定义为：

$$[\Phi(x_0, m)](u) = \begin{cases} m(u)x_0(u) + (1-m(u))\mu_0, & \text{常量} \\ m(u)x_0(u) + (1-m(u))\eta(u), & \text{噪声} \\ \int g_{\sigma_0 m(u)}(v-u)x_0(v)\mathrm{d}v, & \text{模糊} \end{cases} \tag{13}$$

其中，μ_0 表示平均颜色；$\eta(u)$表示注入的噪声，一般为高斯噪声；σ_0 是高斯模糊核 g_σ 的最大各向同性标准偏差。由于式(13)中提供了各种干扰，为了得到图像区域产生的影响，Fong 等[42]提出了“删除”和“保留”规则。它们的主要区别是：“删除”是为了找到信息量最大的区域，“保留”则是为了找到图像中的最小子集。与其他的扰动方法相比，极值扰动方法会导致 DNN 预测结果的变化最大。

3.2.3 基于 Shapley 值的方法

Shapley 值[43]最初是在博弈论背景下提出的一个框架，用于确定一组合作参与者中 P 的个人贡献。该方法针对每个合作参与者的子集 S(S⊆P)，将参与者 i 增加或删除到 S，监测对获得的总回报 v(S)的影响。具体来说，Shapley 值将参与者 i 对整个联盟 P 的贡献定义为：

$$\Phi_i = \sum_{S\subseteq P\setminus\{i\}} \alpha_S \cdot (v(S\cup\{i\}) - v(S)) \tag{14}$$

其中，每个子集 S 由 $\alpha_S = |S|!\cdot(|P|-1-|S|)!/|P|!$ 加权。将该方法应用到解释深度学习模型任务[44-45]时，合作博弈的参与者成为了输入特征，并且回报函数与 DNN 模型的输出相关。所以，式(14)中 $|P|$ 是特征数量，$|S|$ 是集合中特征的个数，$v(S\cup\{i\})$是特征组合 S 的模型输出值，$v(S)$是在子集 S 删除特征 i 的情况下模型的输出值。

Lundberg 等[44]将 LIME 方法和 DeepLift 方法进行改进，提出了基于计算 Shapely 值的 LIMEShap 和 DeepLiftShap 方法。该论文中的 LinearShap 方法假设输入特征相互独立，根据线性模型，近似计算梯度的 SHAP 值，将高斯噪声加入到随机选择的样本点，并计算对应的输出梯度。LIMEShap 方法和 LIME 方法相似，不同之处在于回归代理模型的权重并非通过余弦距离度量，而是式(15)。

$$\pi_{x'}(z') = \frac{M-1}{\left(\frac{M}{|z|}\right)\times(M-|z'|)} \tag{15}$$

其中，M 表示特征的数量，$|z'|$ 是 z' 中非零元素的数量。DeepLiftShap 扩展了 DeepLift 方法，并根据模型的等效性近似计算 SHAP 值。该方法假设输入特征相互独立，使用线性组合规则将神经网络的非线性组件线性化，对于模型 f，每个组件的 SHAP 的有效线性化计算如式(16)所示：

$$\Phi_i(f, \boldsymbol{y}) \approx m_{y_i, f}(y_i - E[y_i]) \tag{16}$$

3.2.4 其他方法

基于梯度的方法无法解决由激活函数引起的梯度饱和或梯度为零的问题。为此，Shrikumar 等[27]提出了计算特征贡献的解释方法 DeepLift，其主要原理是设置一个“参考”激活值，将每个神经元的激活值和“参考”激活值进行比较，并根据差异为每个输入分配特征贡献得分。“参考激活”是通过一些用户定义的参考输入获得的，如式(17)所示：

$$\sum_{i=1}^{n} C_{(x_i)} = A_{x_i} - A'_{x_i} \tag{17}$$

其中，$C_{(x_i)}$ 表示特征贡献得分，A_{x_i} 表示神经元 i 的激活值，A'_{x_i} 表示神经元 i 的“参考”激活值。DeepLift 方法解决了基于梯度方法的局限性，即使梯度为零，它们之间的参考差异也不会为零。

3.3 基于非归因的方法

近年来，基于归因的方法较为流行，但是将这些算法应用于公共卫生、金融银行等领域仍存在一些问题。例如，在基于

相关分数的反向传播过程中，如果一些神经元接收到错误的相关信息，则下层的其他神经元将累积错误，最终导致解释结果不精确；其次，在基于扰动的方法中，不可能对输入的所有扰动样本进行采样。因此许多研究试图通过不同的方法来解决可解释性问题。表1从方法描述、待解释模型和应用3个方面列出了非归因的方法汇总，并对目前比较流行的3个非归因解释方法（概念激活向量、基于实例的解释和基于注意力）的解释做了详细阐述。

表1　非归因方法汇总

Table 1　Summary of non-attribution methods

解释方法	描述	待解释模型	应用	参考文献
概念激活向量	基于人类专家定义的概念，通过计算模型对不同概念的激活程度解释模型的分类决策	Inception ResNet	DR检测 图像分类 口语评估	[46] [59-60]
基于实例	使用特定的输入实例解释复杂的深度学习模型，通常提供局部解释	CNN	图像分类 胸部X射线诊断 智能农业决策	[50-51] [61]
基于注意力	利用注意图解释模型的深层特征，反映图像的哪些区域引起了神经网络的注意	CNN RNN LSTM	神经机器翻译 医疗诊断 自动驾驶	[55]
基于专家知识	利用规则或符号的人工智能系统，将深度学习模型和专家知识融合	U-net VGG16 CNN	图像分类 脑部MLS评估 肺部CT	[62]
基于文本解释	一种使用神经网络生成图像的自然语言描述，实现由文本到图像的转换	CNN &LSTM CNN &BRNN	乳腺肿块分类 多模态情感分析	[63-64]

注：LSTM：长短期记忆网路；CRF：条件随机场；MLS：大脑中线移位；ResNet：残差网络；Inception：Inception网络；U-Net：基于CNN的图像分割网络；VGG16：基于ImageNet图像库预测训练的网络；BRNN：双向递归神经网络

3.3.1　概念激活向量

Kim等[46]提出的使用概念激活向量测试（Testing with Concept Activation Vectors，TCAV）的方法，用于为领域专家解释模型在不同网络层学习到的特征。TCAV方法在概念空间中采用方向导数，根据TCAV分数确定特定概念在分类中的重要性。具体而言，TCAV为DNN模型定义了一些人类易于理解的高级概念，将概念激活向量定义为超平面的法线，把模型激活中没有概念的示例和有概念的示例分隔开。设k为监督学习任务的类标签，则k对概念C的“概念敏感性”被定义为方向导数$S_{C,k,l}(\boldsymbol{x})$：

$$S_{C,k,l}(\boldsymbol{x})=\nabla h_{l,k}(f_l(\boldsymbol{x}))\cdot v_C^l \tag{18}$$

其中，$v_C^l\in\mathbb{R}^m$是l层中概念C的单位，$f_l(\boldsymbol{x})$表示l层中输入$\boldsymbol{x}$的激活值，$h_{l,k}:\mathbb{R}^m\to\mathbb{R}$。$S_{C,k,l}(\boldsymbol{x})$表示在网络的激活层中定量测量模型预测对概念方向输入变化的敏感性。TCAV分数可以用来确定k类输入对概念C的积极影响。

$$TCAV_{C,k,l}=\frac{|\{\boldsymbol{x}\in\boldsymbol{X}_k:S_{C,k,l}(\boldsymbol{x})>0\}|}{|\boldsymbol{X}_k|} \tag{19}$$

其中，$\boldsymbol{X}_k$表示k标签下的所有输入，并且$TCAV_{C,k,l}\in[0,1]$。

3.3.2　基于实例的方法

基于实例推理（Case-Based Reasoning，CBR）是一种类比推理形式，它使用先前的已知案例及其解决方案的案例库来确定新的查询案例的解决方案，从而提供一个与待解释的查询案例最相似的案例。原型选择[47]本质上是一种基于案例的推理，目的是找到能够代表整个数据集实例的最小子集。和原型选择不同的是，CBR从输入和原型中提取特征之间的相似性度量得分值[48-49]，以揭示神经网络的隐藏信息。Chen等[50]构建了一个ProtoPNet模型，该模型由一个卷积层、一个原型层和一个全连接层组成。其中，卷积层由充当特征提取器的标准CNN通道组成，原型层将卷积层的补丁作为输入，在训练过程中学习输入的原型部分（如鸟的头部或身体），然后与每个原型进行比较，计算相关性分数，全连接层根据相关性分数进行预测。但是，由于ProtoPNet模型使用输入图像固定大小的特征映射与原型比较，补丁可能无法捕获类别区分特征。Kim等[51]在ProtoPNet的基础上提出了XProtoNet模型，该模型在动态区域内提取特征，并且与原型比较时不受大小的限制。此外，Wachter等[52]介绍了一种反事实的解释，能够根据输入的最小变化，产生不同的预测结果。例如，在医疗领域，这些解释可以指导医生调整治疗方案，以便患者能够康复出院。因此，反事实解释必须尽量减少当前输入x和反事实示例x'之间的差异。此优化问题可以表示为：

$$\arg\min_{x'}\lambda(f(\boldsymbol{x}')-y')^2+d(\boldsymbol{x},\boldsymbol{x}') \tag{20}$$

其中，λ是一个常数，y'是不同的标签，$d(\cdot,\cdot)$为曼哈顿距离。此外，文献[53-54]对于应该使用何种指标来最小化差异进行了讨论。

3.3.3　基于注意力的解释

在深度学习中，注意力的基本概念来源于人们关注、分析图像或其他数据的不同部分。在深度学习模型中嵌入注意力机制，构建并重新训练网络生成注意图，可以提高模型的可解释性。这里主要介绍如何将注意力图作为提高深度学习可解释性的工具[55-56]。在NLP领域，以神经机器翻译[57]为例，假设输入文件的隐藏层表示为$\boldsymbol{s}=[s_1,s_2,\cdots,s_n]$，每一个$\boldsymbol{s}_i$是一个$d$维向量，$n$为输入长度，解码器在每个时间步长生成预测词语得分$p$。

$$\begin{aligned}&p(y_j\mid y_{<j},\boldsymbol{s})=\mathrm{softmax}(\boldsymbol{W}\tilde{\boldsymbol{h}}_j)\\&\tilde{\boldsymbol{h}}_j=Attention(\boldsymbol{h}_j,\boldsymbol{s}),\boldsymbol{h}_j=f(\boldsymbol{h}_{j-1})\end{aligned} \tag{21}$$

其中，$\boldsymbol{W}$是输出词汇量大小的变换矩阵；f是任何循环结构，可以利用上一时间步的状态计算当前隐藏层状态；$\boldsymbol{h}_j$为循环结构的隐藏层单元；$Attention$为注意力组件，将当前隐藏层

状态和源隐藏层状态作为输入，并输出一个基于注意力的隐藏层状态，最后将其输入到 softmax 层进行模型预测。基于注意力的核心思想是导出一个获取加权源隐藏状态的上下文向量 c_j，并且注意力权重 a_{ji} 决定当前时间步应关注多少源输入词汇。其中，注意力权重向量 a_{ji} 的计算式如式(22)所示，上下文向量 c_j 为注意力权重的加权和：$c_j=\sum_{i=1}^{n}a_{ji}s_i$。

$$a_{ji}=align(h_j,s_i)=\frac{\exp(score(h_j,s_i))}{\sum_{i'}\exp(score(h_j,s_i))} \tag{22}$$

其中，$score$ 为基于内容的函数，对位置 j 周围的输入和位置 i 周围的输出计算匹配分数，该分数的计算是基于输入文件的隐藏层状态 s_i 和循环结构的隐藏层状态 h_j。

对于神经机器翻译，在每个算法步骤中，模型需要关注源句子最相关的部分，并将其翻译成目标语言。Ghader 等[57]研究如何在神经机器翻译系统中操控注意力，构建了一个可以自定义注意力权重的交互式可视化工具，并根据注意力权重可视化输出概率，从而帮助用户了解注意力如何影响模型预测。在医疗诊断中，Zhang 等[2]提出了一个包含图像模型和语言模型的 MDNet 网络，用于在医学图像和诊断之间进行多模式映射。其中，图像模型用于增强多尺度特征集合和提高模型的利用效率，并将语言模型和注意力机制相结合，从而发现有区别的特征，学习图像和诊断报告之间的映射。在自动驾驶领域，Mori 等[58]将注意力分支网络(Attention Branch Network，ABN)引入到自驱动模型，该模型能够使用注意力图直观地分析自驱动决策的原因。

4 可解释性评估方法

针对深度学习模型，评估指标能够全面衡量模型是否满足可解释性。与分类的评估指标(准确度、精确度和召回率)一样，模型可解释性的评估指标应能从特定角度证明模型的性能。但是，由于 DNN 生成解释的性质不同或输入数据的类型不同，目前还没有一个公认的指标用于评估可解释性。然而，专家可以定性评估生成解释的相关性；并且存在一些定量评估方法，可以客观地评估各个领域产生的解释。本节主要从定性评估和定量评估两个方面介绍深度学习模型可解释性评估方法。

4.1 定性评估

Oviedo 等[65]提出了平均类激活映射(average Class Activation Maps，average CAM)方法，该方法提供的全局解释可以由专家通过分析显著性图的形态和细粒度定性分析评估，而且特定用户可能从专家反馈中受益。然而，由于深度学习模型具有高度非线性，领域专家也很难定性评估 XAI 方法生成解释的质量。因此，我们应优先考虑定量评估方法。

4.2 定量评估

定量评估是量化解释的数字指标，为比较不同的解释提供了一种直观的方法。本文主要介绍正确性(Correctness)、连贯性(Coherence)和稳定性(Stability)这几种定量评估指标。

(1)正确性(Correctness)：正确性表示解释在多大程度上忠实于预测模型。为了评估解释对预测模型的可靠性和敏感性，Adebayo 等[66]引入了模型参数随机化检查方法，该方法从上到下破坏已学习的权重，并将可解释性方法应用到每个随机状态。如果随机化后的解释与原始解释相同，则该解释对模型不敏感；如果两种解释不同，则不能保证原始解释完全正确。Yeh 等[67]提出了最大灵敏度(Max-Sensitivity，MS)方法(见式(23))，根据解释 $\Phi(f_c,x)$ 在输入 x' 的微小扰动下的最大变化来衡量可靠性。

$$SENS_{\mathrm{MAX}}(\Phi,f,x,r)=\max_{\|\delta\|\leqslant r}\|\Phi(f,x+\delta)-\Phi(f,x)\| \tag{23}$$

其中，$\delta=x'-x$，f 表示黑箱函数，r 表示输入邻域半径，$\|\cdot\|$ 表示对函数取范数。

此外，Ancona 等[68]提出根据 sensitivity-n 标准来评估 XAI 方法得到的特征重要性得分或热力图的正确性，该标准量化了当删除不同相关性的特征时所导致的输出变化，结果表明，删除最相关的像素时，输出变化更快。

(2)连贯性(Coherence)：连贯性是为了比较 XAI 方法生成的解释是否与领域知识或共识一致。对于图像解释，通常将热力图或解释的位置与真实物体边界框、分割掩码或人类注意力图进行比较来评估“位置一致性”，并使用内外相关比[69]、点定位误差[70]或定点游戏的准确度[71]等量化“基本事实”与解释之间的对应关系。例如，Zhang 等[71]使用定点游戏的准确度评估自上而下的注意力图在视觉场景中定位目标对象的能力。首先，在注意力图上提取最大点，如果最大点位于提示对象类别的注释边界框内，则发生命中。其中，对象类别的定位准确度由 $Acc=\frac{1}{N}\frac{\#Hits}{\#Hits+\#Misses}$ 计算得到，N 为数据集中相关类别的数量。对于文本解释，通常使用标准的自然语言进行评估。例如，Lin[72]提出了 Rouge 度量标准，将待评估摘要与人类创建的标准摘要进行比较，自动确认摘要的质量。ROUGE-N 的计算式如式(24)所示，它统计了计算机生成的待评估摘要和标准摘要的重叠单元数量(n-gram 长度、单词序列或单词对)。

$$ROUGE\text{-}N=\frac{\sum_{S\in\{Ref\}}\sum_{gram_n\in S}Count_{\mathrm{match}}(gram_n)}{\sum_{S\in\{Ref\}}\sum_{gram_n\in S}Count(gram_n)} \tag{24}$$

其中，$gram_n$ 表示 n 个连续片段(n-gram)的长度，$Count_{\mathrm{match}}(gram_n)$ 是待评估摘要和参考摘要中共同出现的 n-gram 的最大数量。

(3)稳定性(Stability)：稳定性用于评估原始输入样本和引入噪声的样本分别得到的解释之间的相似性，也就是说，输入加入微小的白噪声，解释也会引入可见的变化。Alvarez-Melis 等[73]通过归一化距离的方法衡量特定输入 x 及邻域 ε 的自解释模型 f_{expl} 的稳定性，并使用激活最大化方法优化参数。

$$\hat{L}(x_i)=\arg\max_{x_j\in B_\varepsilon(x_i)}\frac{\|f_{\mathrm{expl}}(x_i)-f_{\mathrm{expl}}(x_j)\|_2}{\|h(x_i)-h(x_j)\|_2} \tag{25}$$

其中，$h(\cdot)$为输入 x 的可解释表示。Chu 等[74]通过分析相似输入实例产生的解释是否具有相似性，系统地研究了 XAI

方法在 Fashion-MNIST 数据集上的解释具有高稳定性。假设输入样本 x 的最近邻（欧几里得距离）为 x'，Y 和 Y' 分别为 x 和 x' 的决策特征，Chu 等[74]通过 Y 和 Y' 之间的余弦相似性衡量解释的一致性，较大的余弦相似度表示 LIME 得到了更好的解释一致性。Lakkaraju 等[75]通过比较原始输入和轻微扰动输入的预测评估解释的保真度，并提出解释应该对原始输入数据和稍微偏移的输入都具有高保真度，以确保解释的稳定性。

5　深度学习可解释性应用

5.1　医疗领域的应用

医疗领域越来越依赖深度学习来支持决策，但是一般的深度学习决策系统类似于一个“黑箱”，不能给出输出的决策结果模型内部的诊断过程。然而对于医生而言，只有信任一个模型才能相信它提供的决策结果，因此医疗领域深度学习决策系统要求模型具有可解释性。

Afshar 等[76]研究了胶囊网络的可解释性，提出用已训练的胶囊网络来识别可解释的放射特征。研究表明，胶囊网络从医学图像中提取的特征不仅可以区分肿瘤类型，还与手工制作的特征有很大相关性。而且从医生的角度来说，手工制作的特征更容易被接受和理解。Wu 等[77]提出了 DeepMiner 框架，该框架试图可发现 DNN 中可解释的表征，并为乳腺癌的预测建立解释。这种方法不需要对完整的医疗报告进行训练，而是利用 CNN 模型的内部表示来解释决策。Wang 等[78]提出用 Grad-CAM 方法解释 RAPNN 框架，该框架可以从胸部 CT 图像中学习单个图像级的特征，将模型中 RAP 模块的输出作用于 Grad-CAM[35]，实现医疗领域的决策结果透明化。近期，El-Sappagh 等[79]开发了一个准确且可解释的阿尔茨海默病（Alzheimer's Disease，AD）诊断和进展检测模型。首先，该模型使用 Shapley 附加解释（SHAP）特征归因框架，为模型的每一层提供基于随机森林（Random Forest，RF）分类器的全局解释和实例解释；其次，基于决策树分类器和模糊规则系统得到的 22 个解释器，为每一层的 RF 决策提供一致和可靠的解释，并且为了帮助医生理解预测结果，这些解释以自然语言的形式呈现。

5.2　金融领域的应用

深度学习模型被广泛应用于金融领域。银行系统中的产品推荐、风险评估和异常检测等都由深度学习算法完成，虽然技术上的进步为银行客户带来了巨大的便利，但深度学习模型缺乏透明度和可解释性也会导致用户缺乏对模型决策结果的信任。因此，深度学习模型可解释性和算法透明性格外重要。国内外的金融机构及金融科技企业对 XAI 已展开研究与探索，人们对模型的应用要求不仅停留在准确性层面，模型结果的可解释性、安全性、公平性等同样重要。例如，在金融视觉研究中，交易员可以通过查看烛台模式发现资产的趋势，但由于深度学习模型将其推理隐藏在一个“黑箱”中，交易员无法确定模型学到了什么。因此，Chen 等[80]提出了一个 GASF-CN 模型，用于解释确定时间序列的特定烛台模式的深度学习模型。Chen 等[80]还提出了一种基于局部搜索对抗攻击的方法，解释了 GASF-CN 模型以类似于人类交易者的方式感知烛台的模式。Han 等[81]提出了一种钞票识别和假币检测系统，并基于传统的梯度 CAM，提出像素级梯度 CAM（pixel-wise Grad-CAM，pGrad-CAM）方法，该方法能够以可视化的方式解释模型产生的决策结果。

5.3　自动驾驶领域的应用

目前，深度学习算法推动了自动驾驶在感知、目标检测等方向的飞速发展，许多欧洲国家、美国和加拿大部署到道路网络的自动驾驶汽车数量大幅增加[82]，百度第六代自动驾驶汽车 Apollo RT6 已在全国十多个城市提供自动驾驶出租叫车服务。但是，回顾过去，Uber 自动驾驶汽车在公共道路致路人死亡、特斯拉 Model S 在高速公路追尾消防车等事故，考验着人们的容忍度，且降低了人们对深度学习模型的信任程度。因此，专家致力于构建自动驾驶系统，对车辆的决定性行为做出可理解的解释。一般而言，解释自动驾驶车辆的行为有两个方面：视觉解释和文本解释。视觉解释通过热力图的方式可视化图像的哪些部分影响自动驾驶汽车执行特定的行为，而文本解释旨在使用自然和可理解的语言为人类提供车辆采取决策的理由。

Bojarski 等[83]基于视觉解释方法为自动驾驶决策提供了一种可视化方法（VisualBackProp），该方法显示了哪一组输入像素对 CNN 的预测贡献最大。Bojarski 等[83]在 Udacity 自动驾驶汽车数据集上进行端到端的自动驾驶任务。实验表明，VisualBackProp 方法是调试 CNN 预测的有效工具。Zeng 等[84]提出了一种通过遵守交通规则学习安全驾驶车辆的架构。该架构使用原始激光雷达数据和高清地图作为输入，以 3D（3 Dimension）检测的形式生成可解释的中间表示，并在计划范围内预测未来的运动轨迹。其中，3D 检测确保中间表示的可解释性；以 L1（曼哈顿）和 L2（欧氏）距离度量预测轨迹信息，可以解释车辆运动轨迹偏移的原因是动作错误还是方向错误；通过成本图（Cost Map）自顶向下可视化不同的交通场景。

对于文本解释方面，Kim 等[85]提出了一种基于语义分割的自然语言训练车辆控制系统的方法。该方法通过学习自然语言描述的视觉环境，做出预测控制（红色信号灯亮则停止运动）。为了提高模型的可解释性，Kim 等[85]将语义分割与注意力机制相结合，引入了一种细粒度关注机制，丰富了知识表示。利用 Berkeley Deep Drive eXplanation（BDD-X）数据集进行的实验表明基于语义分割的自然语言训练车辆控制系统提高了车辆预测行为的安全性和可解释性。Omeiza 等[86]提出了一种基于决策树模型解释自动驾驶行为的方法，该方法根据交通规则将视觉检测映射到自动驾驶车辆的行为上，并基于不同的场景生成解释。采用定量和定性的评估方法，依据可理解性和责任性目标评估自动驾驶系统生成的解释。

5.4　司法领域的应用

在刑事司法领域，当使用深度学习模型预测司法判决时，必须确保模型以公平、安全和无歧视的形式为人类呈现判决

结果。例如,在艾瑞克·卢米斯(Eric Loomis)和威斯康星州案件中[87],法官依据专有的、封闭源代码的风险评估模型COMPAS[88],预测出卢米斯的刑事再犯风险高,并判处卢米斯6年有期徒刑和5年延期监督。然而,判决所依据的COMPAS模型被视为商业机密,法官并不清楚其因果判决过程,判决结果引起了学术界和社会舆论的质疑。因此,在司法领域,必须保证模型决策的透明度,但目前仅有少量的研究工作致力于使司法系统的模型决策具备可解释性[89]。

6 深度学习可解释性未来发展方向

针对现有的深度学习可解释性方法,我们主要从可解释的对象、模型性能和复杂性的关系,以及如何提高模型可信度和可靠性这3个方面给出深度学习可解释性的未来发展方向。

(1)现有的可解释性方法主要针对图像和文本分类任务,如何将可解释性方法扩展到其他任务仍是未来深度学习可解释性研究领域的主要方向之一。

(2)深度学习可解释性的目的是在解释DNN的同时,尽可能保持模型性能和复杂度的平衡[90]。现有的一些方法虽然提高了模型的可解释性,但损害了模型性能[91]。因此研究可解释性和模型性能之间的平衡对深度学习可解释性至关重要。

(3)研究可解释性最重要的目的是在一些特定领域能使用解释方法提高模型的可信度和可靠性。但是,当一个可靠的深度学习模型提供了一个错误的视觉或文本解释时,探究错误解释的方法似乎并不存在。未来应该结合应用数学、数据科学等专业知识,为人类提供模型的决策推理过程,以帮助用户判断他们对模型给出结果的信任程度。

结束语 近年来,随着深度学习可解释性的发展,提高深度学习模型的可解释性成为人们日益关注的焦点。本文主要从内在可解释模型、基于归因和非归因解释3个方面对深度学习可解释技术进行了综述。首先,介绍了可解释性的定义及解释原因;其次,介绍了深度学习可解释性方法并从内在可解释模型、基于归因的解释和非归因解释3个角度对这些方法进行概述;然后,从定量和定性两个方面介绍了可解释性的评估指标;最后,阐述了深度学习可解释性在医疗和推荐系统领域的应用,并对深度学习可解释性未来的研究方向进行了展望。

参 考 文 献

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.

[2] ZHANG Z, XIE Y, XING F, et al. Mdnet: A semantically and visually interpretable medical image diagnosis network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2017: 6428-6436.

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

[4] MONGA V, LI Y, ELDAR Y C. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing [J]. IEEE Signal Processing Magazine, 2021, 38(2): 18-44.

[5] SAHBA A, DAS A, RAD P, et al. Image graph production by dense captioning [C] // 2018 World Automation Congress (WAC). New Jersey: IEEE, 2018: 1-5.

[6] ALI M, YOUSUF N, RAHMAN M, et al. Machine translation using deep learning for universal networking language based on their structure [J]. International Journal of Machine Learning and Cybernetics, 2021, 12(8): 2365-2376.

[7] YU K. Deep Learning for Unsupervised Neural Machine Translation[C]// 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). New Jersey: IEEE, 2021: 614-617.

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

[9] ARRIETA A B, DÍAZ-RODRÍGUEZ N, DEL SER J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI [J]. Information Fusion, 2020, 58: 82-115.

[10] MURDOCH W J, SINGH C, KUMBIER K, et al. Interpretable machine learning: definitions, methods, and applications [J]. arXiv: 1901.04592, 2019.

[11] SAMEK W, MONTAVON G, LAPUSCHKIN S, et al. Explaining deep neural networks and beyond: A review of methods and applications [J]. Proceedings of the IEEE, 2021, 109(3): 247-278.

[12] SU J M, LIU H F, XIANG F T, et al. Survey of interpretation methods for deep neural networks[J]. Computer Engineering, 2020, 46(9): 1-15.

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

[14] LEI X, LUO X L. Review on interpretability of deep learning [J]. Journal of Computer Applications, 2022, 42(11): 3588-3602.

[15] LI L M, HOU M R, CHEN K, et al. Survey on interpretability of deep learning[J]. Journal of Computer Applications, 2022, 42(12): 3639-3650.

[16] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning [J]. arXiv: 1702.08608, 2017.

[17] LIPTON Z C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery [J]. Queue, 2018, 16(3): 31-57.

[18] DINGEN D, VAN'T VEER M, HOUTHUIZEN P, et al. Regression Explorer: Interactive exploration of logistic regression models with subgroup analysis [J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 25(1): 246-255.

[19] ZILKE J R, LOZA MENCÍA E, JANSSEN F. Deepred-rule extraction from deep neural networks[C]// International Conference on Discovery Science. Berlin: Springer, 2016: 457-473.

[20] THRUN S. Extracting rules from artificial neural networks with distributed representations[M]// Cambridge: MIT Press, 1994: 505-512.

[21] NGUYEN D T, KASMARIK K E, ABBASS H A. Towards Interpretable Neural Networks: An Exact Transformation to Multi-Class Multivariate Decision Trees [J]. arXiv: 2003.04675, 2020.

[22] BENÍTEZ J M, CASTRO J L, REQUENA I. Are artificial neural networks black boxes? [J]. IEEE Transactions on neural networks, 1997, 8(5): 1156-1164.

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

[24] KENENI B M, KAUR D, AL BATAINEH A, et al. Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles [J]. IEEE Access, 2019, 7: 17001-17016.

[25] ZHENG S, DING C. A group lasso based sparse KNN classifier [J]. Pattern Recognition Letters, 2020, 131: 227-233.

[26] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: Visualising image classification models and saliency maps [J]. arXiv: 1312.6034, 2013.

[27] SHRIKUMAR A, GREENSIDE P, SHCHERBINA A, et al. Not just a black box: Learning important features through propagating activation differences [J]. arXiv: 1605.01713, 2016.

[28] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]// European Conference on Computer Vision. Berlin: Springer, 2014: 818-833.

[29] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[30] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for simplicity: The all convolutional net [J]. arXiv: 1412.6806, 2014.

[31] SUNDARARAJAN M, TALY A, YAN Q. Axiomatic attribution for deep networks[C]// International Conference on Machine Learning. New York: PMLR, 2017: 3319-3328.

[32] SMILKOV D, THORAT N, KIM B, et al. Smoothgrad: removing noise by adding noise [J]. arXiv: 1706.03825, 2017.

[33] 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. New Jersey: IEEE, 2016: 2921-2929.

[34] 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. Washington: IEEE Computer Society, 2017: 618-626.

[35] 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). New Jersey: IEEE, 2018: 839-847.

[36] 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): 1-46.

[37] KINDERMANS P-J, SCHÜTT K T, ALBER M, et al. Learning how to explain neural networks: Pattern net and pattern attribution [J]. arXiv: 1705.05598, 2017.

[38] DABKOWSKI P, GAL Y. Real time image saliency for black box classifiers [C]// Proceeding of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc, 2017: 6970-6979.

[39] 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.

[40] 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. New Jersey: IEEE, 2020: 24-25.

[41] PETSIUK V, DAS A, SAENKO K. Rise: Randomized input sampling for explanation of black-box models [J]. arXiv: 1806.07421, 2018.

[42] FONG R C, VEDALDI A. Interpretable explanations of black boxes by meaningful perturbation[C]// Proceedings of the IEEE International Conference on Computer Vision. New Jersey: IEEE, 2017: 3429-3437.

[43] SHAPLEY L S. A value for n-person games [J]. Classics in game theory, 1952, 2(28): 307-317.

[44] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions [C]// Proceeding of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc, 2017: 4768-4777.

[45] STRUMBELJ E, KONONENKO I. An efficient explanation of individual classifications using game theory [J]. The Journal of Machine Learning Research, 2010, 11: 1-18.

[46] KIM B, WATTENBERG M, GILMER J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav) [C]// International Conference on Machine Learning. New York: PMLR, 2018: 2668-2677.

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

[48] LI O, LIU H, CHEN C, et al. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018: 3530-3537.

[49] WARGNIER-DAUCHELLE V, GRENIER T, DURAND-DUBIEF F, et al. A more interpretable classifier for multiple sclerosis[C]// 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). New Jersey: IEEE, 2021: 1062-1066.

[50] CHEN C, LI O, TAO D, et al. This looks like that: deep learning for interpretable image recognition [C]// Proceeding of the 33rd

International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc, 2019: 8930-8941.

[51] KIM E, KIM S, SEO M, et al. XProtoNet: diagnosis in chest radiography with global and local explanations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2021: 15719-15728.

[52] WACHTER S, MITTELSTADT B, RUSSELL C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR [J]. Harvard Journal of Law & Technology, 2017, 31(2): 841.

[53] 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.

[54] SHARMA S, HENDERSON J, GHOSH J. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models [J]. arXiv: 1905.07857, 2019.

[55] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation [J]. arXiv: 1508.04025, 2015.

[56] STAHLBERG F, SAUNDERS D, BYRNE B. An operation sequence model for explainable neural machine translation[J]. arXiv: 1808.09688, 2018.

[57] GHADER H, MONZ C. What does attention in neural machine translation pay attention to?[J]. arXiv: 1710.03348, 2017.

[58] MORI K, FUKUI H, MURASE T, et al. Visual explanation by attention branch network for end-to-end learning-based self-driving[C]// 2019 IEEE Intelligent Vehicles Symposium (IV). New Jersey: IEEE, 2019: 1577-1582.

[59] WEI X, GALES M J, KNILL K M. Analysing bias in spoken language assessment using concept activation vectors[C]// 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021). New Jersey: IEEE, 2021: 7753-7757.

[60] GRAZIANI M, ANDREARCZYK V, MÜLLER H. Understanding and Interpreting Machine Learning in Medical Image Computing Applications[M]// New York: Springer, 2018: 124-132.

[61] ZHAI Z, ORTEGA J F M, MARTÍNEZ N L, et al. An efficient case retrieval algorithm for agricultural case-based reasoning systems, with consideration of case base maintenance [J]. Agriculture, 2020, 10(9): 387.

[62] ZHU P, OGINO M. Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support[M]// New York: Springer, 2019: 39-47.

[63] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3128-3137.

[64] LEE H, KIM S T, RO Y M. Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support[M]. New York: Springer, 2019: 21-29.

[65] OVIEDO F, REN Z, SUN S, et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks [J]. NPJ Computational Materials, 2019, 5(1): 1-9.

[66] ADEBAYO J, GILMER J, MUELLY M, et al. Sanity checks for saliency maps [C]// Proceeding of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc, 2018: 9525-9536.

[67] YEH C K, HSIEH C Y, SUGGALA A, et al. On the(in) fidelity and sensitivity of explanations [C]// Proceeding of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc, 2019: 10967-10978.

[68] ANCONA M, CEOLINI E, ÖZTIRELI C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks [J]. arXiv: 1711.06104, 2017.

[69] NAM W J, GUR S, CHOI J, et al. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks[C]// Procee-dings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2020: 2501-2508.

[70] JAKAB T, GUPTA A, BILEN H, et al. Self-supervised learning of interpretable keypoints from unlabelled videos[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2020: 8787-8797.

[71] ZHANG J, BARGAL S A, LIN Z, et al. Top-down neural attention by excitation backprop [J]. International Journal of Computer Vision, 2018, 126(10): 1084-1102.

[72] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]// Text summarization branches out. 2004: 74-81.

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

[74] CHU L, HU X, HU J, et al. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1244-1253.

[75] LAKKARAJU H, ARSOV N, BASTANI O. Robust and stable black box explanations[C]// International Conference on Machine Learning. New York: PMLR, 2020: 5628-5638.

[76] AFSHAR P, PLATANIOTIS K N, MOHAMMADI A. Capsule networks' interpretability for brain tumor classification via radiomics analyses[C]// 2019 IEEE International Conference on Image Processing (ICIP). New Jersey: IEEE, 2019: 3816-3820.

[77] WU J, ZHOU B, PECK D, et al. Deepminer: Discovering interpretable representations for mammogram classification and explanation [J]. arXiv: 1805.12323, 2018.

[78] WANG Y, FENG C, GUO C, et al. Solving the sparsity problem

in recommendations via cross-domain item embedding based on co-clustering[C]// Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019:717-725.

[79] EL-SAPPAGH S, ALONSO J M, ISLAM S, 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.

[80] CHEN J H, CHEN S Y C, TSAI Y C, et al. Explainable deep convolutional candlestick learner [J]. arXiv:2001.02767, 2020.

[81] HAN M, KIM J. Joint banknote recognition and counterfeit detection using explainable artificial intelligence [J]. Sensors, 2019, 19(16):3607.

[82] MÜLLER J M. Comparing Technology Acceptance for Autonomous Vehicles, Battery Electric Vehicles, and Car Sharing—A Study across Europe, China, and North America [J]. Sustainability, 2019, 11(16):4333.

[83] BOJARSKI M, CHOROMANSKA A, CHOROMANSKI K, et al. Visualbackprop: visualizing cnns for autonomous driving [J]. arXiv:1611.05418, 2016.

[84] ZENG W, LUO W, SUO S, et al. End-to-end interpretable neural motion planner[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Jersey: IEEE, 2019:8660-8669.

[85] KIM J, ROHRBACH A, AKATA Z, et al. Toward explainable and advisable model for self-driving cars [J]. Applied AI Letters, 2021, 2(4):1-13.

[86] OMEIZA D, WEB H, JIROTKA M, et al. Towards accountability: providing intelligible explanations in autonomous driving [C]// 2021 IEEE Intelligent Vehicles Symposium (IV). New Jersey: IEEE, 2021:231-237.

[87] LIGHTBOURNE J. Damned lies & criminal sentencing using evidence-based tools [J]. Duke Law & Technology Review, 2016, 15:327.

[88] TAN S, CARUANA R, HOOKER G, et al. Distill-and-compare: Auditing black-box models using transparent model distillation [C]// Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018:303-310.

[89] BERK R A, BLEICH J. Statistical procedures for forecasting criminal behavior: A comparative assessment [J]. Criminology & Pubulic Pollcy, 2013, 12(3):511.

[90] SUN Z, FAN C, HAN Q, et al. Self-explaining structures improve nlp models [J]. arXiv:2012.01786, 2020.

[91] BERTSIMAS D, DELARUE A, JAILLET P, et al. The price of interpretability [J]. arXiv:1907.03419, 2019.

CHEN Chong, born in 1987, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. His main research interests include machine learning, information fusion and machine learning interpretability.

(责任编辑:何杨)