软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(1):230−254 [doi: 10.13328/j.cnki.jos.006415]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向图像分类的深度模型可解释性研究综述*

杨朋波[1,2], 桑基韬[1,2], 张 彪[1,2], 冯耀功[1,2], 于 剑[1,2]

[1](北京交通大学 计算机与信息技术学院, 北京 100044)

[2](北京交通大学 人工智能研究院, 北京 100044)

通信作者: 于剑, E-mail: jianyu@bjtu.edu.cn

摘 要: 深度学习目前在计算机视觉、自然语言处理、语音识别等领域得到了深入发展, 与传统的机器学习算法相比, 深度模型在许多任务上具有较高的准确率. 然而, 作为端到端的具有高度非线性的复杂模型, 深度模型的可解释性没有传统机器学习算法好, 这为深度学习在现实生活中的应用带来了一定的阻碍. 深度模型的可解释性研究具有重大意义而且是非常必要的, 近年来许多学者围绕这一问题提出了不同的算法. 针对图像分类任务, 将可解释性算法分为全局可解释性和局部可解释性算法. 在解释的粒度上, 进一步将全局解释性算法分为模型级和神经元级的可解释性算法, 将局部可解释性算法划分为像素级特征、概念级特征以及图像级特征可解释性算法. 基于上述分类框架, 总结了常见的深度模型可解释性算法以及相关的评价指标, 同时讨论了可解释性研究面临的挑战和未来的研究方向. 认为深度模型的可解释性研究和理论基础研究是打开深度模型黑箱的必要途径, 同时可解释性算法存在巨大潜力可以为解决深度模型的公平性、泛化性等其他问题提供帮助.

关键词: 深度学习; 可解释性; 图像分类; 特征

中图法分类号: TP18

中文引用格式: 杨朋波, 桑基韬, 张彪, 冯耀功, 于剑. 面向图像分类的深度模型可解释性研究综述. 软件学报, 2023, 34(1): 230–254. http://www.jos.org.cn/1000-9825/6415.htm

英文引用格式: Yang PB, Sang JT, Zhang B, Feng YG, Yu J. Survey on Interpretability of Deep Models for Image Classification. Ruan Jian Xue Bao/Journal of Software, 2023, 34(1): 230–254 (in Chinese). http://www.jos.org.cn/1000-9825/6415.htm

Survey on Interpretability of Deep Models for Image Classification

YANG Peng-Bo[1,2], SANG Ji-Tao[1,2], ZHANG Biao[1,2], FENG Yao-Gong[1,2], YU Jian[1,2]

[1](School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

[2](Institute of Artificial Intelligence, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Deep learning has made great achievements in various fields such as computer vision, natural language processing, speech recognition, and other fields. Compared with traditional machine learning algorithms, deep models have higher accuracy on many tasks. Because deep learning is an end-to-end, highly non-linear, and complex model, the interpretability of deep models is not as good as traditional machine learning algorithms, which brings certain obstacles to the application of deep learning in real life. It is of great significance and necessary to study the interpretability of depth model, and in recent years many scholars have proposed different algorithms on this issue. For image classification tasks, this study divides the interpretability algorithms into global interpretability and local interpretability algorithms. From the perspective of interpretation granularity, global interpretability algorithms are further divided into model-level and neuron-level interpretability algorithms, and local interpretability algorithms are divided into pixel-level features, concept-level features, and image-level feature interpretability algorithms. Based on the above framework, this study mainly summarizes the common deep model interpretability research algorithms and related evaluation indicators, and discusses the current challenges and future research directions for deep model interpretability research. It is believed that conducting research on the interpretability and theoretical

* 基金项目: 国家重点研发计划 (2017YFC1703506); 国家自然科学基金 (61632004); 中央高校基本科研业务费专项资金 (2020YJS027)
收稿时间: 2020-12-14; 修改时间: 2021-03-21, 2021-05-09; 采用时间: 2021-07-05; jos 在线出版时间: 2021-08-03
CNKI 网络首发时间: 2022-11-15

foundation of deep model is a necessary way to open the black box of the deep model, and interpretability algorithms have huge potential to provide help for solving other problems of deep models, such as fairness and generalization.

Key words: deep learning; interpretability; image classification; feature

1 引 言

深度模型[1,2]凭借其强大的表示能力在许多领域取得了巨大的成功, 如图像处理[3−5]、自然语言处理[6,7]、语音识别[8]等领域. 尽管深度模型具有优异的性能, 但是较高的复杂度和非线性导致其透明度低、可解释性差. 在一些现实任务中模型的可解释性和透明性是必要的, 如医学[9−12]、金融、无人驾驶等领域[13,14]. 此外, 研究发现存在一些人类无法识别的图像, 而深度学习模型以很高的置信度给出预测结果[15], 尤其是对抗样本的发现[16]进一步表明深度模型的内部知识可能和人类的认知是有差异的. 由此可见, 深度模型的可解释性研究是非常必要的, 它可以在用户和模型之间建立信任关系[17].

精度并不是算法的唯一评价指标, 良好的可解释性也是一个优秀算法必不可少的, 传统的机器学习算法如线性模型、决策树模型、支持向量机模型等都具有良好的解释性[18−20]. 近年来深度学习的可解释性研究是一个热门课题, 许多解释性算法从不同的视角对深度模型进行剖析[21,22]. 可解释性研究可以帮助用户理解模型的内部表示, 了解模型从数据中抽取了哪些特征进行决策; 其次有助于我们对模型的异常行为和错误进行追踪和定位, 在此基础上可以帮助开发者融合领域知识对模型进行改进甚至设计新的模型; 可解释性研究也为深度学习的发展提供了新的方向, 比如如何在保证模型精度的前提下同时提高模型的可解释性[23], 同时深度模型的可解释性与鲁棒性、公平性、泛化性之间的关系也是值得我们研究的[24,25]; 另外, 人机交互是当前和未来的研究热点, 可解释性研究可以为人类与机器之间的相互沟通和学习提供帮助从而创造更大的价值.

图像分类任务是计算机视觉中较为基础的任务, 其核心是根据图像中的信息对不同类别的目标进行分类, 并实现最小的分类误差. 深度学习最早在图像分类任务上展现了巨大的潜力, 卷积网络[3−5]的局部连接性和平移不变性完美契合了图像数据的特征. 随着图像分类精度的不断提升, 研究者在图像任务上发现深度学习存在许多问题, 如对抗鲁棒性、泛化性、公平性等问题. 解释性研究为打开深度学习的黑箱和解决深度学习当前存在的问题提供了可能. 一方面图像上的特征是我们比较容易理解的, 另一方面其他领域的可解释性研究工作和图像分类领域具有相似性, 因此本文在图像分类任务场景下总结了当前深度学习的可解释性研究进展.

根据建模过程, 可解释性研究工作可分为建模前的可解释性研究、建模中的可解释性研究以及建模后的可解释性研究 3 个方面. 建模前的可解释性研究主要包括数据的预处理和可视化分析, 可以帮助我们充分了解数据的分布特征[26,27]. 建模中的可解释性研究需要我们在解决问题的过程中, 选择人类可理解的特征并采用具备良好解释性的模型进行建模, 如基于规则的模型[28−30]、线性模型[31,32]等. 建模后的可解释性研究主要针对黑箱模型, 通过各种算法如可视化分析、重要性分析等手段对模型进行分析. 作为端到端的黑箱模型, 当前围绕深度模型的可解释性研究工作主要以建模后的解释性算法为主.

本文将深度模型的可解释性算法分为全局可解释性算法和局部可解释性算法[21,33,34], 全局可解释性算法关注模型内部的知识表达, 而局部可解释性算法关注模型对输入样本决策过程的解释性. 在此分类基础上, 通过分析解释性算法之间的关联, 在解释对象的粒度上, 我们进一步将全局可解释性算法分为模型级和神经元级可解释性算法; 将局部可解释性算法划分为基于像素级特征、基于概念级特征以及基于图像级特征的可解释性算法. 针对图像分类任务, 本文依据上述分类框架对目前常用的深度模型的可解释性算法进行归类和总结.

本文在特征的视角下分析了深度模型在图像分类任务上的学习过程, 同时定义了可解释性研究的主要内容; 其次, 在解释粒度的视角下对可解释性算法进行总结归纳, 并在泰勒展开的框架下将部分局部可解释性算法统一起来; 基于对当前可解释算法的研究和总结, 我们认为深度学习的可解释性研究才刚刚起步, 一方面我们需要进一步开展解释性算法和理论基础的研究, 另一方面可解释性算法在模型测试和调试方面存在巨大潜力需要我们进一步挖掘.

本文第 2 节首先给出了可解释性的定义并从特征角度理解深度模型, 其次介绍了本文的可解释性算法的分类依据和整体框架; 第 3 节和第 4 节我们分别阐述了全局可解释性算法和局部可解释性算法; 第 5 节对深度模型可解释性算法的评估指标进行归纳并对解释性算法进行总结比较; 第 6 节中基于对当前的可解释性研究成果的总结, 我们对未来的可解释性研究趋势进行了探讨.

2 深度学习的可解释性

2.1 可解释性定义

图像分类问题以标签作为监督信息, 模型通过从数据中挖掘与标签相关的信息来对图像进行分类. 传统的图像分类算法使用的特征是经过人工筛选和构建的, 具有良好的语义性和解释性; 而深度模型从数据中挖掘的相关模式是我们难以理解的. 此外, 近年来研究发现神经网络和人类的认知是有差异的[15], 这也为我们理解深度模型带来了一定的挑战. 因此, 为了理解深度模型的行为, 我们需要对深度模型进行可解释性研究. 可解释性并没有明确的数学定义, 一些学者从用户的角度出发来定义可解释性[18,21,34,35]. Montavon 等人[21]把解释定义为将模型中的抽象概念映射到人类可以理解的领域中, Miller 等人[35]定义可解释性为人们能够理解模型决策原因的程度. 在本文, 我们将可解释性定义为用户和模型之间的桥梁, 它将模型中的抽象表示映射到人类可理解的语义空间中.

如图 1 所示, 我们将深度模型从数据中提取到的特征分为任务相关特征和任务无关特征. 深度模型尽可能地利用数据中的相关特征进行决策以提高分类准确率, 然而精度并不能作为模型的唯一评价指标, 用户的可理解性对一个模型来说也是至关重要的. 如图 1 中的红色箭头所示, 可解释性研究所做的事情就是在不破坏模型精度的情况下提升模型的可信赖性, 解释器将位于模型特征空间中的相关特征解释为与人类感知相一致的语义特征. 由于深度模型和人类的认知是有差异的[15], 不可避免的深度模型会从数据中挖掘一些不在人类认知范围内的特征, 即非语义特征. Ilyas 等人[36]将这些非语义特征定义为非鲁棒特征, 非语义特征的存在使得模型在当前任务上具有优异的性能, 但是它们也会影响模型的泛化能力和鲁棒性等问题.

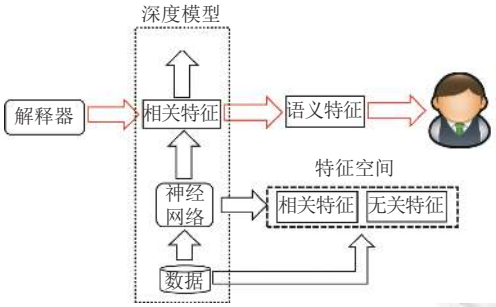


图 1 在特征视角下, 深度模型的学习过程和可解释研究的作用

2.2 深度模型的可解释性算法

由于图像特征比较容易理解, 而且不同领域的解释性算法具有相似性, 因此我们选择在图像分类场景下对深度学习的可解释性研究进行总结. 本文将深度模型的解释性算法分为全局解释性算法和局部解释性算法. 全局解释性算法关注深度模型本身的可解释性, 主要研究模型如何对数据进行表达以及模型中每个神经元的功能; 而局部解释性算法关注模型对单个输入样本决策的解释性, 主要回答模型为什么将输入归为特定类别以及输入中的哪些特征在决策中起到重要作用的问题.

依据解释对象的粒度, 本文将全局可解释性算法分为模型级可解释性算法和神经元级可解释性算法, 模型级可解释性算法主要关注模型对不同类别数据的表达以及模型本身的知识表示, 而神经元级可解释性算法关注深度模型隐藏层神经元的特征表达. 依据解释特征的粒度, 我们将局部可解释性算法进一步划分为以单个像素、像素区域组成的概念以及图像本身作为解释特征基本单位的可解释性算法. 与梯度相关的解释性算法多是以像素为基

本单位来解释模型的[37−41], LIME (local interpretable model-agnostic explanations) 算法[42]以及基于概念激活向量的解释性算法[43−45]均是以概念为基本单位进行决策解释的, 基于实例的可解释性算法以及网络反演的可解释性算法均以图像作为解释性特征[46−49]. 本文的分类框架如图 2 所示, 接下来我们将根据图 2 对深度模型的可解释性算法进行总结.

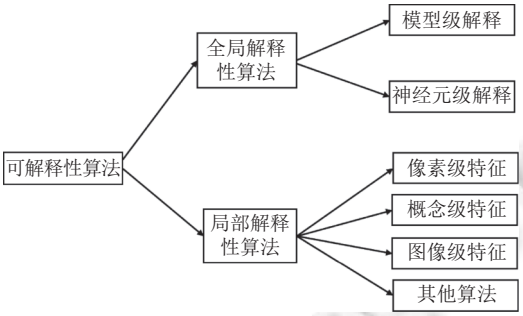


图 2 本文可解释性算法的分类框架

3 全局可解释性算法

深度模型的全局可解释性算法主要关注模型本身的解释性以及神经元粒度上的特征表示, 通过将模型内部的特征表示映射为图像空间的语义特征, 从而提升模型的透明度. 常见的模型级可解释性算法包括激活最大化算法[16,37,50−53]、网络压缩[54,55]、知识蒸馏[56,57]等, 神经元级的解释性算法有基于激活最大化的算法和基于 Network Dissection 框架[58−61]的可解释性算法.

3.1 模型级可解释性算法

3.1.1 模型级激活最大化算法

激活最大化 (activation maximization)[16,37,50−52]算法是分析神经元感兴趣特征的一类算法. Simonyan 等人[37]将该算法应用到深度模型中, 研究神经网络输出层学习到的类别特征. 通过激活最大化来寻找类别语义特征的具体步骤是: 首先, 固定网络的权重参数并输入一张噪声图像 x, 并对其进行优化使得网络输出层类别 c 的输出得分 $S_c(x)$ 达到最大, 优化目标如公式 (1):

$$x^* = \arg\max_x S_c(x) - \lambda \|x\|_2^2 \tag{1}$$

其中, λ 是一个正则化参数, 通过反向传播算法和梯度上升策略不断调整输入图像 x 使得 $S_c(x)$ 达到最大. 通过可视化最终的 x^* 来展示网络所学习的类别特征. 激活最大化是一种模型相关的解释性算法, 能够准确反映模型的真实行为, 通过这种方法获得的语义图像是非自然图像, 如图 3 所示.

(a) Dumbbell (b) Cup (c) Dalmatian

图 3 卷积网络最后一层神经元激活最大化对应的可视化特征[37]

为了使得激活最大化算法找到的语义特征图像接近自然图像, Nguyen 等人[53]将深度生成模型 (GAN)[5]和激活最大化算法结合构建了 DGN-AM 模型, 优化目标如下:

$$z^* = \arg\max_z \ S_c(G(z)) - \lambda\|z\|^2 \tag{2}$$

其中, z 是生成器 G 的输入编码向量, $G(z)$ 是生成器的输出图像. 利用 GAN 的生成能力, 可以生成逼真的自然图像同时捕获神经元的语义特征.

模型级激活最大化算法探究了模型是如何对数据中的类别特征进行抽象表示的, 在一定程度上让用户了解到模型确实从数据中归纳出了与类别相关的决策规则. 然而, 激活最大化算法难以训练, 通常会生成带有噪声和高频模式的非自然图像; 其次, 激活最大化算法适合优化连续性数据, 无法直接应用于离散型数据如文本、图结构数据等.

3.1.2 代理模型

由于深度模型体量较大且结构复杂使得我们难以理解模型的行为, 因此可以通过降低模型的复杂度来提升模型的解释性. 代理模型在原始网络的基础上, 采用复杂度低、解释性好的替代模型来模仿原始模型进行决策. 网络压缩 (network compression)[54,55]和知识蒸馏 (knowledge distillation)[56,57]是这类算法的代表, 代理模型在继承原始模型优异性能的同时降低了模型的复杂度.

压缩算法在保证网络精度的前提下对深度网络的参数、结构等进行压缩, 常见的压缩算法有网络剪枝优化、矩阵分解等算法. Abbasi-Asl 等人[54]分析了压缩算法在卷积网络可解释性方面发挥的作用, 他们将删除卷积核后网络精度的下降程度作为指标对卷积核进行裁剪. 通过可视化剪枝前后网络中每个神经元对应的最大激活图像, Abbasi-Asl 等人发现压缩后的网络特征更加紧凑一致. 此外, 部分工作将卷积核的可解释性作为评估指标来指导模型压缩[55], 在降低模型复杂度的同时保证压缩后的模型具备一定的可解释性.

Hinton 等人提出知识蒸馏算法[56], 通过训练一个小网络来模拟原始网络的决策行为, 将原始模型的知识通过蒸馏提炼到小网络中, 训练小网络时将训练集在原始网络上的伪概率输出作为软标签来监督小网络的学习. 为了更好地理解深度模型的工作原理, Frosst 等人[57]尝试将决策树作为解释性载体, 通过蒸馏算法将网络所学习的分布式特征表示转移到同样是分层模式的决策树中. 文献 [57] 通过使用二元决策树来模仿神经网络的决策过程, 其中决策树内部节点 i 对应着神经网络中一个神经元的过滤器 w_i 和偏执 b_i, 在决策树的每一个内部节点处, 对于输入 x 来说选择右侧分支的概率为:

$$p_i(x) = \sigma(w_i x + b_i) \tag{3}$$

其中, σ 为 Sigmoid 激活函数, 相应的选择左侧分支的概率为 $1 - p_i(x)$. 决策树的每一个叶子节点 l 的输出是 k 个类的概率分布 Q_l, 最终使用两种不同的方式给出测试样本的预测分布——最大路径概率的叶子分布或将叶子分布按其各自路径概率的加权平均值作为最终的预测分布. 通过这样一个二元决策树的学习, 可以将深度模型的层级结构以及网络学习到的知识蒸馏到决策树模型中.

压缩算法和蒸馏算法本质上无法直接为原始模型提供解释性, 因为他们只是原始模型的一个全局近似, 它们无法保证替代模型和原始模型之间的行为完全一致, 因此在解释性上可能无法完全反映原始模型的行为.

模型级可解释性算法通过挖掘模型的类别特征以及简化模型的整体结构, 让用户对模型有一个整体的认识, 为了进一步提升模型的透明度我们需要对模型中每个神经元的功能进行分析.

3.2 神经元级可解释性算法

神经元级可解释性算法主要关注网络中每个神经元所对应的语义特征, 通过将神经元在特征空间中的表示映射到人类可理解的语义空间来揭示神经元所学习到的特征, 这类算法主要有基于激活最大化的可解释性算法[16,51,62]以及基于 Network Dissection 框架的可解释性算法[58−61].

3.2.1 神经元级的激活最大化算法

最初激活最大化算法用来对模型的类别特征进行学习, Yosinski 等人[51]将该算法用来分析隐层神经元的感兴趣特征, 将最大程度激活特定神经元输出的图像特征定义为神经元的语义特征, 优化目标如公式 (4)[16]:

$$x^* = \arg\max_x < \phi(x), e_i > -r(x) \tag{4}$$

其中, $\phi(x)$ 是某层神经元输出的向量化表示, e_i 是标准正交基坐标向量且维度和 $\phi(x)$ 大小一致. $\phi(x)$ 和 e_i 的内积即为该层第 i 个神经元的输出, 通过梯度上升最大化 $\phi(x)$ 和 e_i 的内积即可找到第 i 个神经元所对应的语义特征图像. 我们也可以通过对 e_i 进行组合来找到部分神经元的组合语义特征. 式 (4) 中的第二项是关于 x 的一个正则化项, 主要作用是通过先验约束来提高合成图像 x^* 的质量和可理解性, $r(x)$ 的主要形式有 L_2 正则化[37]、全变分[51]等各种形式的正则化.

一个神经元可以被多种模式激活, 而激活最大化算法倾向于找到一种最大程度激活神经元的模式, 从而忽略了神经元激活模式的多样性. 为了展现神经元特征的多样性, 文献 [52] 的工作结合降维技术和 K-均值聚类算法生成模型特征表示空间的多个聚类中心, 然后利用激活最大化算法来寻找每个聚类中心的最大激活语义图像, 从而揭示深度神经网络学习特征的多样性. DeepDream 算法[62]将任意一张自然图像作为网络初始化输入, 利用激活最大化算法放大单个神经元或部分神经元组合的语义特征, 最终将输入图像变化成各种有趣特征的杂糅.

同样的, 通过激活最大化算法获得的神经元对应的语义特征图像大都是非自然图像, 可以通过文献 [53] 中的算法来找到隐层神经元对应的带有语义信息的自然图像. 通过激活最大化算法探究每层神经元的语义特征, 可以发现神经网络所捕获的特征在语义上有层级的现象, 底层神经元会捕捉到颜色、纹理等信息, 中高层神经元可以逐渐的抓取局部和类别特征.

3.2.2 Network Dissection 框架

Bau 等人[58]提出 Network Dissection 框架, 通过评估隐藏层神经元和语义概念之间的一致性来量化网络神经网络特征表示的可解释性. 为了将单个神经元和语义概念对应起来, 他们构建了一个像素级分割的数据集 (Broden 数据集), 该数据集包含颜色、纹理、场景、材料、物体部分、物体等语义概念. 对于 Broden 数据集中的每个图像 x, 根据语义概念进行语义分割形成概念 c 的二进制模板 $L_c(x)$.

将 Broden 数据集中的每个图像 x 输入到训练好的深度模型中, 提取某个隐藏层中神经元 k 的特征激活图 $A_k(x)$, 然后统计神经元 k 在所有图像上的激活分布 a_k, 通过分布为神经元 k 确定一个激活阈值 T_k 满足 $P(a_k > T_k) = 0.005$, T_k 的作用是为了从激活图中提取具有显著激活的区域. 使用双线性插值将激活图 $A_k(x)$ 扩大到输入图像分辨率大小得到 $S_k(x)$; 然后将 $S_k(x)$ 通过阈值 T_k 进行处理并转化为二进制模板: $M_k(x) \equiv S_k(x) \geqslant T_k$. 针对每个 (k,c) 对, 通过计算 $M_k(x)$ 和 $L_c(x)$ 的交并比来评估神经元 k 对语义 c 的分割能力, 通过公式 (5) 定义神经元和语义概念之间的一致性:

$$IoU_{k,c} = \frac{\sum |M_k(x) \cap L_c(x)|}{\sum |M_k(x) \cup L_c(x)|} \tag{5}$$

$IoU_{k,c}$ 的值反映了神经元 k 检测概念 c 的精度, 为了保证神经元 k 确实能够对语义概念 c 进行分割, 需要进一步对 $IoU_{k,c}$ 进行阈值筛选确保得到有意义的结果. 如果 $IoU_{k,c}$ 满足一定的条件, 则称神经元 k 为概念 c 的检测器.

Network Dissection 框架可以为每个神经元定义其对应的语义特征, 这些特征在概念上和人类的认知是一致的, 因此该算法可以提升深度模型的透明度和解释性. 实验发现在网络训练的过程中, 低级语义特征检测器先出现如颜色、纹理等特征检测器, 随着训练的进行物体部分检测器和对象检测器逐渐出现; 不同网络框架下模型语义特征的分布是不一样的, 不同的训练方式也会导致神经元的语义特征发生变化; 其次, 有的神经元可能对应着多个语义概念特征, 一个语义概念也可能被多个神经元检测到, 这说明神经元和语义概念之间的关系存在一对多和多对一的可能. 在后续工作中, Bau 等人[61]将 Network Dissection 框架应用到生成模型中, 对神经元的功能进行剖析, 并且通过操控神经元来生成带有对应语义特征的图像.

神经元级可解释性算法可以让我们清楚地了解到深度模型中每个神经元的功能和作用, 然而这类算法主要是单独地对每个神经元进行分析缺乏对神经元之间的交互分析, 此外通过神经元级可解释性分析难以获得每个特征在模型决策中的重要程度.

4 局部可解释性算法

局部可解释性算法旨在解释模型对单个输入样本的决策行为, 回答为什么在特定的测试样本上模型会做出特定的分类决定. 与全局可解释性算法不同, 局部解释性算法不需要关注模型的整体结构, 它们专注于模型在测试样本点附近局部空间上的决策行为.

4.1 像素级特征可解释性算法

像素级特征可解释性算法以像素为基本单位来解释模型对输入图像的决策行为. 对于给定的分类器 f 和测试图像 x, 像素级可解释性算法为 x 中的每个像素 x_i 分配一个重要度值或者输出贡献值, 从而得到一个与输入图像大小相同的热力图. 该热力图中的数值大小反映 x 中的像素特征对网络决策的重要程度. 一般来说, 像素级特征和人类容易理解的高级概念并不对应, 但热力图的整体视觉效果能够凸显出输入图像中的一些局部特征, 如轮廓、纹理等语义信息.

本文将像素级特征可解释性算法统一到基于泰勒展开的框架下. 对于给定的分类器 f 和测试图像 x, 可以在 x 的局部邻域内找到一个 x 的参考点 x_0, 则我们可以将函数 f 利用泰勒展开进行近似表示:

$$f(x)=f(x_0)+\left(\frac{\partial f}{\partial x_0}\right)^{\mathrm{T}}(x-x_0)+o((x-x_0)^{\mathrm{T}}(x-x_0)) \tag{6}$$

基于公式 (6) 可以产生两类解释性算法:

(1) 敏感性分析算法: 梯度可以反映输出相对输入中每个维度的敏感程度, 因此可以将局部梯度作为一种解释性[37], 即 $\frac{\partial f}{\partial x_0}$, 在局部小邻域内 f 是近似线性的, 即 $\frac{\partial f}{\partial x_0}=\frac{\partial f}{\partial x}$;

(2) 归因算法: 将 f 的一阶微分作为一种解释, 即 $\frac{\partial f}{\partial x_0}\odot(x-x_0)$, $\odot$ 代表逐元素相乘.

相关的可解释性算法总结在表 1 中, 接下来我们将根据表 1 对该类算法进行总结. 此类算法通过反向传播算法获得最终的解释结果, 因此计算较为简便高效.

表 1 像素级特征可解释性算法及其分类

算法类型	相关的可解释性算法
敏感性分析	Saliency Map[37]、SmoothGrad[63]、DeconvNet Visualization[38]、Guided Backpropagation (GBP)[64]
归因算法	Input⊙Gradients[65,66]、Integrated Gradients (IG)[67]、LRP[41]、DeepLIFT[68]、TD[41]、DTD[69]

4.1.1 敏感性分析算法

敏感性分析算法是比较常见的解释性算法, 用于分析输入样本每一维度的变化对输出的影响. Simonyan 等人[37]提出 Saliency Map 算法, 该算法直接将深度模型的输出对输入图像的梯度作为解释结果. 对原始梯度进行绝对值处理、阈值过滤等操作[37]可以提升梯度可视化的视觉效果. 由于深度模型的复杂度和高度非线性导致梯度存在很多问题.

(1) 梯度存在噪声, 一方面显著性图中无规则的噪点会造成视觉扩散, 从而影响解释结果的有效性, 另一方面这些噪点的存在也说明模型学习到了一些无关的特征模式;

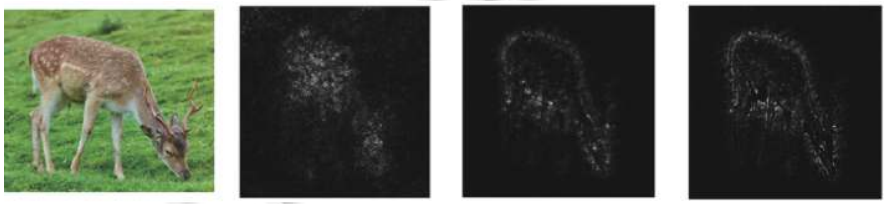
(2) 梯度饱和现象, 由于激活函数的存在会导致反向传播的过程中出现梯度为 0 的现象, 从而忽略掉图像中的重要特征信息[67];

(3) 梯度的连续性较差, 对于微小扰动产生的相似样本, 梯度的解释结果可能会出现差异较大的现象.

为了改进原始梯度存在的视觉扩散效应, Smilkov 等人[63]提出了平滑梯度算法 (SmoothGrad) 来消除原始梯度可视化产生的噪点, 该算法主要是通过在 x 的邻域内进行多次采样得到 n 个样本 $\{x_i\}_{i=1,\ldots,n}$, 其中 x_i 是通过对 x 加入高斯噪声生成的 $x_i=x+N(0,\sigma^2)$, 对这些样本的梯度进行平均来平滑掉原始梯度中出现的噪点[38]. SmoothGrad 算法可以和许多与梯度相关的可视化算法相结合, 如类激活图 (class activation mapping)[70]等.

为了提升梯度可视化的效果, 反卷积可视化 (DeconvNet visualization) 算法[38]和导向反向传播 (guided back-propagation) 算法[64]对梯度的反向传播过程进行修改, 他们将高层激活信号通过反向传播机制逐层传递到输入空间, 并在输入空间重构与特定神经元激活或分类决策相关的模式. 反卷积可视化[38]的流程主要由反池化、反ReLU、反卷积 3 个操作组成, 该算法和梯度反向传播过程类似, 唯一不同的地方是经过 ReLU 时的操作. 如果前向传播过程中神经元被抑制, 则正常的梯度反向传播过程中梯度经过该神经元时变为 0. 反卷积可视化算法在反向传播时, 不考虑前向传播过程中神经元的激活情况, 直接对反向传递回来的信号用 ReLU 函数进行激活处理. GBP 算法[64]将原始梯度和反卷积可视化算法结合, 梯度在反向传播经过 ReLU 时, 不仅考虑前向传播的激活情况同时考虑反相传播的激活情况.

相比于原始梯度信息, 反卷积可视化和 GBP 算法在梯度反向传播过程中尽可能地保留正的梯度信息, 突出强调那些对分类结果有积极作用的特征, 能够更好地展示网络学习到的特征. 如图 4 所示, 对于一张输入图像, 经过不同的反向传播过程得到的解释性结果. 可以发现反卷积可视化和 GBP 算法的可视化结果要比原始梯度效果好, 它们能够更好地刻画出对象的轮廓信息. 利用反卷积可视化算法对隐藏层的神经元特征进行分析, 可以观察到神经网络学习的特征逐层变得复杂[38], 浅层神经元主要关注颜色、边缘等低层特征, 随着层数的加深模型会学习到复杂的纹理特征、局部特征甚至是类别对象特征.



(a) Input (b) VanillaGrad (c) DeconvNet (d) GBP

图 4 对于一个卷积模型和一张输入图像, 利用不同可视化算法获取的特征

敏感性分析算法从视觉上来说能够为用户提供清晰的语义信息, 然而一些工作发现这类算法存在缺陷: Nie 等人[71]实验发现 GBP 算法和 DeconvNet 可视化算法对网络的重参数化不明感, 即改变网络参数解释结果几乎不变, 他们从理论上证明了这两个算法在对输入样本进行部分重构; Ghorbani 等人[72]和 Dombrowski 等人[73]提出了针对基于反向传播方法的攻击算法, 在测试样本上添加扰动在不改变输出预测的情况下导致截然不同的解释结果. 这些发现为可解释性算法的可信赖性提出了质疑.

4.1.2 归因算法

归因算法将网络的预测分数 $f(x)$ 通过逐层反向传播到输入的每个维度上, 为输入的每个维度分配一个贡献值 R_d. 归因算法具有完备性, 即输出得分可以近似由输入的各个维度的贡献之和表示:

$$f(x) \approx \sum_{d=1}^{N} R_d \tag{7}$$

从泰勒展开的角度来说, 如果我们找到一个参考点 x_0 满足 $f(x_0)=0$, 那么我们可以将网络输出通过一阶微分进行近似:

$$f(x) \approx \left(\frac{\partial f}{\partial x_0}\right)^{\mathrm{T}} (x - x_0) \tag{8}$$

一阶微分近似在一定程度上反映了输入样本和参考点在网络输出上的差异. 如果 $f(x_0)=0$ 且 $f(x)>0$, 那么我们可以认为由于 x 中存在 x_0 中不存在的特征从而为分类提供证据, 这就是简单泰勒分解对深度模型进行解释的依据[41].

直接使用 Input$\odot$Gradients[65], 即 $\dfrac{\partial f}{\partial x}\odot x$, 在一定程度上能够缓解梯度扩散的问题. Kim 等人[66]利用 GBP 算法原理, 在梯度逐层反向传播的过程中设置阈值过滤掉较小的梯度, 最终使用修正的梯度和输入样本的逐元素乘积作为解释结果. 为了解决原始梯度可视化存在的饱和问题, 文献 [67] 提出了积分梯度 (integrated gradients) 算法,

主要思想是在参考点 $\bar{x}$ 和测试图像 x 之间对梯度函数进行路径积分. IG (integrated gradients) 的解释结果通过公式 (9) 计算:

$$IG(x)=(x-\bar{x})\times\int_0^1\frac{\partial S_c(\bar{x}+\alpha(x-\bar{x}))}{\partial x}d\alpha \tag{9}$$

其中, $S_c(x)$ 是网络在目标类别上的输出结果, $\bar{x}$ 是参考点一般选择全黑图像. 公式 (9) 中积分项是对原始图像的像素在 [0, 1] 尺度上进行多次缩放, 然后计算这些样本梯度的均值, 因此 $IG(x)$ 是 Input⊙Gradients 的一种特殊形式. 通过引入与输入相似的多个样本来计算梯度, 在一定程度上能够解决梯度弥散和饱和的问题. 但是, 通过这种方式会引入虚假相关性, 因为深度模型内部对这些尺度缩放图像的处理可能存在很大的差异, 尽管他们可能会被网络识别为同一类别.

分层相关性传播 (layer-wise relevance propagation) 算法[41]将网络的输出得分 $f(x)$ 作为相关性得分逐层反向传播, 最终将 $f(x)$ 分配到输入样本 x 的各个维度上, 通过定义每个像素点的贡献来衡量像素级特征的重要性, 最终通过热力图将解释结果呈现给用户, 从而提供模型决策的依据. 假设 l 层神经元预激活的输出 z^l, 为 z^l 的每一维度 z_d^l 分配一个相关性得分 R_d^l, LRP 在传播过程中满足完备性假设: 每层神经元对应的相关性得分之和相等, 即满足公式 (10) 条件.

$$f(x)=\ldots=\sum\nolimits_{d\in l+1}R_d^{l+1}=\sum\nolimits_{d\in l}R_d^l=\ldots=\sum\nolimits_d R_d^1 \tag{10}$$

在反向传播的过程中, $l+1$ 层第 k 个神经元分配给 l 层第 i 个神经元的相关性得分用 $R_{i\leftarrow k}^{(l,l+1)}$ 表示, 那么反向传播过程中应该遵循以下规则:

$$R_i^l=\sum\nolimits_{k\in l+1}R_{i\leftarrow k}^{(l,l+1)} \tag{11}$$

基于以上的传播规则, Bach 等人[41]提出了一个合理的相关性传播方案:

$$R_{i\leftarrow k}^{(l,l+1)}=\frac{z_{ij}}{z_j}\cdot R_j^{l+1} \tag{12}$$

为了避免出现 z_j 过小导致 $R_{i\leftarrow k}^{(l,l+1)}$ 无界的情况, 可以通过引入预定义的稳定器 ε 来克服无界的问题 (ε-LRP):

$$R_{i\leftarrow k}^{(l,l+1)}=\begin{cases}\dfrac{z_{ij}}{z_j+\varepsilon}\cdot R_j^{l+1}, & z_j\geqslant 0\\ \dfrac{z_{ij}}{z_j-\varepsilon}\cdot R_j^{l+1}, & z_j<0\end{cases} \tag{13}$$

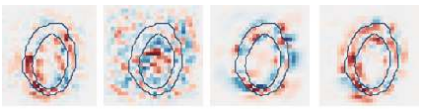
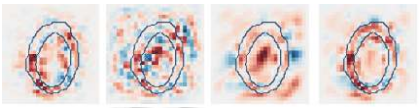
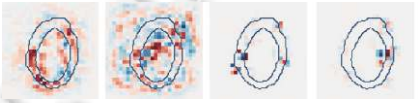
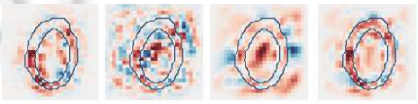
本质上来说, LRP 是基于泰勒展开的一种特殊的反向传播形式. 为了避免相关性的遗漏, Bach 等人[41]建议将神经元激活 z_{ij} 进一步划分为积极的和消极的激活并分别进行反向传播. LRP 算法不依赖梯度信息, 因此可以对不可微的模型进行解释分析, 该算法在图像分类、视频分析、医疗诊断等领域具有广泛的应用场景[21].

Shrikumar 等人[68]提出了 DeepLIFT (deep learning important features) 算法同样是将网络的预测结果分配到输入的每个维度上, 不同的是 DeepLIFT 将每个神经元的激活 z 与其"参考激活" $\bar{z}$ 进行比较, 并根据差异将高层神经元的贡献得分分配给下一层神经元, 给定参考点 $\bar{x}$ 将其输入网络中并记录网络中每个神经元的参考激活 $\bar{z}$, 则 DeepLIFT 算法的传播规则如公式 (14) 所示[74]:

$$R_i^l=\sum\nolimits_j\frac{z_{ij}-\bar{z}_{ij}}{\sum_{i'}z_{i'j}-\sum_{i'}\bar{z}_{i'j}}\cdot R_j^{l+1} \tag{14}$$

类似于 LRP 算法, DeepLIFT 算法也可以分别考虑积极的贡献和消极的贡献来避免相关性遗漏. 从传播方式上来看, LRP 算法是 DeepLIFT 算法的一个特例, 即考虑参考点 $\bar{x}=0$. DeepLIFT 算法[68]建议参考点的选择要根据领域知识, 且最好针对多个不同的参考计算 DeepLIFT 得分. Ancona 等人[40,74]将 Integrated Gradients、LRP、DeepLIFT 统一了起来, 他们形式上和输入⊙梯度是一致的, 不同之处是梯度的求取方式上存在差异, 表 2 给出了这些算法的比较以及它们在 MNIST 数据上的可视化效果.

表 2 基于一阶微分近似的归因算法及其比较[40,74]

方法	归因 $R_i(x)$	MNIST样本归因结果 ReLU	tanh	Sigmoid	Softplus
Input⊙Gradients	$x_i \cdot \frac{\partial S_c(x)}{\partial x_i}$				
Integrated Gradients	$(x_i-\bar{x}_i)\cdot\int_{\alpha=0}^{1}\frac{\partial S_c(\tilde{x})}{\partial \tilde{x}_i}\Big\vert_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})}d\alpha$				
ε-LPR	$x_i \cdot \frac{\partial^g S_c(x)}{\partial x_i},\ g=\frac{f(z)}{z}$				
DeepLIFT	$(x_i-\bar{x}_i)\cdot\frac{\partial^g S_c(x)}{\partial x_i},\ g=\frac{f(z)-f(\bar{z})}{z-\bar{z}}$				

参考点的选择是一阶泰勒近似类解释算法的关键, 不同的参考点选择会得到不同的解释性结果. 为了获得良好的解释结果, 泰勒分解 TD (Taylor decomposition)[41]建议我们选择根点作为参考点, 根点位于分类边界上可以很好的为输入样本提供局部的解释性, 然而寻找根点是一个非凸优化的过程, 而且并不是所有的样本都能够找到对应的根点. Montavon 等人[69]基于泰勒分解和分治思想的启发提出了深度泰勒分解算法 DTD (deep Taylor decomposition) 对神经网络进行解释. DTD 方法将深度网络学习的复杂函数在结构上分解为一组子函数的和, 同时该算法提供了在不同情况下如何寻找根点或者近似根点的方法.

4.2 概念级特征可解释性算法

概念级特征可解释性算法需要构建语义概念集合, 并通过重要性分析来衡量每个概念特征在模型决策中重要度. 利用超像素分割算法可以将图像划分为不同的超像素区域, 这些区域可以和语义概念对应起来, 通过衡量测试图像中的每个概念特征在模型输出上的贡献来对深度模型的决策进行解释. LIME 算法和基于概念激活向量 CAV (concept activation vectors) 的可解释性算法[43−45]都是在概念特征的基础上对深度模型的决策行为进行解释的. 在本文第 3.2.2 节介绍的基于 network dissection 框架[58−61]的可解释性算法也是一种概念级特征的解释性算法.

4.2.1 LIME 算法

LIME 算法[42]使用复杂性低的可解释模型来对深度模型的局部映射关系进行近似, 进而使得我们可以理解深度模型的局部决策行为. 其次, 该算法是一个模型无关的算法, 意味着无论模型多复杂, 如 SVM 或神经网络, 该解释器都可以工作. LIME 算法的具体步骤如下: 对于一个输入图像 x 和分类网络 f, 首先对 x 进行超像素分割形成 d 个图像块区域; 然后对这 d 个超像素区域进行随机采样, 生成 x 局部邻域内的 n 个样本 $\{z_i\}_{i=1,\ldots,n}$, 其中 $z_i \in \{0,1\}^d$ 的每个维度代表着生成的样本是否包含对应的超像素块区域; 将生成的样本输入到模型 f 中等到 $f(z_i)$, 将 $f(z_i)$ 作为监督信息训练一个可解释模型 $g(z_i)$; 最后将那些拥有较大权重系数的超像素区域进行融合生成测试图像的一个解释性区域. LIME 算法的关键是保证在测试样本的局部区域内解释器和原始模型的行为一致, 其次要保证解释器拥有良好的可理解性, 因此算法优化目标如下:

$$\xi(x)=\arg\min_{g\in G}\mathcal{L}(f,g,\pi_x)+\Omega(g) \tag{15}$$

其中, π_x 用来衡量样本 z_i 和 x 之间的紧邻程度, $\Omega(g)$ 是一个衡量解释器 g 的复杂性的度量, 例如: 对于决策树 $\Omega(g)$ 可以是树的深度, 而对于线性模型 $\Omega(g)$ 可以是非零权重的数量. LIME 算法是一种简单易操作的算法, 然而由于算法本身存在很多不确定因素, 如采样的随机性、采样样本与测试样本近似程度的衡量等[75], 这些不确定性在一定

程度上影响解释模型的效果, 其次局部近似也很难保证解释模型和深度模型在行为上完全一致. 为了改进 LIME 算法的缺陷提升解释的准确性, 作者在后续工作中提出了基于 if-then 规则的 Anchors 算法[76]来描述模型的局部行为, 使得用户可以通过充分条件来推测模型的行为.

4.2.2 基于概念激活向量的可解释性算法

基于概念激活向量的可解释性算法[43−45]核心思想是通过度量一个概念对网络输出的重要程度来解释模型的预测结果. 该类解释性算法需要构建一个概念集合以及与每个概念相关的正负样本集合, 同时需要学习每个概念对应的概念激活向量, 然后计算深度模型的输出在对应概念激活向量上的方向梯度, 通过方向梯度的大小来反映该概念对测试样本预测输出的相关性.

Kim 等人[43]提出概念激活向量的解释性算法, 首先需要定义一个感兴趣的概念集合 C (例如, 条纹), 同时构建概念相关的正样本集 P_C (例如, 条纹相关的图像) 和负样本集 N (例如, 随机的图像集合); 然后针对特定的网络中间层 l, 获得正负样本集的隐层表示 $\{f_l(x): x \in P_C\}$ 和 $\{f_l(x): x \in N\}$, 之后训练一个线性二分类器来区分 $\{f_l(x): x \in P_C\}$ 和 $\{f_l(x): x \in N\}$, 该分类器的法向量即为此概念的概念激活向量 v_C^l. 针对测试样本 x (例如, 斑马图像), 假设网络 l 层到输出层的函数为 h_l, 则概念 C (条纹) 对测试样本预测的重要程度可以通过下式来衡量:

$$S_{C,k,l}(x) = \lim_{\varepsilon \to 0} \frac{h_{l,k}(f_l(x) + \varepsilon v_C^l) - h_{l,k}(f_l(x))}{\varepsilon} = \nabla h_{l,k}(f_l(x)) \cdot v_C^l \tag{16}$$

其中, $h_{l,k}$ 代表网络在类别 k 上的输出值, $S_{C,k,l}(x)$ 可以反映概念 C 对测试样本预测的重要性. 为了衡量一个概念对整个类别的重要程度, Kim 等人[43]提出了 $TCAV$ 度量计算如公式 (17):

$$TCAV_{Q_{C,k,l}} = \frac{|\{x \in X_k : S_{C,k,l}(x) > 0\}|}{|X_k|} \tag{17}$$

即, 对于所有标签为 k 的图像, 通过计算与概念 C 相关的图像占整个图像集的比例来反映概念和类别的相关性. 为了防止某些概念对测试类产生虚假的相关性, 文献 [43] 提出需要对 $TCAV$ 进行统计假设检验. 该算法的关键是构建一个高质量的概念集并搜集与概念相关的数据集 P_C, 这样才能保证学到有效的概念激活向量. 对于测试图像 x, 将那些 $TCAV$ 分数较高的概念进行拼接, 可以得到模型在测试样本中所关注的决策区域.

为了避免人工设计概念集和构建数据集带来的误差, 文献 [44,45] 提出利用超像素分割来提取概念集和构建对应的数据集. 由于每层神经元的感受野是不同的, 因此我们需要使用不用的超参数实施超像素分割从而形成不同层级的概念. 为了生成概念集合, 我们需要对分割生成的图像进行聚类, 然后对聚类后的图像集进行概念标注. 为了避免超像素分割生成无意义的概念和重复的概念样本, 在聚类前我们需要对数据集进行预处理剔除异常值和重复图像.

概念级特征可解释性算法能够将图像中的特征和语义概念对应起来, 并估计每个概念对于类别决策的重要性. 然而, 概念级特征可解释性算法过分依赖人为定义的概念集, 这在一定程度上会低估深度模型的表达能力.

4.3 图像级特征可解释性算法

图像级特征可解释性算法通过在输入空间中寻找与测试样本具有相似特征的样本, 从而对深度模型的决策行为进行解释. 网络反演[46,47]和基于实例的可解释性算法[48,49]均以图像级特征作为解释依据. 网络反演通过将神经元的激活特征图进行重构, 从而寻找具有相同激活效果的样本来对测试样本以及神经元的特征进行解释. 而基于实例的可解释性算法则是通过在训练集中寻找具有代表性的样本来对测试样本进行解释.

4.3.1 网络反演

网络反演 (network inversion)[46,47]通过对神经元的激活模式进行重构, 以此说明每个神经元学习到的特征. 网络反演算法主要有基于正则化的算法[46]和基于反卷积网络的算法[47], 算法流程如图 5 所示.

基于正则化的网络反演算法[46]流程如下: 首先针对测试样本 x, 通过原始网络得到其在某层单个神经元或多个神经元的特征激活图 $A(x)$; 然后, 将一个随机噪声 x_0 作为初始化输入原始网络, 通过限制目标神经元的激活模式 $A(x_0)$ 和原始输入图像激活模式 $A(x)$ 相同来重构神经元的特征, 其优化目标如公式 (18):

$$x^* = \arg\min_x \mathcal{L}(A(x), A(x_0)) - \lambda r(x) \tag{18}$$

其中, $\mathcal{L}$ 损失可以是均方误差损失等, $r(x)$ 的作用主要是为了保证生成图像的质量, 主要形式有 L_2 正则化、全变分等各种形式的正则化.

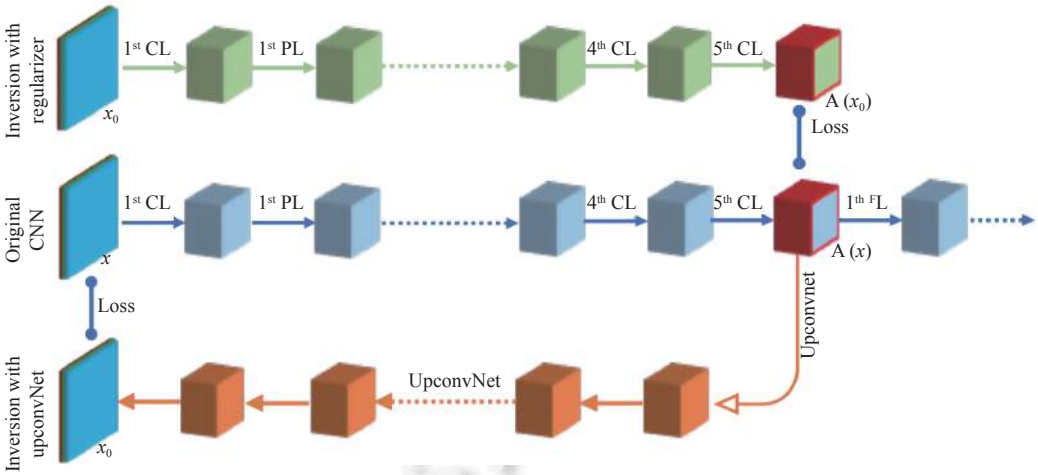


图 5 两种网络反演的算法框架[77]

基于反卷积的输入重构算法[47], 需要重新训练一个反卷积网络 D, D 的作用是将特征激活图中的信息进行上采样并在输入空间进行重构, 通过优化下式来训练网络 D:

$$W^* = \arg\min_W \sum_i \|x_i - D(A(x_i), W)\|^2 \tag{19}$$

其中, W 是反卷积网络 D 的参数. 训练完成后, 将特定神经元的激活特征图 $A(x_i)$ 作为 D 的输入来对神经元的激活进行重构.

网络反演算法可以分析每层神经元提取的语义特征, 通过比较不同隐层特征图重构的语义图像可以发现底层特征图能够清晰地重构输入图像, 随着网络层数的升高特征图的重构结果越来越模糊, 但是仍然能够保留类别相关特征. 这说明神经网络在特征提取过程中逐渐会丢弃掉一些不重要的特征, 而保留那些重要的具有判别性的特征. 网络反演算法的缺陷在于针对不同层的激活特征需要训练不同的优化算法或者反卷积网络进行特征重构; 其次, 此类算法不能为每个神经元的特征提供重要性评估.

4.3.2 基于实例的可解释性算法

基于实例的可解释性算法通过从训练样本中寻找具有代表性的样本来对测试样本进行解释性分析, 隐含的假设是测试样本和代表性样本有较好的相似性, 因此网络会给出相同的预测结果[19]. 如何寻找具有代表性的实例是此类算法的关键, 一些算法通过设计评价指标来评估训练集中每个样本对测试样本的影响力来找到代表样本.

原型 (prototypes) 通常被认为是样本空间中具有代表性的数据点, 然而 Kim 等人[48]认为仅凭原型并不能完全代表模型内部的复杂表示, 为了更好地理解模型的抽象表示能力, 我们还需要解释原型未捕获的内容. 文献 [48] 将那些原型不能很好代表的数据点称为批评 (criticisms), 原型和批评都来自样本空间中的真实样例. Kim 等人[48]提出基于 MMD-critic 的评判标准从训练集中选择具有代表性的原型和批评. 最大均值差异 (MMD) 是两个分布之间的差异的度量, 由两个分布在函数空间 $\mathcal{F}$ 上差异的上确界给出. Kim 等人[48]选择能够最小化数据分布和原型分布之间 MMD 距离的数据点作为原型代表, 而批评则对应着最大化 MMD 距离的样本点. 给定训练集中属于同类的 n 个样本 $X = \{x_i, i \in [n]\}$, $[n]$ 代表整数集合 $\{1,2,\ldots,n\}$, 从 X 中寻找一个子集 $X_S = \{x_i, \forall i \in S\}$ 其中 $S \subseteq [n]$, 给定一个核函数 $k(\cdot,\cdot)$ 来度量样本间的相似性, Kim 等人[48]根据 X 和 X_S 之间的 MMD 距离定义了如下损失函数 $J_b(S)$:

$$J_b(S)=\frac{1}{n^2}\sum_{i,j=1}^{n}k(x_i,x_j)-MMD^2(\mathcal{F},X,X_S)=\frac{2}{n|S|}\sum_{i\in[n],j\in S}k(x_i,y_j)-\frac{1}{|S|^2}\sum_{i,j\in S}k(y_i,x_j) \tag{20}$$

通过最大化 $J_b(S)$ 从 X 中选择 m_* 个原型, 通过优化公式 (21):

$$\max_{S\in 2^{[n]},|S|\leqslant m_*} J_b(S) \tag{21}$$

优化公式 (21) 一般是困难的因为子集的搜索空间很大, 因此 Kim 等人[48]设计了一个贪婪选择算法来有效的寻找原型集合. 为了提取批评样本集合 X_C, 即那些不能被原型很好代表的样本, Kim 等人[48]基于见证函数设计了另一个损失函数:

$$L(C)=\sum_{l\in C}\left|\frac{1}{n}\sum_{i\in[n]}k(x_i,x_l)-\frac{1}{m}\sum_{j\in S}k(x_j,x_l)\right| \tag{22}$$

通过选择最大化公式 (22) 定义的损失函数来寻找 X_C. 实验发现基于 MMD-critic 寻找到的原型样本一般对应着数据分布中高密度区域的样本点, 而批评样本对应着数据分布中远离分布中心的一些带有异常特征的样本点.

如果删除训练集中的一个训练实例, 模型的参数和预测结果会受到较大的影响, 那么我们称这样的实例为有影响力的实例[19]. 要验证一个实例是否具有影响力最直接办法是将实例从训练集中剔除然后重新训练模型, 然而这对深度模型来说是很不现实的, 文献 [49] 通过使用影响函数 (influence functions) 来衡量实例如何影响模型的参数和预测, 从而避免了这种耗时的训练过程. 删除训练集中的一个数据 x 对测试样本 x_{test} 的影响通过公式 (23) 计算:

$$\mathcal{I}(x,x_{\text{test}})=-\nabla_\theta L(x_{\text{test}},\theta)^{\mathrm{T}}H_\theta^{-1}\nabla_\theta L(x,\theta) \tag{23}$$

其中, θ 是网络训练后的参数, H 为损失函数 L 对网络参数 θ 的二阶导数. 公式 (23) 中逆矩阵的计算量是耗时的, 文献 [49] 提出通过共轭梯度法和随机估计法来解决公式 (23) 中的计算问题. Ribeiro 等人通过实验发现针对相同的测试样本, 相同性能的模型根据影响函数找到的具有影响力的实例可能不同; 同时与传统机器学习算法相比, 神经网络更关注于在概念上具有相似性的实例[49].

Yeh 等人[78] 基于再生核希尔伯特空间中的表示定理 (representer theorem) 对深度模型的预测进行解释, 他们证明了测试样本的预测值可以分解为训练点特征表示的线性组合, 并通过线性组合的权重从训练集中提取兴奋样本和抑制样本来解释模型对测试样本的决策. Yeh 等人的工作[78]将深度分类模型分为两部分: 输出层 (logits) 和特征表示层 (logits 层之前的所有层), 其中特征表示层的输出为 $f_i=\Phi_2(x_i,\Theta_2)$, 输出层的输出为 $\Phi(x_i,\Theta)=\Theta_1 f_i$, 最终网络的输出为 $\hat{y}_i=softmax(\Phi(x_i,\Theta))$. 根据表示定理, 测试样本的网络预测值可以分解为训练样本点特征表示 f_t 的加权线性组合, 即 $\Phi(x_t,\Theta)=\sum_{i=1}^{n}\alpha_i k(x_t,x_i)=\sum_{i=1}^{n}\alpha_i f_i^{\mathrm{T}}f_t$, 权重 α_i 叫做 x_i 对 x_t 的表示值. 文献 [78] 中将正、负表示值对应的训练样本分别定义为兴奋样本和抑制样本. 与影响函数[49]算法相比, 该算法计算相对简单且通过提供抑制样本可以加深对模型决策的理解.

基于实例的解释性方法通常选择数据集中有代表性的实例, 可以帮助人们理解复杂的数据分布, 同时有助于我们改进和构建合适的机器学习模型.

4.4 其他局部可解释性算法

除上述基于特征粒度总结的局部可解释性算法外, 还有很多其他的可解释性算法, 如类激活图算法[79]、基于扰动的可解释性算法[38,80−83]、局部近似模型[84]等. 类激活图算法[79]和基于扰动的可解释性算法[38,80−83]通过从输入中定位出模型决策所依赖的特征区域来对模型的决策进行解释, 局部近似模型[84]是通过使用解释性较好的简单模型来局部近似深度模型的行为.

4.4.1 类激活图算法

类激活图 CAM (class activation mapping) 算法[79]通过在输入图像中定位出与网络决策相关的区域来解释网络的决策机制. 2016 年 Zhou 等人提出类激活图算法[79], 阐明了用于分类任务的卷积神经网络具有非凡的定位能力, 同时展示了网络高层神经元所关注的特征. CAM 算法需要将卷积网络中的全连接层用全局平均池化层[85]代替, 该算法的框架如图 6 所示, 首先将最后一个卷积层的输出特征图进行上采样得到和输入图像大小相同的特征图 $f_k(x,y)$, 最终通过计算 $f_k(x,y)$ 的线性组合得到类激活图:

$$M_c(x,y)=\sum_k w_k^c f_k(x,y) \tag{24}$$

其中, w_k^c为输出类别c所对应的权重. 从图 6 中可以看到 CAM 算法可以准确定位到对象所在的区域.

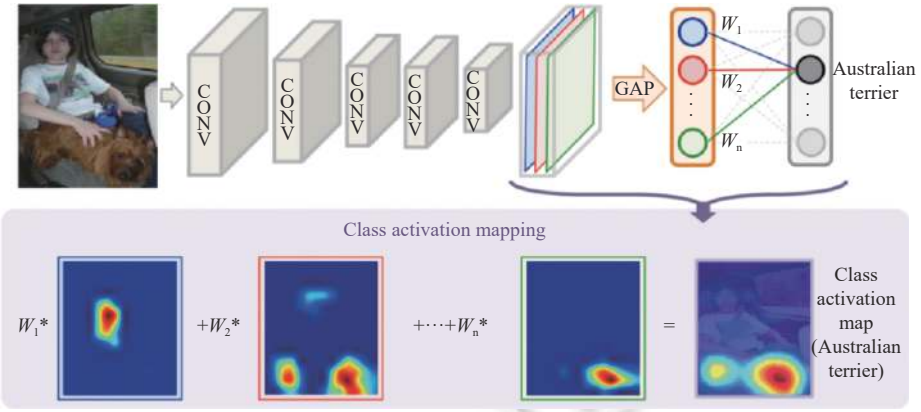


图 6 类激活图算法的框架和可视化效果[79]

为了将 CAM 算法应用到一般的网络中, 文献 [25] 提出了 Guided Grad-CAM 算法将 GBP 算法和 CAM 算法相结合, 可以对每一层神经元在输入上获取的特征进行定位, 算法流程如图 7 所示. Guided Grad-CAM 不再对网络有任何限制, 通过求得输出类得分对某一层所有特征图的梯度, 然后通过对梯度进行通道粒度上的平均, 可以得到该层特征图的线性组合系数, 最终的类激活图也是通过特征图的线性组合得到的, 不同之处是需要对组合后的结果使用 ReLU 函数进行激活提取正向信息:

$$w_k^c=\frac{1}{z}\sum_i\sum_j\frac{\partial Y^c}{\partial A_{ij}^k}\quad L_{\text{Grad-CAM}}=\text{ReLU}\left(\sum_k w_k^c A^k\right) \tag{25}$$

其中, z 为一个常数, 代表特征图中的像素点数目.

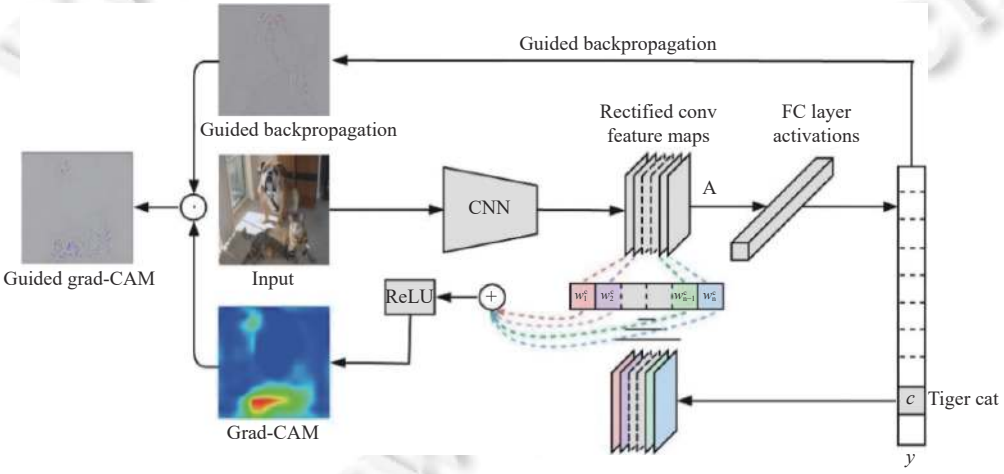


图 7 Guided Grad-CAM 算法框架和可视化效果[25]

CAM 的算法在弱标签物体定位方面和可解释性方面具有较大的潜力, 该算法可以在各种流行的架构上被广泛应用. 由于 CAM 算法在计算过程中使用了梯度信息, 在第 4.1.1 节中我们总结到梯度具有扩散和饱和的现象, 因此梯度的这些缺陷会使得 CAM 算法对特征的重要性产生错误的估计, 从而导致获得错误的解释性结果. 其次, CAM 算法是易受攻击的, Subramanya 等人[86]设计了对 Grad-CAM 解释性结果的攻击算法, 他们使用特定的攻击

策略对测试图像进行扰动, 不仅能够愚弄模型的输出结果, 同时可以操控扰动后图像的解释结果偏离扰动区域.

在 CAM 算法的框架基础上后续还有许多改进的算法, 诸如 Grad-CAM++算法[87]、SmoothGrad-CAM++[70]、RISE 算法[88]、SS-CAM 算法[89]、Score-CAM 算法[90]等. Grad-CAM++算法[87]拥有更优异的定位能力, 该算法可以定位出图像中同一类别的所有对象区域, 从而为深度模型的预测提供可靠的解释性结果. 在第 4.1.1 节中我们提到 SmoothGrad 技术可以降低梯度中的噪声[63], 因此 SS-CAM 算法[89]将 SmoothGrad 技术应用到 CAM 算法中来解决梯度存在的噪声问题. 为了彻底解决 CAM 算法中的梯度缺陷问题, Score-CAM 算法[90]抛弃了通过梯度来估计特征图重要性的做法, 它利用激活特征图生成掩码对测试样本进行扰动, 通过评估网络输出相对扰动的变化程度来估计每个特征图的重要性. RISE[88]算法是模型无关的算法可以为任何黑箱算法进行解释, 该算法通过蒙托卡诺采样生成多个掩码对测试样本进行扰动, 将扰动后的样本输入网络中来估计掩码的重要性, 最后通过掩码的线性组合生成测试样本的解释性区域. 总的来说, 基于 CAM 的可解释性算法计算简单而且能够有效地反映网络在输入中关注的特征.

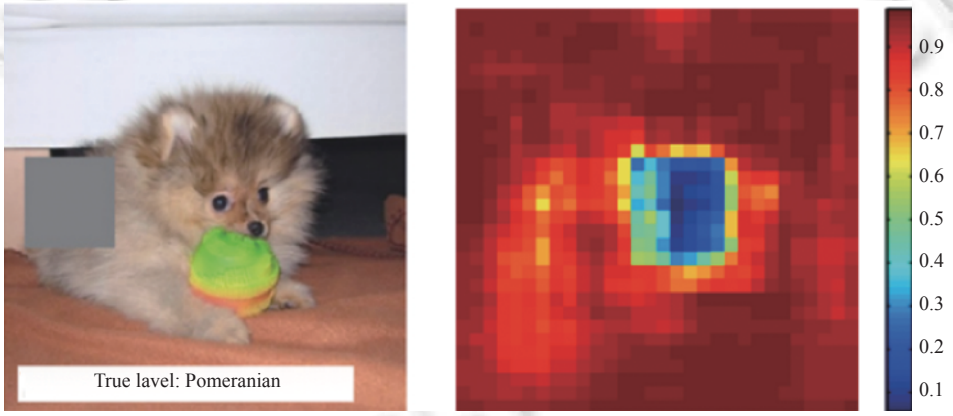
4.4.2 基于扰动的特征重要性分析

基于扰动的特征重要性分析的可解释性算法[38,80−83]主要思想是对输入图像 x 进行多次扰动, 通过对比扰动前后的图像在网络输出上的差异来判定被扰动特征对网络输出的重要程度, 扰动的方式主要分为遮挡[38,80]、擦除[80]、掩码[88,91]等方式. 我们可以将基于扰动的算法统一到泰勒展开的范式下, 将扰动后的图像视为 x_0, 那么神经元或分类器输出 f 的变化可以用公式 (26) 进行估计:

$$f(x)-f(x_0)\approx\left(\frac{\partial f}{\partial x_0}\right)^{\mathrm{T}}(x-x_0) \tag{26}$$

与基于反向传播算法相比, 基于扰动的算法是模型无关的算法, 它们不需要访问模型内部的参数.

遮挡是最为常见的一种扰动方式, Zeiler 等人[38]使用一个灰色的方块对图像进行滑动式的遮挡, 将连续遮挡产生的图像输入网络中, 将分类器输出的类别概率作为像素空间位置的函数进行可视化, 通过这种方式可以定位出原始图像中对输出类别有较大影响的像素点集合. 如图 8 所示, 原始图像被分类器正确分类为博美犬, 图 8(b)展示了对原始图像在不同位置进行遮挡后正确类别输出的概率图, 可以发现在红色区域对图像进行遮挡时正确类别的输出概率变化不大, 当对博美犬所在区域进行遮挡时尤其是面部进行遮挡时正确类别概率会急剧下降.



(a) Input image (b) Classifier, probability of correct class

图 8 (a) 对于给定的测试样本, 使用灰色方块遮挡图像的不同区域, (b) 将扰动后图像输入网络得到的正确类别概率[38]

Zhou 等人的工作[80]中采用了像素删除和遮挡的策略研究深度模型的对象检测能力. 给定一个被网络正确分类的图像, 文献 [80] 中采用简化图像 (simplifying images) 的策略, 迭代删除图像中的像素区域尽可能少地保留视觉信息, 同时保证仍然对正确类别有较高的预测分数, 将最终得到的简化图像命名为最小图像表示, 它突出显示导致高分类输出的特征. 如图 9 所示, 在分类器判别为卧室的情况下, 最小图像表示通常包含床的区域. 同样的,

Zhou 等人[80]采用了类似[38]中的遮挡策略, 可视化出了一个神经元所对应的真实感受野区域, 发现随着层数的增加神经元的感受野逐渐扩大, 但是真实有效的感受野比理论的感受野范围小很多.

图 9 测试样本及其对应的简化图像表示[80]

文献 [81,82] 中将扰动和贝叶斯概率估计相结合来评价扰动特征对输出的影响, 从而得到对输出有重要贡献的图像区域. 基于扰动的解释性算法一般都是模型无关的, 它们可以直接估计特征的边际效应, 但是随着扰动和测试的特征数量的增加, 它们的计算效率往往会非常慢.

4.4.3 局部近似模型

局部近似的可解释性模型主要思想是在局部范围内, 对深度模型的行为使用简单的解释性好的模型进行近似, LIME 算法和 OpenBox 算法[84]是该类算法的代表. OpenBox 算法主要针对以分段线性函数 (如 MaxOut [92]、ReLU 以及 ReLU 的一些变体[93−95]) 作为激活函数的神经网络, 该算法可以将这类神经网络等价为一系列的分段线性函数.

以 ReLU 作为激活函数的神经网络是分段线性的网络 (PLNN)[96], 因此 OpenBox 算法[84]的主要思想就是将神经网络所拟合的函数进行分段表示. 对于输入到网络的单个样本, 将所有神经元的激活状态用 0 和 1 进行表示, 可以得到关于网络所有神经元是否激活的 0、1 序列, 该序列称为输入样本的激活配置. 根据样本的激活配置可以将所有的样本进行划分, 将具有相同激活配置的输入样本作为一个集合 P, 集合 P 的边界是一个凸多面体. 因此, 属于同一集合 P 的所有样本具有相同的激活配置和线性映射函数. 实际上, PLNN 网络对输入空间进行区域划分并在每个区域内拟合数据和标签之间的关系, 因此部分工作[97]将线性区域的数目作为衡量一个网络容量和表达能力的指标. OpenBox 算法[84]是对分段线性网络准确一致的解释, 通过提取每个区域上线性映射函数的梯度以及区域边界的特征可以为深度模型的决策提供依据. 该算法具有一定的局限性, 首先只能针对 PLNN 网络, 其次一个网络的激活配置理论上有很多种组合这会带来计算时间上的消耗.

局部近似模型通过使用解释性好的简单模型来对深度模型的行为进行局部近似, 然而这类算法无法窥探模型内部的知识表达, 其次如何保证近似模型和原始模型在决策时行为的一致性也是值得我们思考和研究的.

5 可解释性算法性能评估及总结

5.1 可解释性算法的性能评估

目前关于深度模型的解释性算法越来越多, 不同的解释性算法在同一模型上的表现也不尽相同, 因此需要评价指标来衡量解释性算法的好坏. 评估解释性算法的质量具有一定的挑战性, 因为我们不清楚哪些特征对模型来说是重要的, 导致我们无法提供相关的监督信息来对解释性算法进行统一的评估. 因此, 一些性能评估算法选择在不同的角度下对解释性算法进行比较, 常见的评价指标如: 敏感性[65]、忠实性/准确性[21,98]、一致性/鲁棒性/稳定性[21]等.

● 可解释性算法的敏感性[65]关注解释性算法是否对模型的参数以及类别标签敏感. 与模型相关的解释性算法需要访问模型参数, 那么我们应该期望这些解释性算法在不同网络参数下的解释结果应该有差异. 同样的, 一般来说解释性算法依赖于数据标签, 那么数据标签的变化也应该会导致不同的解释性结果.

Adebayo 等人[65]提供了模型参数随机化试验和数据标签随机化试验对解释性算法的敏感性进行分析, 同时利用一些相似性指标来衡量不同实验条件下解释结果的相似性, 如 Spearman 相关系数、结构相似性指数 (SSIM) 等. 文献 [65] 中的实验发现, 导向反向传播算法[64]和导向 Grad-CAM 算法[25]对网络参数和数据标签不敏感.

● 可解释性算法的忠实性/准确性[21,98]关注的是可解释性算法检测到的模型决策特征, 对当前任务来说, 是否是模型真正依赖的特征, 该评价指标主要是基于扰动的思想来实施的. 部分解释性算法会生成和输入图像大小一致的热力图, 热力图中的数值大小反映着该像素特征的重要程度. 我们将热力图中的值从高到低排列, 按照排序依次从原始图像的对应位置移除相关像素, 并将扰动后的图像 x' 输入网络中计算当前图像的预测结果 $f(x')$.

一般情况下, 会得到一条随着移除像素点的数目的增加而减小的 $f(x')$ 曲线. 通过计算曲线和坐标轴围成的区域的面积 (AUC) 可以比较各种解释性算法的质量. 较低的 AUC 分数表明预测输出函数值 f 急剧下降, 说明解释性算法所发现的特征确实是被模型认定为与任务相关的特征. 然而, 这不是一个理想的指标, 因为使用黑色像素替换图像像素会导致在图像中出现伪像, 这也可能导致模型输出的变化.

● 可解释性算法的稳定性/一致性[21,99]要求输入相似的样本具有相似的解释结果. 对于输入图像 x 附近领域内的图像 x' 且模型对它们有相同的预测结果, 那么一个好的解释性算法生成的解释性结果 $R(x)$ 和 $R(x')$ 应该也是相似的. 因此, 定义如下的度量:

$$\max_{x\neq x'} \frac{\|R(x)-R(x')\|_1}{\|x-x'\|_2} \tag{27}$$

这个指标衡量输入相似的样本在解释性输出结果上的差异程度, 如果解释性输出差异较大那么解释性算法的一致性 (鲁棒性/稳定性) 较差, 则可设计攻击算法对输入图像进行微小扰动从而生成矛盾的解释结果[72], 这会使得用户怀疑解释性算法的可靠性.

相较于可解释性算法来说, 当前对解释性算法的性能评估研究较少, 而且部分评价指标也存在一定的缺陷, 如忠实性的评价指标需要扰动输入图像, 删除像素后的样本和原始样本的分布是不一致的, 这会导致忠实性的评估存在一定的误差. 评估解释性算法的解释质量是非常必要的, 只有这样我们才能有针对性地对解释性算法以及深度模型进行改进. 因此, 设计合理的解释性算法评价指标仍是当前的一个挑战问题.

5.2 可解释性算法总结

近年来深度学习的可解释性研究越来越多, 本文从全局和局部的视角下对图像分类的可解释性算法进行梳理, 基于解释粒度对相关的可解释性算法进行了进一步的划分和总结. 在每一小结详细地介绍了相关算法的原理、机制以及他们的优缺点, 通过模型级可解释性算法我们可以获得模型的整体知识表达, 通过神经元级的可解释性研究可以告诉用户每个神经元对应的激活语义特征, 通过局部可解释性算法我们可以清楚地了解到模型在决策过程中所依赖的特征. 总体来说, 可解释性研究搭建了用户和模型之间的桥梁, 让用户深入了解模型内部的知识和逻辑同时提升了用户对模型的信赖. 其次, 可解释性算法也可以被应用到不同的实际问题中辅助解决当前任务, 如弱监督定位[25]、数据集偏见检测[24,25]、小样本问题[100]等.

我们将本文所涉及的可解释性算法以及它们的性质总结在表 3 中, 主要在全局/局部、解释粒度以及模型相关/无关的视角下对所有算法进行划分, 并使用第 5.1 节中介绍的评价指标对不同的解释性算法进行比较. 其中敏感性反映的是解释模型是否对网络参数和数据标签敏感; 忠实性/准确性主要反映算法定位到的关键特征是否是模型所依赖的特征; 稳定性/一致性体现的是解释模型在相似输入上是否具有相似的解释结果, 稳定性差的可解释性算法容易受到攻击从而导致解释性算方法失效. 总的来说, 深度模型的可解释研究目前处于初级阶段, 该方向仍有许多科学问题需要我们深入探讨和研究.

6 挑战与展望

虽然本文仅对图像分类任务的可解释性研究算法进行了梳理, 但是不同领域的可解释性研究具有一定的相似性, 如多数可解释性算法都是从深度模型本身、神经元以及模型决策等方面对模型的行为进行理解, 在解释性算法的设计上可以通过重要性分析、可视化分析、注意力机制等来对深度模型进行剖析[101−105].

表 3 可解释性算法及其性质总结

一级分类	二级分类	算法	模型相关(Y)/无关(N)	敏感性	忠实性/准确性	稳定性/一致性
全局可解释性算法	模型级	Activation Maximization[37]	Y	■	☆☆	▬
		Network compression[54,55]	Y	—	☆	—
		Knowledge distillation[56,57]	Y	—	☆	—
	神经元级	Activation Maximization[16,51,62]	Y	■	☆☆	▬
		Network Dissection[58−61]	Y	—	☆☆	—
局部可解释性算法	像素级特征	Saliency Map[37]	Y	▌	☆	▬
		SmoothGrad[63]	Y	▌	☆☆	▬
		DeconvNet Visualization[38]	Y	▌	☆☆	▬
		GBP[64]	Y	▌	☆☆	▬
		Input⊙Gradients[65]	Y	▌	☆☆	▬
		Integrated Gradients (IG) [67]	Y	▌	☆☆	▬
		LRP[41]	Y	▌	☆☆	▬
		DeepLIFT[68]	Y	■	☆☆	■
	概念级特征	LIME	N	▌	☆☆	▬
		CAV [43−45]	Y	▌	☆☆	▬
	图像级特征	Network Inversion[46,47]	Y	▌	☆☆	■
		MMD-critic[48]	Y	▌	☆☆	■
		Influence Functions[49]	Y	▌	☆☆	■
		Based on the Representation Theorem [78]	Y	▌	☆☆	■
	其他算法	CAM[79]	Y	▌	☆☆	▬
		Grad-CAM[25]	Y	▌	☆☆	▬
		Simplifying images[80]	N	▌	☆☆	■
		OpenBox[84]	Y	■	☆☆☆	■

注: —代表不可知; ▌, ▌, ■分别代表解释算法对参数和标签的敏感性较差, 一般, 良好; ☆, ☆☆和☆☆☆分别表示可解释性算法解释结果准确性为一般, 良好, 准确; ▬, ▬, ■分别代表解释性算法的一致性较差, 一般, 良好

目前可解释性研究已经有相当多的研究基础, 然而该领域还是存在很多的问题和挑战, 其中如何设计一个忠实于模型的、准确可靠的可解释性算法, 以及如何设计一个合理的可解释性算法评价指标是当前面临的两个挑战. 其次, 可解释性算法除了可以对深度模型进行解释外, 它也可以作为辅助工具对模型进行测试和调试, 这说明可解释性算法有巨大的潜在价值需要我们深入挖掘. 此外, 理论基础研究的缺乏也导致深度模型的解释性、泛化性等问题, 因此深度模型的理论基础研究也是非常必要的. 基于目前对深度学习模型的认识和理解, 本文认为将来深度学习的可解释性研究可以从以下几个方面入手.

(1) 深度模型的可解释性算法研究: 可解释性研究的难点在于深度模型的参数空间庞大, 其次模型的特征空间存在组合爆炸和语义纠缠的现象. 因此在设计解释性算法的时候, 既要关注模型的整体框架, 又要降低特征空间的维度同时需要对语义特征进行解纠缠. 一个可信赖的可解释性算法要忠实于模型本身, 即算法的解释结果要和深度模型的真实行为保持一致, 否则解释结果是没有意义的; 其次, 一个可靠的解释性算法应该具有稳定性, 当前的部分解释性算法是易受攻击的, 这会导致解释性算法在相似的输入样本上具有不同的解释结果. 此外, 当前的可解释性研究大多停留在感知的层面上, 将深度模型的特征表示空间映射到与人类感知相吻合的语义空间来进行解释, 这是一种相对较弱的解释结果. 如何利用外部知识将深度模型的逻辑推理过程完全透明化, 甚至利用知识指导深度模型的推理过程使得建模过程具备解释性是将来的一个研究方向. 深度学习是一种统计上的关联学习, 将深度模型和因果学习相结合也是研究深度学习可解释性的一种思路, 可以通过因果学习来挖掘数据中的本质模式从而提升模型的解释性.

(2) 可解释性算法的评估指标研究: 目前可解释性算法的性能评估研究比较缺乏, 主要是因为我们对深度模型

的认知存在局限性. 由于深度模型过于复杂, 我们不清楚模型从数据中提取了哪些特征, 因此评价指标的设计缺乏监督信息; 其次, 许多评价指标的设计引入了强烈的偏见, 人类倾向于选择接近自己期望的解释方法, 代价是惩罚那些可能反映模型真实行为的算法; 另外, 现在的评价指标很难将模型的错误和解释性算法的错误区分开来, 如一致性评价指标, 我们知道神经网络对微小的形变 (连续平移、缩放、不同视角、对抗噪声等) 不具有不变性[106], 这和我们希望解释性算法的解释结果具有一致性是有矛盾的. 因此, 我们认为在开发评估指标时, 要充分考虑模型、数据、解释性算法以及人类认知等多方面的因素; 评估指标的设计要客观合理, 能够反映解释性算法的忠实性、一致性以及不同解释性算法的差异性.

(3) 利用解释性算法探索和改进深度模型存在的问题: 可解释性算法的直接应用就是对深度模型进行解释, 当前可解释性算法已经应用到推荐系统、医疗、金融等领域. 可解释性算法也可以用来对网络预测错误的样本进行解释和纠正[78]以及监督网络训练是否收敛等[38]. 除此之外, 我们可以使用可解释性算法对深度模型进行测试和调试. 一方面解释性算法可以作为探针来测试模型内部存在的问题, 如公平性问题、数据偏见[24,25]、虚假相关性[24,75]等. 另一方面, 当使用解释性算法诊断出深度模型存在的问题后, 我们可以采取相应的措施对模型进行改进. 如, Li 等人[24]利用 CAM 算法指导模型的训练过程, 最终不仅可以实现较高的分类性能以及弱监督下的图像分割, 而且利用解释性引导网络关注对象区域从而避免模型依赖虚假相关性特征. 此外, 利用可解释性算法来改进模型甚至建立具有自解释性的深度模型是将来的一个研究方向[98,107].

(4) 深度模型的理论基础研究: 虽然我们可以将深度模型中每个神经元的激活情况和他们所感兴趣的特征都展示出来, 但是要真正理解深度模型我们还需要在理论方面展开研究. 一个单隐含层的神经网络可以以任意精度逼近任意连续函数[108,109], 相比于浅层网络, 深层网络具有更强的表达能力和优异的性能[110]. 深度模型为什么具有良好的泛化性是值得我们进一步研究的, 目前部分工作也提出了他们的观点. Choromanska 等人的工作[111]和 Wu 等人[112]的工作证明了随着网络规模的增大, 神经网络的损失函数具有良好泛化性的极小值点的数目要比泛化性差的极小值点要多很多, 从而导致随机初始化训练的网络总能收敛到好的极小值点. 部分文献[113−117]表明在一定假设条件下深度模型的局部极小值接近于全局最小值, 文献 [118−120] 实验发现并证明深度模型的泛化能力和优化算法找到的最终收敛点的锐度有关, 平滑的局部极小值点有助于模型获得好的泛化性. 神经网络的训练过程可以认为是一个离散的动力系统, 因此可以通过微分方程来理解神经网络, 同时也可以利用神经网络来求解微分方程问题[121−123]. Xu 等人[124,125]从频域的视角下解释了神经网络的训练过程, 发现神经网络在训练的过程中倾向于先拟合数据和标签间的低频映射关系, 随着训练的进行网络会逐渐拟合映射中的高频分量. 以 ReLU 作为激活函数的神经网络是分段线性的, 每个神经元的作用是在对输入空间进行分割, 因此可以将输入空间所划分的区域的数目作为神经网络复杂度和容量的度量[84,97]. 一些工作证明了具有独立同分布参数的无限宽深度网络是一个高斯过程, 而且高斯过程是深度模型性能的极限[126−128]. 文献 [129−132] 从流形的视角来理解深度生成学习, 并将最优传输理论应用到生成模型中来解决生成模型存在的模式坍塌等问题. 总的来说, 理论基础还是深度学习的一个巨大软肋, 这也导致模型的解释性差等其他问题.

7 结束语

深度模型的可解释性研究是必要的也是非常有意义的, 每年有很多优秀的相关工作在计算机领域的顶级会议和期刊上发表. 本文从全局解释性算法和局部解释性算法的角度对目前的可解释性算法进行了归类和总结, 并介绍了可解释性算法的评价指标及当前研究的挑战与未来的研究方向. 总的来说, 深度学习的可解释性研究才刚刚起步, 一方面我们仍需要进一步研究深度模型的可解释性算法和基础理论, 另一方面可解释性算法在其他领域也具有巨大的潜在价值需要我们深入挖掘.

References:

[1] Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: The MIT Press, 2016.

[2] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]

[3] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[4] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3065386]

[5] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

[6] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015.

[7] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 3104–3112.

[8] Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. In: Proc. of the 2013 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013. 6645–6649. [doi: 10.1109/ICASSP.2013.6638947]

[9] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 2015, 33(8): 831–838. [doi: 10.1038/nbt.3300]

[10] Sturm I, Lapuschkin S, Samek W, Müller KR. Interpretable deep neural networks for single-trial EEG classification. Journal of Neuroscience Methods, 2016, 274: 141–145. [doi: 10.1016/j.jneumeth.2016.10.008]

[11] Tjoa E, Guan CT. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(11): 4793–4813. [doi: 10.1109/TNNLS.2020.3027314]

[12] Li L, Qin LX, Xu ZG, Yin YB, Wang X, Kong B, Bai JJ, Lu Y, Fang ZH, Song Q, Cao KL, Liu DL, Wang GS, Xu QZ, Fang XS, Zhang SQ, Xia J, Xia J. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology, 2020: 200905. [doi: 10.1148/radiol.2020200905]

[13] Wu F, Liao BB, Han YH. Interpretability for deep learning. 2019, 26(1): 39–46 (in Chinese with English abstract). [doi: 10.12132/ISSN.1673-5048.2018.0065]

[14] Cheng KY, Wang N, Shi WX, Zhan YZ. Research advances in the interpretability of deep learning. Journal of Computer Research and Development, 2020, 57(6): 1208–1217 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2020.20190485]

[15] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 427–436. [doi: 10.1109/CVPR.2015.7298640]

[16] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff, 2014.

[17] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82–115. [doi: 10.1016/j.inffus.2019.12.012]

[18] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608, 2017.

[19] Molnar C. Interpretable Machine Learning. Lulu.com, 2020.

[20] Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Interpretable machine learning: Definitions, methods, and applications. arXiv: 1901.04592, 2019.

[21] Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 2018, 73: 1–15. [doi: 10.1016/j.dsp.2017.10.011]

[22] Fan FL, Xiong JJ, Li MZ, Wang G. On interpretability of artificial neural networks: A survey. arXiv:2001.02522, 2020.

[23] Tsipras D, Santurkar S, Engstrom L, Turner A, Madry A. Robustness may be at odds with accuracy. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[24] Li KP, Wu ZY, Peng KC, Ernst J, Fu Y. Tell me where to look: Guided attention inference network. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9215–9223. [doi: 10.1109/CVPR.2018.00960]

[25] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 618–626. [doi: 10.1109/ICCV.2017.74]

[26] Rakhlin A, Caponnetto A. Stability of *k*-means clustering. In: Proc. of the 19th Int'l Conf. on Neural Information Processing Systems. Vancouver: MIT Press, 2006. 1121–1128.

[27] Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 2010, 31(8): 651–666. [doi: 10.1016/j.patrec.2009.09.011]

[28] Wang T, Rudin C, Doshi-Velez F, Liu YM, Klampfl E, Macneille P. A Bayesian framework for learning rule sets for interpretable classification. The Journal of Machine Learning Research, 2017, 18(1): 2357–2393.

[29] Letham B, Rudin C, Mccormick TH, Madigan D. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics, 2015, 9(3): 1350–1371. [doi: 10.1214/15-AOAS848]

[30] Loh WY. Classification and regression trees. WIREs Data Mining and Knowledge Discovery, 2011, 1(1): 14–23. [doi: 10.1002/widm.8]

[31] Hastie T, Tibshirani R. Generalized additive models: Some applications. Journal of the American Statistical Association, 1987, 82(398): 371–386. [doi: 10.2307/2289439]

[32] Nelder JA, Wedderburn RWM. Generalized linear models. Journal of the Royal Statistical Society: Series A, 1972, 135(3): 370–384. [doi: 10.2307/2344614]

[33] Ji SL, Li JF, Du TY, Li B. Survey on techniques, applications and security of machine learning interpretability. Journal of Computer Research and Development, 2019, 56(10): 2071 –2096 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2019.20190540]

[34] Chen KR, Meng XF. Interpretation and understanding in machine learning. Journal of Computer Research and Development, 2020, 57(9): 1971–1986 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2020.20190456]

[35] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2019, 267: 1–38. [doi: 10.1016/j.artint.2018.07.007]

[36] Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Mądry A. Adversarial examples are not bugs, they are features. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 12.

[37] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: 2nd Int'l Conf. on Learning Representations. Banff, 2014.

[38] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 818–833. [doi: 10.1007/978-3-319-10590-1_53]

[39] Srinivas S, Fleuret F. Full-gradient representation for neural network visualization. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 4126–4135.

[40] Ancona M, Ceolini E, Öztireli C, Gross M. Gradient-based attribution methods. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer, 2019. 169–191. [doi: 10.1007/978-3-030-28954-6_9]

[41] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, 2015, 10(7): e0130140. [doi: 10.1371/journal.pone.0130140]

[42] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: Association for Computing Machinery, 2016. 1135–1144. [doi: 10.1145/2939672.2939778]

[43] Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, Sayres R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholmsmassan: PMLR, 2018. 2668–2677.

[44] Ghorbani A, Wexler J, Kim B. Automating interpretability: Discovering and testing visual concepts learned by neural networks. arXiv: 1902.03129, 2019.

[45] Ghorbani A, Wexler J, Zou JY, Kim B. Towards automatic concept-based explanations. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 9273–9282.

[46] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5188–5196. [doi: 10.1109/CVPR.2015.7299155]

[47] Dosovitskiy A, Brox T. Inverting visual representations with convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4829–4837. [doi: 10.1109/CVPR.2016.522]

[48] Kim B, Khanna R, Koyejo O. Examples are not enough, learn to criticize! Criticism for interpretability. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 2288–2296.

[49] Koh PW, Liang P. Understanding black-box predictions via influence functions. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: PMLR, 2017. 1885–1894.

[50] Erhan D, Bengio Y, Courville A, Vincent P. Visualizing higher-layer features of a deep network. Technical Report 1341, Montreal: University of Montreal, 2009.

[51] Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. arXiv:1506.06579, 2015.

[52] Nguyen A, Yosinski J, Clune J. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. arXiv:1602.03616, 2016.

[53] Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 3395–3403.

[54] Abbasi-Asl R, Yu B. Interpreting convolutional neural networks through compression. arXiv:1711.02329, 2017.

[55] Li YC, Lin SH, Zhang BC, Liu JZ, Doermann D, Wu YJ, Huang FY, Ji RR. Exploiting kernel sparsity and entropy for interpretable CNN compression. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2795–2804. [doi: 10.1109/CVPR.2019.00291]

[56] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

[57] Frosst N, Hinton GE. Distilling a neural network into a soft decision tree. In: Proc. of the 1st Int'l Workshop on Comprehensibility and Explanation in AI and ML. Bari: CEUR-WS.org, 2017.

[58] Bau D, Zhou BL, Khosla A, Oliva A, Torralba A. Network dissection: Quantifying interpretability of deep visual representations. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3319–3327. [doi: 10.1109/CVPR.2017.354]

[59] Zhou BL, Bau D, Oliva A, Torralba A. Interpreting deep visual representations via network dissection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2131–2145. [doi: 10.1109/TPAMI.2018.2858759]

[60] Zhou BL, Bau D, Oliva A, Torralba A. Comparing the interpretability of deep networks via network dissection. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, eds. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer, 2019. 243–252. [doi: 10.1007/978-3-030-28954-6_12]

[61] Bau D, Zhu JY, Strobelt H, Zhou BL, Tenenbaum JB, Freeman WT, Torralba A. Gan dissection: Visualizing and understanding generative adversarial networks. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[62] Chatonsky G. Deep dream (the network's dream). SubStance, 2016, 45(2): 61–77. [doi: 10.3368/ss.45.2.61]

[63] Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: Removing noise by adding noise. arXiv:1706.03825, 2017.

[64] Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA. Striving for simplicity: The all convolutional net. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015.

[65] Tomsett R, Harborne D, Chakraborty S, Gurram P, Preece A. Sanity checks for saliency metrics. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 6021–6029. [doi: 10.1609/aaai.v34i04.6064]

[66] Kim B, Seo J, Jeon S, Koo J, Choe J, Jeon T. Why are saliency maps noisy? Cause of and solution to noisy saliency maps. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision Workshop (ICCVW). Seoul: IEEE, 2019. 4149–4157. [doi: 10.1109/ICCVW.2019.00510]

[67] Sundararajan M, Taly A, Yan QQ. Axiomatic attribution for deep networks. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: PMLR, 2017. 3319–3328.

[68] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: PMLR, 2017. 3145–3153.

[69] Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition, 2017, 65: 211–222. [doi: 10.1016/j.patcog.2016.11.008]

[70] Omeiza D, Speakman S, Cintas C, Weldermariam K. Smooth grad-CAM++: An enhanced inference level visualization technique for deep convolutional neural network models. arXiv:1908.01224, 2019.

[71] Nie WL, Zhang Y, Patel A. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholmsmassan: PMLR, 2018. 3809–3818.

[72] Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 3681–3688. [doi: 10.1609/aaai.v33i01.33013681]

[73] Dombrowski AK, Alber M, Anders CJ, Ackermann M, Müller KR, Kessel P. Explanations can be manipulated and geometry is to blame. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 13567–13578.

[74] Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[75] Zhang YJ, Song KY, Sun YM, Tan S, Udell M. "Why should you trust my explanation?" Understanding uncertainty in LIME explanations. In: Proc. of the Int'l Conf. on Machine Learning AI for Social Good Workshop. Long Beach, 2019.

[76] Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI Press, 2018. 1527–1535.

[77] Qin ZW, Yu FX, Liu CC, Chen X. How convolutional neural networks see the world-A survey of convolutional neural network visualization methods. American Institute of Mathematical Sciences, 2018, 1(2): 149–180. [doi: 10.3934/mfc.2018008]

[78] Yeh CK, Kim JS, Yen IEH, Ravikumar P. Representer point selection for explaining deep neural networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 9311–9321.

[79] Zhou BL, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2921–2929. [doi: 10.1109/CVPR.2016.319]

[80] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Object detectors emerge in deep scene CNNs. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015.

[81] Zintgraf LM, Cohen TS, Adel T, Welling M. Visualizing deep neural network decisions: Prediction difference analysis. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[82] Zintgraf LM, Cohen TS, Welling M. A new method to visualize deep neural networks. arXiv:1603.02518, 2016.

[83] Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 3449–3457. [doi: 10.1109/ICCV.2017.371]

[84] Chu LY, Hu X, Hu JH, Wang LJ, Pei J. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: Association for Computing Machinery, 2018. 1244–1253. [doi: 10.1145/3219819.3220063]

[85] Lin M, Chen Q, Yan SC. Network in network. arXiv: 1312.4400, 2013.

[86] Subramanya A, Pillai V, Pirsiavash H. Fooling network interpretation in image classification. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 2020–2029. [doi: 10.1109/ICCV.2019.00211]

[87] Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: Proc. of the 2018 IEEE Winter Conf. on Applications of Computer Vision (WACV). Lake Tahoe: IEEE, 2018. 839–847. [doi: 10.1109/WACV.2018.00097]

[88] Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models. In: British Machine Vision Conf. 2018. Newcastle: BMVA Press, 2018.

[89] Wang HF, Naidu R, Michael J, Kundu SS. SS-CAM: Smoothed score-CAM for sharper visual feature localization. arXiv:2006.14255, 2020.

[90] Wang HF, Wang ZF, Du MN, Yang F, Zhang ZJ, Ding SR, Mardziel P, Hu X. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020. 111–119. [doi: 10.1109/CVPRW50498.2020.00020]

[91] Wang Y, Hu X, Su H. Learning attributions grounded in existing facts for robust visual explanation. XAI, 2018: 178.

[92] Goodfellow I, Warde-Farley D, Mirza M, Courville A, Bengio Y. Maxout networks. In: Proc. of the 30th Int'l Conf. on Machine Learning. Atlanta: PMLR, 2013. 1319–1327.

[93] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proc. of the 14th Int'l Conf. on Artificial Intelligence and Statistics. Fort Lauderdale: JMLR, 2011. 315–323.

[94] He KM, Zhang XY, Ren SQ, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imageNet classification. In: Proc. of the IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1026–1034. [doi: 10.1109/ICCV.2015.123]

[95] Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proc. of the 27th Int'l Conf. on Machine Learning. Haifa: Omni Press, 2010.

[96] Harvey N, Liaw C, Mehrabian A. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In: Proc. of the 2017 Conf. on Learning Theory. Amsterdam: PMLR, 2017. 1064–1068.

[97] Montúfar GF, Pascanu R, Cho K, Bengio Y. On the number of linear regions of deep neural networks. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2924–2932.

[98] Alvarez-Melis D, Jaakkola TS. Towards robust interpretability with self-explaining neural networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 7786–7795.

[99] Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. arXiv:1806.08049, 2018.

[100] Sun JM, Lapuschkin S, Samek W, Zhao YQ, Cheung NM, Binder A. Explanation-guided training for cross-domain few-shot classification. In: Proc. of the 25th Int'l Conf. on Pattern Recognition (ICPR). Milan: IEEE, 2021. 7609–7616. [doi: 10.1109/ICPR48806.2021.9412941]

[101] Ding SY, Xu HN, Koehn P. Saliency-driven word alignment interpretation for neural machine translation. In: Proc. of the 4th Conf. on Machine Translation. Florence: Association for Computational Linguistics, 2019. 1–12. [doi: 10.18653/v1/W19-5201]

[102] Jain S, Wallace BC. Attention is not explanation. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 3543–3556. [doi: 10.18653/v1/N19-1357]

[103] Wiegreffe S, Pinter Y. Attention is not not explanation. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 11–20. [doi: 10.18653/v1/D19-1002]

[104] Li JW, Chen XL, Hovy E, Jurafsky D. Visualizing and understanding neural models in NLP. In: Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 681–691. [doi: 10.18653/v1/N16-1082]

[105] Li JW, Monroe W, Jurafsky D. Understanding neural networks through representation erasure. arXiv:1612.08220, 2016.

[106] Azulay A, Weiss Y. Why do deep convolutional networks generalize so poorly to small image transformations? Journal of Machine Learning Research, 2019, 20: 1–25.

[107] Chen CF, Li O, Tao CF, Barnett AJ, Su J, Rudin C. This looks like that: Deep learning for interpretable image recognition. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 801Pages.

[108] Cybenko G. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems, 1989, 2(4): 303–314. [doi: 10.1007/BF02551274]

[109] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks, 1989, 2(5): 359–366. [doi: 10.1016/0893-6080(89)90020-8]

[110] Lu Z, Pu HM, Wang FC, Hu ZQ, Wang LW. The expressive power of neural networks: A view from the width. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6232–6240.

[111] Choromanska A, Henaff M, Mathieu M, Arous GB, LeCun Y. The loss surfaces of multilayer networks. In: Proc. of the 8th Int'l Conf. on Artificial Intelligence and Statistics. San Diego: PMLR, 2015. 192–204.

[112] Wu L, Zhu ZX, E WN. Towards understanding generalization of deep learning: Perspective of loss landscapes. arXiv:1706.10239, 2017.

[113] Laurent T, Brecht J. Deep linear networks with arbitrary loss: All local minima are global. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 2902–2907.

[114] Kawaguchi K, Huang JY, Kaelbling LP. Every local minimum value is the global minimum value of induced model in nonconvex machine learning. Neural Computation, 2019, 31(12): 2293–2323. [doi: 10.1162/neco_a_01234]

[115] Dauphin YN, Pascanu R, Gulcehre C, Cho K, Ganguli S, Bengio Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2933–2941.

[116] Haeffele BD, Vidal R. Global optimality in neural network training. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4390–4398. [doi: 10.1109/CVPR.2017.467]

[117] Soudry D, Carmon Y. No bad local minima: Data independent training error guarantees for multilayer neural networks. arXiv: 1605.08361, 2016.

[118] Yao ZW, Gholami A, Keutzer K, Mahoney MW. Hessian-based analysis of large batch training and robustness to adversaries. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 4954–4964.

[119] Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP. On large-batch training for deep learning: Generalization gap and sharp minima. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[120] Santurkar S, Tsipras D, Ilyas A, Madry A. How does batch normalization help optimization? In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 2488–2498.

[121] E WN, Yu B. The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. Communications in Mathematics and Statistics, 2018, 6(1): 1–12. [doi: 10.1007/s40304-018-0127-z]

[122] Wang B, Li Z, Shi ZQ, Luo XY, Zhu W, Osher SJ. Deep neural nets with interpolating function as output activation. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 751–761.

[123] Chen RTO, Rubanova Y, Bettencourt J, Duvenaud DK. Neural ordinary differential equations. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 6572–6583.

[124] Xu ZQJ, Zhang YY, Luo T, Xiao YY, Ma Z. Frequency principle: Fourier analysis sheds light on deep neural networks. arXiv: 1901.06523, 2019.

[125] Xu ZQJ, Zhang YY, Xiao YY. Training behavior of deep neural network in frequency domain. In: Proc. of the 26th Int'l Conf. on Neural Information Processing. Sydney: Springer, 2019. 264–274. [doi: 10.1007/978-3-030-36708-4_22]

[126] Matthews AGDG, Hron J, Rowland M, Turner RE, Ghahramani Z. Gaussian process behaviour in wide deep neural networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[127] Lee J, Bahri Y, Novak R, Schoenholz SS, Pennington J, Sohl-Dickstein J. Deep neural networks as gaussian processes. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[128] Lee J, Xiao LC, Schoenholz SS, Bahri Y, Novak R, Sohl-Dickstein J, Pennington J. Wide neural networks of any depth evolve as linear models under gradient descent. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 769.

[129] Lei N, An DS, Guo Y, Su KH, Liu SX, Luo ZX, Yau ST, Gu XF. A geometric understanding of deep learning. Engineering, 2020, 6(3): 361–374. [doi: 10.1016/j.eng.2019.09.010]

[130] Lei N, Su KH, Cui L, Yau ST, Gu XD. A geometric view of optimal transportation and generative model. Computer Aided Geometric Design, 2019, 68: 1–21. [doi: 10.1016/j.cagd.2018.10.005]

[131] An DS, Guo Y, Lei N, Luo ZX, Yau ST, Gu XF. AE-OT: A new generative model based on extended semi-discrete optimal transport. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[132] An DS, Guo Y, Zhang M, Qi X, Lei N, Gu XF. AE-OT-GAN: Training GANs from data specific latent distribution. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 548–564. [doi: 10.1007/978-3-030-58574-7_33]

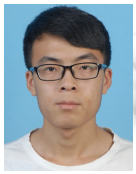
附中文参考文献:

[13] 吴飞, 廖彬兵, 韩亚洪. 深度学习的可解释性. 航空兵器, 2019, 26(1): 39–46. [doi: 10.12132/ISSN.1673-5048.2018.0065] [doi: 10.12132/ISSN.1673-5048.2018.0065]

[14] 成科扬, 王宁, 师文喜, 詹永照. 深度学习可解释性研究进展. 计算机研究与发展, 2020, 57(6): 1208–1217. [doi: 10.7544/issn1000-1239.2020.20190485]

[33] 纪守领, 李进锋, 杜天宇, 李博. 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展, 2019, 56(10): 2071–2096. [doi: 10.7544/issn1000-1239.2019.20190540]

[34] 陈珂锐, 孟小峰. 机器学习的可解释性. 计算机研究与发展, 2020, 57(9): 1971–1986. [doi: 10.7544/issn1000-1239.2020.20190456]



杨朋波(1993－), 男, 博士生, 主要研究领域为计算机视觉, 对抗鲁棒性研究, 可解释性研究.

冯耀功(1992－), 男, 硕士, 主要研究领域为计算机视觉, 零样本学习.

桑基韬(1985－), 男, 博士, 教授, CCF 高级会员, 主要研究领域为多媒体计算, 网络数据挖掘, 可信赖机器学习.

于剑(1969－), 男, 博士, 教授, CCF 会士, 主要研究领域为人工智能, 机器学习.

张彪(1995－), 男, 硕士, 主要研究领域为计算机视觉, 模型压缩.