机器学习模型可解释性方法、应用与安全研究综述

纪守领[1] 李进锋[1] 杜天宇[1] 李博[2]

[1]（浙江大学网络空间安全研究中心 浙江大学计算机科学与技术学院 杭州 310027）

[2]（伊利诺伊大学香槟分校计算机科学学院 厄巴纳香槟 61822）

（lijinfeng0713@zju.edu.cn）

A Survey on Techniques, Applications and Security of Machine Learning Interpretability

Ji Shouling[1], Li Jinfeng[1], Du Tianyu[1], Li Bo[2]

[1]（*Institute of Cyberspace Research and College of Computer Science and Technology*, *Zhejiang University*, *Hangzhou* 310027）

[2]（*Department of Computer Science, University of Illinois at Urbana–Champaign*, *Urbana-Champaign* 61822）

Abstract While machine learning has achieved great success in various domains, the lack of interpretability has limited its widespread applications in real-world tasks, especially security-critical tasks. To overcome this crucial weakness, intensive research on improving the interpretability of machine learning models has emerged, and a plethora of interpretation methods have been proposed to help end users understand its inner working mechanism. However, the research on model interpretation is still in its infancy, and there are a large amount of scientific issues to be resolved. Furthermore, different researchers have different perspectives on solving the interpretation problem and give different definitions for interpretability, and the proposed interpretation methods also have different emphasis. Till now, the research community still lacks a comprehensive understanding of interpretability as well as a scientific guide for the research on model interpretation. In this survey, we review the explanatory problems in machine learning, and make a systematic summary and scientific classification of the existing research works. At the same time, we discuss the potential applications of interpretation related technologies, analyze the relationship between interpretability and the security of interpretable machine learning, and discuss the current research challenges and potential future research directions, aiming at providing necessary help for future researchers to facilitate the research and application of model interpretability.

Key words machine learning; interpretability; interpretation method; interpretable machine learning; security

收稿日期：2019-05-28

基金项目：国家自然科学基金项目（61772466，U1836202）；浙江省自然科学基金杰出青年项目（LR19F020003）；浙江省科技计划项目（2017C01055）
This work was partly supported by NSFC under No. 61772466 and U1836202, the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under No. LR19F020003, and the Provincial Key Research and Development Program of Zhejiang, China under No. 2017C01055.

通信作者：纪守领（sji@zju.edu.cn）

摘要 尽管机器学习在许多领域取得了巨大的成功，但缺乏可解释性严重限制了其在现实任务尤其是安全敏感任务中的广泛应用。为了克服这一弱点，许多学者对如何提高机器学习模型可解释性进行了深入的研究，并提出了大量的解释方法以帮助用户理解模型内部的工作机制。然而，可解释性研究还处于初级阶段，依然还有大量的科学问题尚待解决。并且，不同的学者解决问题的角度不同，对可解释性赋予的含义也不同，所提出的解释方法也各有侧重。迄今为止，学术界对模型可解释性仍缺乏统一的认识，可解释性研究的体系结构尚不明确。在本综述中，我们回顾了机器学习中的可解释性问题，并对现有的研究工作进行了系统的总结和科学的归类。同时，我们讨论了可解释性相关技术的潜在应用，分析了可解释性与可解释机器学习的安全性之间的关系，并且探讨了可解释性研究当前面临的挑战和未来潜在的研究方向，以期进一步推动可解释性研究的发展和应用。

关键词 机器学习；可解释性；解释方法；可解释机器学习；安全性

0 引言

近年来，随着人工智能和机器学习的发展，机器学习相关技术在计算机视觉、自然语言处理、语音识别等多个领域取得了巨大的成功，机器学习模型也被广泛地应用到一些重要的现实任务中，如人脸识别[1-3]、自动驾驶[4]、恶意软件检测[5]和智慧医疗分析[6]等。在某些场景中，机器学习模型的表现甚至超过了人类。

尽管机器学习在许多有意义的任务中胜过人类，但由于缺乏可解释性，其表现和应用也饱受质疑[7]。对于普通用户而言机器学习模型尤其是深度神经网络（Deep Neural Networks，DNN）模型如同黑盒一般，我们给它一个输入，其反馈一个决策结果，没人能确切地知道它背后的决策依据以及它做出的决策是否可靠。而缺乏可解释性将有可能给实际任务中尤其是安全敏感任务中的许多基于 DNN 的现实应用带来严重的威胁。比如说，缺乏可解释性的自动医疗诊断模型可能给患者带来错误的治疗方案，甚至严重威胁患者的生命安全。此外，最近的研究表明，DNN 本身也面临着多种安全威胁——恶意构造的对抗性样本可以轻易让 DNN 模型分类出错[8-10]，而他们针对对抗样本的脆弱性同样也缺乏可解释性。因此，缺乏可解释性已经成为机器学习在现实任务中的进一步发展和应用的主要障碍之一。

为了提高机器学习模型的可解释性和透明性，建立用户与决策模型之间的信任关系，消除模型在实际部署应用中的潜在威胁，近年来学术界和工业界进行了广泛和深入的研究并且提出了一系列的机器学习模型可解释性方法。然而，由于不同的研究者解决问题的角度不同，因而给“可解释性”赋予的含义也不同，所提出的可解释性方法也各有侧重。因此，亟需对现有工作进行系统的整理和科学的总结、归类，以促进该领域的研究。

在本文中，我们首先详细地阐述可解释性的定义和所解决的问题。然后，我们对现有的可解释性方法进行系统的总结和归类，并讨论相关方法的局限性。接着，我们简单地介绍模型可解释性相关技术的实际应用场景，同时详细地分析可解释性中的安全问题。最后，我们讨论模型可解释性相关研究所面临的挑战以及未来可行的研究方向。

1 机器学习可解释性问题

在介绍具体的可解释问题与相应的解决方法之前，我们先简单地介绍什么是可解释性以及为什么需要可解释性。在数据挖掘和机器学习场景中，可解释性被定义为向人类解释或以呈现可理解的术语的能力[11]。从本质上讲，可解释性是人类与决策模型之间的接口，它既是决策模型的准确代理，又是

人类所可以理解的[12]。在自上而下的机器学习任务中，模型通常建立在一组统计规则和假设之上，因而可解释性至关重要，因为它是所定义的规则和假设的基石。此外，模型可解释性是验证假设是否稳健，以及所定义的规则是否完全适合任务的重要手段。与自上而下的任务不同，自下而上的机器学习通常对应于手动和繁重任务的自动化，即给定一批训练数据，通过最小化学习误差，让模型自动地学习输入数据与输出类别之间的映射关系。在自下而上的学习任务中，由于模型是自动构建的，我们不清楚其学习过程，也不清楚其工作机制，因此，可解释性旨在帮助人们理解机器学习模型是如何学习的，它从数据中学到了什么，针对每一个输入它为什么会做出如此决策以及它所做的决策是否可靠。

在机器学习任务中，除了可解释性，我们常常会提到另外两个概念：模型准确性（Accuracy）和模型复杂度（Model Complexity）。准确性反映了模型的拟合能力以及在某种程度上准确预测未知样本的能力。模型复杂度反映了模型结构上的复杂性，只与模型本身有关，与模型训练数据无关。在线性模型中，模型的复杂度由非零权重的个数来体现；在决策树模型中，模型的复杂度由树的深度体现；在神经网络模型中，模型复杂度则由神经网络的深度、宽度、模型的参数量以及模型的计算量来体现[13]。模型的复杂度与模型准确性相关联，又与模型的可解释性相对立。通常情况下，结构简单的模型可解释性好，但拟合能力差，往往准确率不高。结构复杂的模型，拟合能力强，准确性高，但由于模型参数量大、工作机制复杂、透明性低，因而可解释性又相对较差。

那么，在实际的学习任务中，我们是选择结构简单易于解释的模型然后训练它，还是训练复杂的最优模型然后开发可解释性技术解释它呢？基于这两种不同的选择，机器学习模型可解释性总体上可分为两类：ante-hoc 可解释性和 post-hoc 可解释性。其中，ante-hoc 可解释性指通过训练结构简单、可解释性好的模型或将可解释性结合到具体的模型结构中的自解释模型使模型本身具备可解释能力。Post-hoc 可解释性指通过开发可解释性技术解释已训练好的机器学习模型。根据解释目标和解释对象的不同，post-hoc 可解释性又可分为全局可解释性（Global Interpretability）和局部可解释性（Local Interpretability）。全局可解释性旨在帮助人们理解复杂模型背后的整体逻辑以及内部的工作机制[12]，局部可解释性旨在帮助人们理解机器学习模型针对每一个输入样本的决策过程和决策依据[14]。

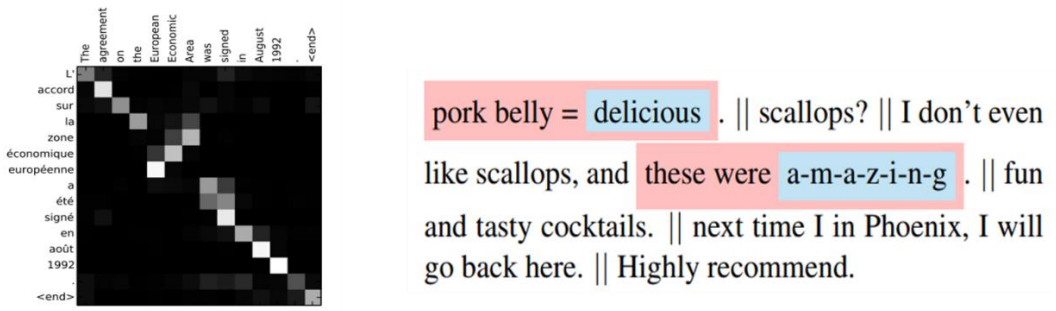
2 Ante-hoc 可解释性

Ante-hoc 可解释性指模型本身内置可解释性，即对于一个已训练好的学习模型，我们无需额外的信息就可以理解模型的决策过程或决策依据。模型的 ante-hoc 可解释性发生在模型训练之前，因而也称为事前可解释性。在学习任务中，我们通常采用结构简单、易于理解的自解释模型来实现 ante-hoc 可解释性，如朴素贝叶斯、线性回归、决策树、基于规则的模型。此外，我们也可以通过构建将可解释性直接结合到具体的模型结构中的学习模型来实现模型的内置可解释性[15]。

2.1 自解释模型

对于自解释模型，我们从两个角度考虑模型的可解释性和透明性，即模型整体的可模拟性（Simulatability）和模型单个组件的可分解性（Decomposability）。

严格意义上来讲，如果我们认为某个模型是透明的，那么我们一定能从整体上完全理解一个模型，也应该能够将输入数据连同模型的参数一起，在合理的时间步骤内完成产生预测所需的每一个计算（即整体上的可模拟性）。比如，在朴素贝叶斯模型中，由于条件独立性的假设，我们可以将模型的决策过程转化为概率运算[16; 17]。在线性模型中，我们可以基于模型权重，通过矩阵运算线性组合样本的特征值，复现线性模型的决策过程，其中模型权重体现了特征之间的相关关系[13; 17; 18]。而在决策树模型中，每一棵决策树都由表示特征或者属性的内部节点和表示类别的叶子节点组成，树的每一个分支代表一种可能的决策结果[19; 20]。决策树中每一



(a) sentence alignment in English-French translation[29] (b) word importance in sentiment analysis[34]

Fig.1 Visualization of attention weight in natural language processing applications.

图 1 自然语言处理应用中的注意力权重可视化

条从根节点到不同叶子节点的路径都代表着一条不同的决策规则，因而每一棵决策树都可以被线性化为一系列由 if-then 形式组成的决策规则[20-23]。因此，对于新的观测样本，我们可以通过从上到下遍历决策树，结合内部节点中的条件测试，基于 if-then 决策规则判定样本是否必须遵循左或右分支来模拟决策树的决策过程。

自解释模型的可分解性要求模型的每个部分，包括模型结构、模型参数，模型的每一个输入以及每一维特征都允许直观的解释[24]。在朴素贝叶斯模型中，由于条件独立性的假设，模型的预测可以很容易地转化为单个特征值的贡献—特征向量，特征向量的每一维表示每个特征值对最终分类结果的贡献程度[17]。在线性模型中，模型的权重直接反映了样本特征重要性，既包括重要性大小也包括相关性方向[25]。如果权重绝对值越大，则该特征对最终预测结果的贡献越大，反之则越小。如果权重值为正，则该特征与最终的预测类别正相关，反之则负相关。在决策树模型中，每个节点包含了特征值的条件测试，判定样本属于哪一分支以及使用哪一条规则，同时，每一条规则也为最终的分类结果提供了解释。此外，决策树模型自带的基于信息理论的筛选变量标准也有助于帮助我们理解在模型决策过程中哪些变量起到了显著的作用。

然而，由于人类认知的局限性，自解释模型的内置可解释性受模型的复杂度制约，这要求自解释模型结构一定不能过于复杂。因此，上述模型只有具有合理的规模才能具有有效的可解释性。例如，对于高维的线性模型，其内置可解释性未必优于 DNN。此外，对于决策树模型和基于规则的模型，如果树深度太深或者模型的规则太复杂，人类也未必能理解[12; 20]。但如果模型结构太简单，模型的拟合能力必然受限，因此模型可能会学习错误的特征来最小化在训练集上的经验误差，而这些特征可能与人类认知相违背，对于人类而言同样也很难解释。因此，自解释模型的内置可解释性与模型准确性之间始终存在一个平衡[13]。

2.2 广义加性模型

在实际学习任务中，简单模型（如线性模型）因为准确率低而无法满足需要，而复杂模型的高准确率又通常是牺牲自身可解释性为代价的。作为一种折中，广义加性模型既能提高简单线性模型的准确率，又能保留线性模型良好的内置可解释性[24; 26; 27]。广义加性模型一般形式如下：

$$g(y) = f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$$

其中，$f_i(\cdot)$ 为单特征（Single-feature）模型，也称为特征 x_i 对应的形函数（Shape Function）。广义加性模型通过线性函数组合每一单特征模型得到最终的决策形式。在广义加性模型中，形函数本身可能是非线性的，每一个单特征模型可能采用一个非常复杂的形函数 $f_i(x_i)$ 来量化每一个特征x_i与最终决策目标之间的关系，因而可以捕获到每一个特征与最终决策目标之间的非线性关系，因此广义加性模型准确率高于简单线性模型。又因为广义加性模型通过简单的线性函数组合每一个单特征模型得到最终的决策形式，消除了特征之间的相互作用，因此可

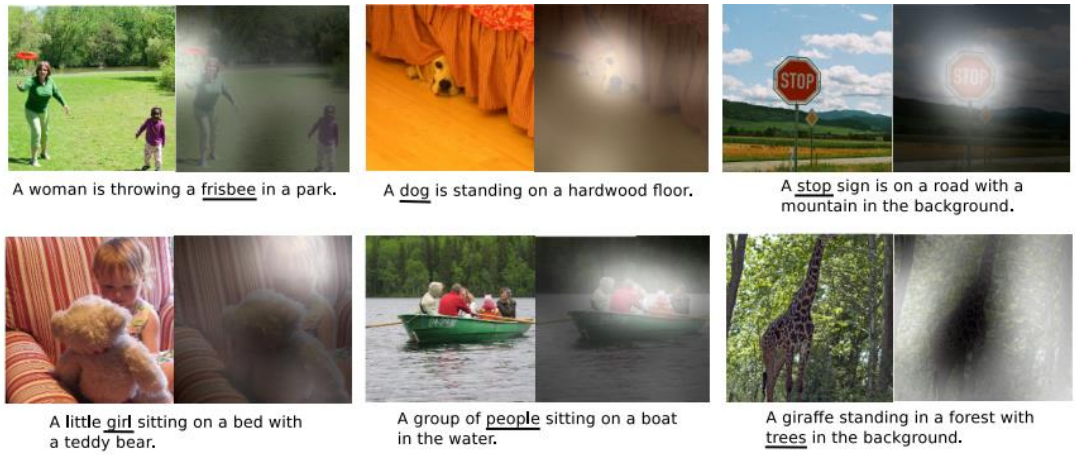


Fig.2 Alignment of words and images by attention in image caption task.

图 2 看图说话任务中注意力实现单词与图片的对齐[32]

以保留简单线性模型良好的可解释性，从而解决了复杂模型因为特征之间复杂的相关关系而削弱自身可解释性的问题。

Lou 等人[24]提出了一种新的基于有限大小的梯度提升树加性模型方法，该方法在回归和分类问题上精度显著优于传统方法，同时还保持了 GAM 模型的可解释性。Ravikumar 等人[28]结合稀疏线性建模和加性非参数回归的思想，提出了一种称之为稀疏加性模型的高维非参数回归分类方法，解决了高维空间中加性模型的拟合问题，同时基于ℓ_1正则的稀疏性，可实现特征的有效选择。 Poulin 等人[16]开发了一个图形化解释框架，提供了对加性模型的图形化解释，包括对模型整体的理解以及决策特征的可视化，以帮助建立用户与决策系统之间的信任关系。

2.3 注意力机制

神经网络模型由于模型结构复杂，算法透明性低，因而模型本身的可解释性差。因此，神经网络模型的自身可解释性只能通过额外引入可解释性模块来实现，一种有效的方法就是引入注意力机制（Attention Mechanism）[29-31]。

注意力机制源于对人类认知神经学的研究。在认知科学中，由于信息处理的瓶颈，人脑可以有意或无意地从大量输入信息中选择小部分有用信息来重点处理，同时忽略其他可见的信息，这就是人脑的注意力机制[32]。在计算能力有限的情况下，注意力机制是解决信息超载问题的一种有效手段，通过决定需要关注的输入部分，将有限的信息处理资源分配给更重要的任务。此外，注意力机制具有良好的可解释性，注意力权重矩阵直接体现了模型在决策过程中感兴趣的区域。

近年来，基于注意力机制的神经网络已成为神经网络研究的一大热点，并在自然语言处理、计算机视觉、推荐系统等领域有着大量的应用[33]。在自然语言处理领域，Bahdanau 等人[29]将注意力机制引入到基于编码器-解码器架构的机器翻译中，有效地提高了“英语-法语”翻译的性能。在编码阶段，机器翻译模型采用双向循环神经网络（Bi-RNN）将源语言编码到向量空间中；在解码阶段，注意力机制为解码器的隐藏状态分配不同的权重，从而允许解码器在生成法语翻译的每个步骤选择性地处理输入句子的不同部分。最后通过可视化注意力权重（如图 1（a）所示），用户可以清楚地理解一种语言中的单词是如何依赖另一种语言中的单词进行正确翻译的。Yang 等人[34]将分层注意力机制引入到文本分类任务中，显著提高了情感分析任务的性能，同时注意力权重量化了每一个词的重要性，可帮助人们清晰地理解每一个词对最终情感分类结果的贡献（如图 1（b）所示）。在计算机视觉领域，Xu 等人[32]将注意力机制应用于看图说话（Image Caption）任务中以产生对图片的描述。首先利用卷积神经网络（CNN）提取图片特征，然后基于提取的特征，利用带

Table 1 Summary of classic post-hoc interpretation methods.
表 1 经典的 post-hoc 解释方法总结

Method	G/L	MA/MS	TML	FCN	CNN	RNN	Fidelity	Security	Domain
inTree [23]	G	MS	✓	×	×	×	○	-	n/a
SGL [47]	G	MS	✓	×	×	×	○	-	n/a
GIRP [53]	G	MA	✓	✓	✓	✓	○	×	CV/NLP
MAGIX [58]	G	MA	✓	✓	✓	✓	○	-	n/a
DeepVID [70]	G	MA	×	×	✓	×	○	×	CV
AM [75]	G	MS	×	✓	✓	×	◐	×	CV
Nguyen et al. [79]	G	MS	×	✓	✓	×	◐	×	CV
Yuan et al. [82]	G	MS	×	×	×	✓	◐	×	NLP
Saliency Mask [93]	L	MA	×	✓	✓	×	○	×	CV
RSRS [94]	L	MA	×	✓	✓	×	○	×	CV
LIME [13]	L	MA	✓	✓	✓	✓	◐	×	CV/NLP
LORE [96]	L	MA	✓	✓	✓	✓	○	×	n/a
Anchor [98]	L	MA	✓	✓	✓	✓	◐	×	CV/NLP
LEMNA [99]	L	MS	×	×	×	✓	◐	×	NLP/Malware
Grad [73]	L	MS	×	✓	✓	✓	○	×	CV/NLP
DeconvNet [80]	L	MS	×	×	✓	×	◐	×	CV
GuidedBP [100]	L	MS	×	×	✓	×	◐	×	CV
Integrated [101]	L	MS	×	✓	✓	✓	○	×	CV/NLP
SmoothGrad [102]	L	MS	×	✓	✓	✓	◐	×	CV/NLP
LRP [105]	L	MS	×	✓	✓	✓	◐	×	CV/NLP
DeepLIFT [106]	L	MS	×	✓	✓	✓	◐	×	CV/Genomics
Guided Inversion [103]	L	MS	×	×	✓	×	◐	✓	CV
CAM [112]	L	MS	×	×	✓	×	◐	×	CV
Grad-CAM [113]	L	MS	×	×	✓	×	◐	×	CV
AI^2 [115]	L	MS	×	×	✓	×	○	✓	CV
OpenBox [116]	G, L	MS	×	✓	×	×	●	✓	CV

Note：G = global, L = local, MA = model-agnostic, MS = model-specific, TML = traditional machine learning, ○ = low, ◐ = middle, ● = high, - = unknown, CV = computer vision, NLP = natural language processing, and n/a = not mentioned in the literature.

注意力机制的循环神经网络（RNN）生成描述。在这个过程中，注意力实现了单词与图片之间的对齐，因此，通过可视化注意力权重矩阵，人们可以清楚地了解到模型在生成每一个单词时所对应的感兴趣的图片区域（如图 2 所示）。此外，注意力机制还被广泛地应用于推荐系统中，以研究可解释的推荐系统[35-39]。具体地，这些方法首先基于历史记录，利用注意力机制计算针对每一条记录的注意力分值，从而给不同的偏好设置不同的权重，或者通过注意力机制对用户行为、用户表征进行建模来学习用户的长期偏好，以推荐用户可能感兴趣的下一个项目；最后，通过可视化用户历史记录列表中每一条记录的注意力分值来提供对推荐结果的解释，以增强推荐系统自身的可解释性。

3 Post-hoc 可解释性

Post-hoc 可解释性也称事后可解释性，发生在模型训练之后。对于一个给定的训练好的学习模型，post-hoc 可解释性旨在利用解释方法或构建解释模型，解释学习模型的工作机制、决策行为和决策依据。因此，post-hoc 可解释性的重点在于设计高保真的解释方法或构建高精度的解释模型。根据解释目的和解释对象的不同，post-hoc 可解释性又分为全局可解释性和局部可解释性，所对应的方法分别称为全局解释方法和局部解释

方法。经典的 post-hoc 解释方法及其满足的属性如表 1 所示。

3.1 全局解释

机器学习模型的全局可解释性旨在帮助人们从整体上理解模型背后的复杂逻辑以及内部的工作机制，例如模型是如何学习的、模型从训练数据中学到了什么、模型是如何进行决策的等，这要求我们能以人类可理解的方式来表示一个训练好的复杂学习模型。典型的全局解释方法包括解释模型/规则提取、模型蒸馏、激活最大化解释等。

3.1.1 规则提取

早期针对模型可解释性的研究主要集中于解释规则或解释模型提取，即通过从受训模型中提取解释规则的方式，提供对复杂模型尤其是黑盒模型整体决策逻辑的理解[40-43]。规则提取技术以难以理解的复杂模型或黑盒模型作为入手点，利用可理解的规则集合生成可解释的符号描述，或从中提取可解释模型（如决策树、基于规则的模型等）[44-46]，使之具有与原模型相当的决策能力。解释模型或规则提取是一种有效的开箱技术，有效地提供了对复杂模型或黑盒模型内部工作机制的深入理解。根据解释对象不同，规则提取方法可分为针对树融合（Tree Ensemble）模型的规则提取[23; 47-50]和针对神经网络的规则提取。

针对复杂的树融合模型（例如随机森林、提升树等）的规则提取方法通常包含以下几个部分：首先，从树融合模型中提取规则，一个集成的树模型通常由多个决策树构成，每棵树的根节点到叶子节点的每一条路径都表示一条决策规则，将从每一棵决策树中提取的规则进行组合即可得到从树融合模型中提取的规则；其次，基于规则长度、规则频率、误差等指标对提取的规则进行排序，其中规则长度反映了规则的复杂度，规则频率反映满足规则的数据实例的比例，误差则反映了规则的决策能力；接着，基于排序结果，对规则中的无关项和冗余项进行剪枝并选择一组相关的非冗余规则；最后，基于挑选的规则构建一个可解释的规则学习器，用于决策和解释。

针对神经网络的规则提取方法可以分为两类：分解法（Decompositional Method）[51-53]和教学法（Pedagogical Method）[54-56]。分解法的显著特点是注重从受训神经网络中提取单个单元（如隐含单元、输出单元）层次上规则，这要求神经网络是“透明”的，即我们可以接触到模型的具体架构和参数。分解法要求受训神经网络中的每一个隐含单元和输出单元的计算结果都能映射成一个对应于一条规则的二进制结果。因此，每一个隐含单元或输出单元都可以被解释为一个阶跃函数或一条布尔规则。分解法通过聚合在单个单元级别提取的规则，形成整个受训神经网络的复合规则库，最后基于复合规则库提供对神经网络的整体解释。与分解法不同，教学法将受训神经网络模型当作是一个黑盒，即神经网络是“不透明”的，我们无法利用其结构和参数信息，只能操纵模型的输入和输出[57; 58]。因此，教学法旨在提取将输入直接映射到输出的规则，基本思想是结合符号学习算法，利用受训神经网络来为学习算法生成样本，最后从生成的样例中提取规则[55]。

然而，规则提取方法提取的规则往往不够精确，因而只能提供近似解释，不一定能反映待解释模型的真实行为。此外，规则提取方法提供的可解释性的质量受规则本身复杂度的制约，如果从待解释模型中提取的规则很复杂或者提取的决策树模型深度很深，那么提取的规则本身就不具备良好的可解释性，因而无法为待解释模型提供有效的解释。

3.1.2 模型蒸馏

当模型的结构过于复杂时，要想从整体上理解受训模型的决策逻辑通常是很困难的。解决该问题的一个有效途径是降低待解释模型的复杂度，而模型蒸馏（Model Distillation）则是降低模型复杂度的一个最典型的方法[59]。

模型蒸馏，也称知识蒸馏或模型模拟学习，是一种经典的模型压缩方法，其目的在

于将复杂模型学习的函数压缩为具有可比性能的更小，更快的模型[60]。模型蒸馏的核心思想是利用结构紧凑的学生模型（Student Model）来模拟结构复杂的教师模型（Teacher Model），从而完成从教师模型到学生模型的知识迁移过程，实现对复杂教师模型的知识“蒸馏”。蒸馏的难点在于压缩模型结构的同时如何保留教师模型从海量数据中学习到的知识和模型的泛化能力。一种有效的解决办法是利用软目标来辅助硬目标一起训练学生模型，其中硬目标为原始数据的类别信息，软目标为教师模型的分类概率值，包含的信息量大，体现了不同类别之间相关关系的信息[61]。给定一个复杂的教师模型和一批训练数据，模型蒸馏方法首先利用教师模型生成软目标，然后通过最小化软目标和硬目标的联合损失函数来训练学生模型，损失函数定义如下：

$$L_{stdudent} = \alpha L^{(soft)} + (1-\alpha)L^{(hard)}$$

其中，$L^{(soft)}$为软目标损失，要求学生模型生成的软目标与教师模型生成的软目标要尽可能的接近，保证学生模型能有效地学习教师模型中的暗知识（Dark Knowledge）；$L^{(hard)}$为硬目标损失，要求学生模型能够保留教师模型良好的决策性能。

由于模型蒸馏可以完成从教师模型到学生模型的知识迁移，因而学生模型可以看作是教师模型的全局近似，在一定程度上反映了教师模型的整体逻辑，因此我们可以基于学生模型，提供对教师模型的全局解释。在利用模型蒸馏作为全局解释方法时，学生模型通常采用可解释性好的模型来实现，如线性模型、决策树、广义加性模型以及浅层神经网络等[62-64]。Hinton 等人[61]提出了一种知识蒸馏方法，通过训练单一的相对较小的网络来模拟原始复杂网络或集成网络模型的预测概率来提炼复杂网络的知识，以模拟原始复杂网络的决策过程，并且证明单一网络能达到复杂网络几乎同样的性能。为了进一步提升蒸馏知识的可解释性，Frosst 等人[63]扩展了 Hinton 提出的知识蒸馏方法，提出利用决策树来模拟复杂深度神经网络模型的决策。Tan 等人[64]基于广义加性模型的良好可解释性，提出利用模型蒸馏的方法来学习描述输入特征与复杂模型的预测之间关系的全局加性模型，并基于加性模型对复杂模型进行全局解释。Che 等人[65]将基于模型蒸馏的可解释方法应用于医疗诊断模型的可解释性研究中，提出利用梯度提升树进行知识蒸馏的方式来学习可解释模型，不仅在急性肺损伤病人无呼吸机天数预测任务中取得了优异的性能，而且还可以为临床医生提供良好的可解释性。Ding 等人[66]利用知识蒸馏解释基于社交媒体的物质使用预测模型，通过运用知识蒸馏框架来构建解释模型，取得了与最先进的预测模型相当的性能，而且还可以提供对用户的社交媒体行为与物质使用之间的关系深入理解。Xu 等人[67]开发了 DarkSight 可解释方法，通过利用模型蒸馏的方式从黑盒模型中提取暗知识，并以可视化的形式对提取的暗知识进行呈现，以帮助分析师直观地了解模型决策逻辑。

此外，基于模型蒸馏的解释方法还被广泛地应用于模型诊断与验证[68-70]。Tan 等人[68]提出了一种针对黑盒风险评分模型的两阶段模型审计方法，对于一个给定的黑盒风险评分模型和一批审计数据，该方法首先利用模型蒸馏的方法得到一个解释模型，同时基于审计数据和其真实标签训练一个透明的结果预测模型，并通过比较解释模型和结果预测模型来理解特征与风险评分之间的相关关系；最后，通过使用统计测试的方式来确定黑盒模型是否使用了审计数据中不存在的其他特征。同时，通过评估受保护特征对风险评分的贡献与其对实际结果的贡献的差异，可以检测黑盒风险评分模型中是否存在偏差[69]。

模型蒸馏解释方法实现简单，易于理解，且不依赖待解释模型的具体结构信息，因而作为一种模型无关的解释方法，常被用于解释黑盒机器学习模型。然而，蒸馏模型只是对原始复杂模型的一种全局近似，它们之间始终存在差距。因此，基于蒸馏模型所做出的解释不一定能反映待解释模型的真实行为。此外，知识蒸馏过程通常不可控，无法保障待解释模型从海量数据中学到的知识有效地迁移到蒸馏模型中，因而导致解释结果质量较低无法满足精确解释的需要。

Fig.3 Class-discriminative prototypes generated by combining generative model with activation maximization.

图 3 利用生成模型与激活最大化相结合生成的类别对应原型样本[79]

3.1.3 激活最大化

在自下而上的深度学习任务中，给定一批训练数据，DNN 不仅可以自动地学习输入数据与输出类别之间的映射关系，同时也可以从数据中学到特定的特征表示（Feature Representation）。然而，考虑到数据集中存在偏差，我们无法通过模型精度来保证模型表征的可靠性，也无法确定 DNN 用于预测的内部工作模式[71]。因此，深入理解并呈现 DNN 中每一个隐含层的神经元所捕获的表征，有助于从语义上、视觉上帮助人们理解 DNN 内部的工作逻辑[72]。为此，许多研究者探索如何在输入空间实现对 DNN 任意层神经单元计算内容的可视化，并使其尽可能通用，以便能够深入了解神经网络不同单元代表的特定含义。其中，最有效和使用最广泛的一种方法是通过在特定的层上找到神经元的首选输入最大化神经元激活，因此该方法也称为激活最大化（Activation Maximization, AM）方法[73]。

激活最大化方法思想较为简单，即通过寻找有界范数的输入模式，最大限度地激活给定的隐藏单元，而一个单元最大限度地响应的输入模式可能是一个单元正在做什么的良好的一阶表示[74-76]。给定一个 DNN 模型，寻找最大化神经元激活的原型样本$\boldsymbol{x}^*$的问题可以被定义成一个优化问题，其形式化定义如下：

$$x^* = \arg\max_x f_l(x) - \lambda \|x\|^2$$

其中，优化目标第一项 $f_l(x)$ 为 DNN 第l层某一个神经元在当前输入$\boldsymbol{x}$下的激活值；第二项为ℓ_2正则，用于保证优化得到的原型样本（Prototype）与原样本尽可能的接近。整个优化过程可以通过梯度上升来求解。最后，通过可视化生成的原型样本$\boldsymbol{x}^*$，可以帮助我们理解该神经元在其感受野中所捕获到的内容。当然，我们可以分析任意层的神经元，以理解 DNN 不同层所编码的不同表示内容。当我们分析输出层神经元的最大激活时，我们可以找到某一类别所对应的最具代表性的原型样本。

激活最大化方法虽然原理简单，但如何使其正常工作同样面临着一些挑战。由于样本搜索空间很大，优化过程可能产生含有噪声和高频模式的不现实图像，导致原型样本虽能最大化神经元激活却难以理解。为了获取更有意义、更自然的原型样本，优化过程必须采用自然图像先验约束，为此，一些研究者创造性地提出了人工构造先验，包括α范数、高斯模糊等[77; 78]。此外，一些研究者将激活最大化框架与生成模型相结合，利用生成模型产生的更强的自然图像先验正则化优化过程。Nguyen 等人[79]提出利用生成对抗网络与激活最大化优化相结合的方法来生成原型样本，优化问题被重定义为：

$$z^* = \arg\max_{z \in Z} f_l(g(z)) - \lambda \|z\|^2$$

其中，第一项为解码器与原神经元激活值的结合，第二项为代码空间中的ℓ_2正则。该方法不直接优化图像，转而优化代码空间以找

到可以最大化神经元激活的解$\boldsymbol{z}^*$，一旦最优解$\boldsymbol{z}^*$找到，则可以通过解码得到原型样本$\boldsymbol{z}^*$，即$x^* = g(z^*)$。实验结果表明（如图 3 所示），将激活最大化与生成模型相结合的方法可以产生更真实、更具有可解释性的原型样本。从图 3 可以看出，模型成功捕获了与类别相对应的特征表示。

对不同层生成的原型样本的可视化结果表明，DNN 在若干抽象层次上进行表示学习，从模型的第一层到最后一层，模型学习到的特征表征由局部过渡到整体，由一般任务过渡到特定任务。以图像分类任务中的 CNN 为例，低层神经元通常可以捕获到图片中的颜色、边缘等信息；中间层神经元有更复杂的不变性，可以捕获相似的纹理；中高层神经元可以捕获图片中的显著变化，并可以聚焦到特定类别对应的局部特征，如狗的脸部，鸟的脚部等；最后，高层神经元则通过组合局部特征表征，从而学习到整个分类目标的整体表征[80]。此外，神经元具有多面性，可以对与同一语义概念相关的不同图像做出反应，例如，人脸检测神经元可以同时对人脸和动物面孔做出反应[81]。

激活最大化解释方法是一种模型相关的解释方法，相比规则提取解释和模型蒸馏解释，其解释结果更准确，更能反映待解释模型的真实行为。同时，利用激活最大化解释方法，可从语义上、视觉上帮助人们理解模型是如何从数据中进行学习的以及模型从数据中学到了什么。然而，激活最大化本身是一个优化问题，在通过激活最大化寻找原型样本的过程中，优化过程中的噪音和不确定性可能导致产生的原型样本难以解释。尽管可以通过构造自然图像先验约束优化过程来解决这一问题，但如何构造更好的自然图像先验本身就是一大难题。此外，激活最大化方法只能用于优化连续性数据，无法直接应用于诸如文本、图数据等离散型数据[82]，因而该方法难以直接用于解释自然语言处理模型和图神经网络模型。

3.2 局部解释

机器学习模型的局部可解释性旨在帮助人们理解学习模型针对每一个特定输入样本的决策过程和决策依据。与全局可解释性不同，模型的局部可解释性以输入样本为导向，通常可以通过分析输入样本的每一维特征对模型最终决策结果的贡献来实现。在实际应用中，由于模型算法的不透明性、模型结构的复杂性以及应用场景的多元性，提供对机器学习模型的全局解释通常比提供局部解释更困难，因而针对模型局部可解释性的研究更加广泛，局部解释方法相对于全局解释方法也更常见。经典的局部解释方法包括敏感性分析解释、局部近似解释、梯度反向传播解释、特征反演解释以及类激活映射解释等。

3.2.1 敏感性分析

敏感性分析（Sensitivity Analysis）是指在给定的一组假设下，从定量分析的角度研究相关自变量发生某种变化对某一特定的因变量影响程度的一种不确定分析技术[83]，其核心思想是通过逐一改变自变量的值来解释因变量受自变量变化影响大小的规律。敏感性分析被广泛地应用于机器学习及其应用中，如机器学习模型分析[84-86]、生态建模[87]、医学诊断[88]等。近年来，敏感性分析作为一种模型局部解释方法，被用于分析待解释样本的每一维特征对模型最终分类结果的影响[89-91]，以提供对某一个特定决策结果的解释。根据是否需要利用模型的梯度信息，敏感性分析方法可分为模型相关方法和模型无关方法。

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

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

其中，$f(\boldsymbol{x})$为模型的决策函数，x_i为待解释样本$\boldsymbol{x}$的第i维特征。直观地，相关性分数$R_i(\boldsymbol{x})$可以看作是模型梯度的ℓ_2范数的分解，即$\sum_{i=1}^{d} R_i(x) = \left\|\nabla f(x)\right\|^2$。在模型相关方法中，相关性分数$R_i(\boldsymbol{x})$可通过梯度反向传播来求解。最后，通过以热力图的形式可视化相关性分数可以直观地理解输入的每一维特征对决策结果的影响程度。

Fig.4 Learn a saliency mask (right) by blurring an image (middle) to minimize the probability of its target class.

图 4 通过图像模糊的方式（中间图）最小化分类概率来学习显著性掩码（右图）[93]

在模型无关敏感性分析方法中，待解释模型可以看作是黑盒，我们无需利用模型的梯度信息，只关注待解释样本特征值变化对模型最终决策结果的影响。Robnik-Sikonja 等人[92]提出通过对输入样本单个属性值的预测进行分解的方式来观察属性值对该样本预测结果的影响。具体地，该方法通过观察去掉某一特定属性前后模型预测结果的变化来确定该属性对预测结果的重要性，即

$$R_i(x) = f(x) - f(x \setminus x_i)$$

类似地，Liu 等人[94]提出了“限制支持域集”的概念，它被定义为一组受大小限制且不重叠的区域，并且满足如下属性：删除任何一个区域将会导致模型分类出错。其本质思想是，如果某个特定的区域的缺失导致模型分类结果发生反转，则该区域必定为模型正确决策提供支持。因此，最终可通过分析特定图像区域是否存在与模型决策结果之间的依赖关系来可视化模型决策规则。Fong 等人[93]提出了一种基于有意义扰动的敏感性分析方法，通过添加扰动或删除待解释图片的不同区域来最小化模型目标类别分类概率的方式学习一个显著性掩码，以识别对模型决策结果影响最大的图像部分，并可视化显著性掩码作为对该决策结果的解释（如图 4 所示）。Li 等人[95]则提出通过观察修改或删除特征子集前后模型决策结果的相应变化的方式来推断待解释样本的决策特征。

然而，敏感性分析方法解释的是决策函数 $f(x)$ 局部变化对决策结果的影响，而不是解释决策函数本身，只能捕获到单个特征对最终决策结果的影响程度，而不一定关注实际的决策相关特征，因而相关性分值 $R_i(x)$ 对应的热力图在空间上是分散而不连续的。因此，敏感性分析方法提供的解释结果通常相对粗糙且难以理解。此外，敏感性分析方法无法解释特征之间的相关关系对最终决策结果的影响。

3.2.2 局部近似

局部近似解释方法的核心思想是利用结构简单的可解释模型拟合待解释模型针对某一输入实例的决策结果，然后基于解释模型对该决策结果进行解释。该方法通常基于如下假设：给定一个输入实例，模型针对该实例以及该实例邻域内样本的决策边界可以通过可解释的白盒模型来近似。在整个数据空间中，待解释模型的决策边界可以任意的复杂，但模型针对某一特定实例的决策边界通常是简单的，甚至是近线性的[13]。我们通常很难也不需要对待解释模型的整体决策边界进行全局近似，但可在给定的实例及其邻域内利用可解释模型对待解释模型的局部决策边界进行近似，然后基于可解释模型提供对待解释模型的决策依据的解释。

Ribeiro 等人[13]基于神经网络的局部线性假设，提出了一种模型无关局部可解释方法（LIME）。具体地，对于每一个输入实例，LIME 首先利用该实例以及该实例的一组近邻训练一个易于解释的线性回归模型来拟合待解释模型的局部边界，然后基于该线性模型解释待解释模型针对该实例的决策依据，其中，线性模型的权重系数直接体现了当前决策中该实例的每一维特征重要性。Guidotti 等人[96]提出了一种适用于关系表数据的基于局部规则的黑盒模型决策结果解释方法（LORE）。给定一个二分类模型 f 及一个由 f 标记的特定实例 $\boldsymbol{x}$，LORE 首先

利用 ad-hoc 遗传算法生成给定实例$\boldsymbol{x}$的一组平衡邻居实例来构建一个简单的、可解释的预测模型，以逼近二分类模型f针对实例$\boldsymbol{x}$的决策边界；然后，基于该解释模型，从生成的实例集合中提取一个决策树模型；最后，从决策树模型中提取决策规则作为对实例$\boldsymbol{x}$的分类结果的局部解释。Ribeiro 等人[97; 98]提出了一种称之为锚点解释（Anchor）的局部解释方法，针对每一个输入实例，该方法利用被称之为“锚点”的 if-then 规则来逼近待解释模型的局部边界。Anchor 方法充分地结合了模型无关局部解释方法的优点和规则的良好的可解释性，在 Anchor 方法中用于解释的“锚点”通常是直观、易于理解的，而且解释覆盖范围非常清晰。通过构造，“锚点”不仅可以与待解释模型保持一致，而且还可以以确保正确理解和高保真的方式将待解释模型的决策行为传达给用户。

然而，LIME、LORE 以及 Anchor 等解释方法均假设输入样本的特征是相互独立，因而无法准确地解释诸如 RNN 等专门对序列数据中的依赖关系进行建模的模型。为此，Guo 等人[99]提出了 LEMNA，一种专用于安全应用场景中的 RNN 模型的高保真解释方法，其核心思想与 LIME 等方法相似，即利用可解释模型来近似 RNN 的局部决策边界，并针对每一个输入实例，产生一组可解释的特征以解释针对该实例的决策依据。与 LIME 不同的是，LEMNA 假设待解释模型的局部边界是非线性的，为了保证解释的保真度，LEMNA 通过训练混合回归模型来近似 RNN 针对每个输入实例的局部决策边界。此外，LEMNA 引入了融合 Lasso 正则来处理 RNN 模型中的特征依赖问题，有效地弥补了 LIME 等方法的不足。

基于局部近似的解释方法实现简单，易于理解且不依赖待解释模型的具体结构，适于解释黑盒机器学习模型。但解释模型只是待解释模型的局部近似，因而只能捕获模型的局部特征，无法解释模型的整体决策行为。针对每一个输入实例，局部近似解释方法均需要重新训练一个解释模型来拟合待解释模型针对该实例的决策结果，因而此类方法的解释效率通常不高。此外，大多数的局部近似解释方法假设待解释实例的特征相互独立，因此无法解释特征之间的相关关系对决策结果的影响。

3.2.3 反向传播

基于反向传播（Back Propagation）的解释方法的核心思想是利用 DNN 的反向传播机制将模型的决策重要性信号从模型的输出层神经元逐层传播到模型的输入以推导输入样本的特征重要性。

Simonyan 等人[73]最先提出了利用反向传播推断特征重要性的解释方法（Grad），具体地，Grad 方法通过利用反向传播算法计算模型的输出相对于输入图片的梯度来求解该输入图片所对应的分类显著图（Saliency Map）。与 Grad 方法类似，Zeiler 等人[80]提出了反卷积网络（DeconvNet），通过将 DNN 的高层激活反向传播到模型的输入以识别输入图片中负责激活的重要部分。不同的是，在处理线性整流单元（ReLU）过程中，当使用 Grad 方法反向传播重要性时，如果正向传播过程中 ReLU 的输入为负，则反向传播过程中传入 ReLU 的梯度值为零。而在反卷积网络中反向传播一个重要信号时，当且仅当信号值为负，进入 ReLU 的重要信号被置零，而不考虑前向传播过程中输入到 ReLU 的信号的符号。Springenberg 等人[100]将 Grad 方法与反卷积网络相结合提出了导向反向传播方法（GuidedBP），通过在反向传播过程中丢弃负值来修改 ReLU 函数的梯度。与只计算输出针对当前输入的梯度不同，Sundararajan 等人[101]提出了一种集成梯度方法（Integrated），该方法通过计算输入从某些起始值按比例放大到当前值的梯度的积分代替单一梯度，有效地解决了 DNN 中神经元饱和问题导致无法利用梯度信息反映特征重要性的问题。

然而，Grad、GuidedBP 以及 Integrated 等方法通过反向传播所得到的显著图通常包含很多视觉可见的噪音（如图 5 所示），而我们无法确定这种噪音是否真实地反映了模型在分类过程中的决策依据。为此，Smilkov 等人[102]提出了一种平滑梯度的反向传播解释方法（SmoothGrad），该方法通

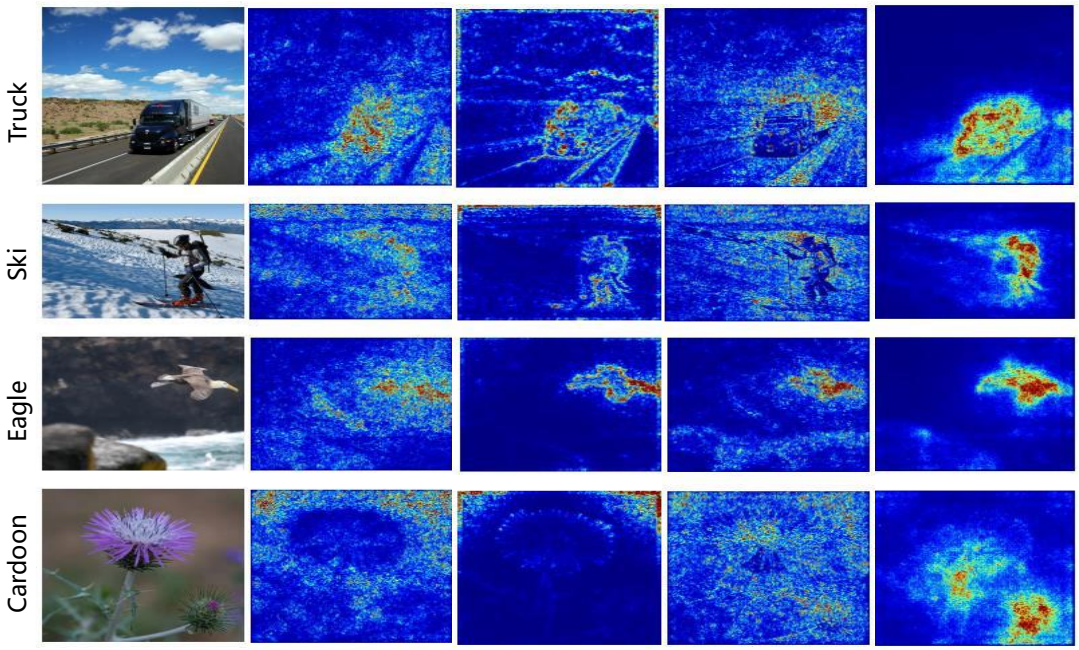


Fig.5 Comparison of interpretation quality of four gradient back-propagation based interpretation methods.

图 5 四种梯度反向传播解释方法解释效果对比[103]

过向输入样本中引入噪声解决了 Grad 等方法中存在的视觉噪音问题。SmoothGrad 方法的核心思想是通过向待解释样本中添加噪声对相似的样本进行采样，然后利用反向传播方法求解每个采样样本的决策显著图，最后将所有求解得到的显著图进行平均并将其作为对模型针对该样本的决策结果的解释。

尽管上述基于梯度反向传播的方法可以定位输入样本中决策特征，但却无法量化每个特征对模型决策结果的贡献程度。因此，Landecker 等人[104]提出一种贡献传播方法，该方法首先利用加性模型计算 DNN 高层特征对模型分类结果的贡献，然后通过反向传播将高层特征的贡献逐层传递到模型的输入，以确定每一层的每一个神经元节点对其下一层神经元节点的相对贡献。给定一个待解释样本，该方法不仅可以定位样本中的重要特征，而且还能量化每一个特征对于分类结果的重要性。Bach 等人[105]则提出了一种分层相关性传播方法（LRP），用于计算单个像素对图像分类器预测结果的贡献。一般形式的 LRP 方法假设分类器可以被分解为多个计算层，每一层都可以被建模为一个多维向量并且该多维向量的每一维都对应一个相关性分值，LRP 的核心则是利用反向传播将高层的相关性分值递归地传播到低层直至传播到输入层。Shrikumar 等人[106] 对 LRP 方法进行了改进（DeepLIFT），通过在输入空间中定义参考点并参考神经元激活的变化按比例传播相关分数。其研究结果表明，在不进行数值稳定性修正的情况下，原始 LRP 方法的输出结果等价于 Grad 方法所求显著图与输入之间的乘积。与梯度反向传播方法不同的是，LRP 方法不要求 DNN 神经元的激活是可微的或平滑的。基于此优点，Ding 等人[107]首次将 LRP 方法应用于基于注意力机制的编码器-解码器框架，以度量神经网络中任意两个神经元之间关联程度的相关性。在汉英翻译案例中的研究表明，该方法有助于解释神经机器翻译系统的内部工作机制并分析翻译错误。类似地，Arras 等人[108]将 LRP 方法引入到自然语言处理任务中，并且从定性和定量的角度证明 LRP 方法既可以用于文档级别的细粒度分析，也可以作为跨文档的数据集级别的分析，以识别对分类器决策很重要的单词。

基于反向传播的解释方法通常实现简单、计算效率高且充分利用了模型的结构特性。然而，从理论上易知，如果预测函数在

(a) input (b) inversion map (c) elephant (d) zebra

Fig.6 Interpretation example of guided feature inversion method.

图 6 导向特征反演方法解释示例[103]

输入附近变得平坦，那么预测函数相对于输入的梯度在该输入附近将变得很小，进而导致无法利用梯度信息定位样本的决策特征。尽管 Integrated 方法在一定程度上解决了该问题，但同时也增加了计算开销，并且 Integrated 方法的解释结果中依然存在许多人类无法理解的噪音。此外，梯度信息只能用于定位重要特征，而无法量化特征对决策结果的重要程度，利用基于重要性或相关性反向传播的解释方法则可以解决该问题。

3.2.4 特征反演

尽管敏感性分析、局部近似以及梯度反向传播等方法在一定程度上可以提供对待解释模型决策结果的局部解释，但它们通常忽略了待解释模型的中间层，因而遗漏了大量的中间信息。而利用模型的中间层信息，我们能更容易地表征模型在正常工作条件下的决策行为，进而可提供更准确的解释结果。特征反演（Feature Inversion）作为一种可视化和理解 DNN 中间特征表征的技术，可以充分利用模型的中间层信息，以提供对模型整体行为及模型决策结果的解释。

特征反演解释方法可分为模型级（Model-level）解释方法和实例级（Instance-level）解释方法。模型级解释方法旨在从输入空间中寻找可以表示 DNN 神经元所学到的抽象概念的解释原型（如激活最大化方法），并通过可视化和理解 DNN 每一层特征表示的方式，提供对 DNN 每一层所提取信息的理解[73; 77; 109; 110]。然而，模型级解释方法的反演结果通常相对粗糙且难以理解，此外，如何从输入样本中自动化提取用于模型决策的重要特征仍然面临着巨大的挑战。针对模型级方法的不足，实例级特征反演方法试图回答输入样本的哪些特征被用于激活 DNN 的神经元以做出特定的决策。其中，最具代表性的是 Du 等人[103]提出的一个实例级特征反演解释框架，该框架通过在执行导向特征反演过程中加入类别依赖约束，不仅可以准确地定位待输入实例中的用于模型决策的重要特征（如图 6 所示），还可以提供对 DNN 模型决策过程的深入理解。

3.2.5 类激活映射

最新研究表明，CNN 不同层次的卷积单元包含大量的位置信息，使其具有良好的定位能力[111]。基于卷积单元的定位能力，我们可以定位出输入样本中用于 CNN 决策的核心区域，如分类任务中的决策特征、目标检测任务中的物体位置等。然而，传统 CNN 模型通常在卷积和池化之后采用全连接层对卷积层提取的特征图进行组合用于最终决策，因而导致网络的定位能力丧失。

为解决这一问题，Zhou 等人[112]提出了类激活映射（Class Activation Mapping，CAM）解释方法，该方法利用全局平均池化（Global Average Pooling）层来替代传统 CNN 模型中除 softmax 层以外的所有全连接层，并通过将输出层的权重投影到卷积特征图来识别图像中的重要区域。具体地，CAM 首先利用全局平均池化操作输出 CNN 最后一个卷积层每个单元的特征图的空间平均值，并通过对空间平均值进行加权求和得到 CNN 的最终决策结果。同时，CAM 通过计算最后一个卷积层的特征图的加权和，得到 CNN 模型的类激活图，而一个特定类别所对应的类激活图则反映了 CNN 用来识别该类别的

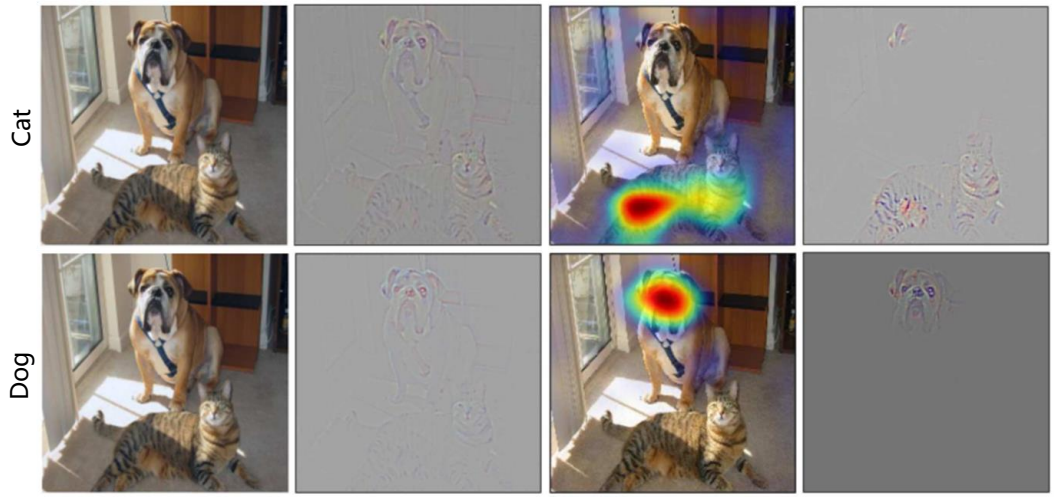


Fig.7 Visualization of interpretation results of Grad-CAM and Guided Grad-CAM methods.

图 7 Grad-CAM 与 Guided Grad-CAM 方法解释结果可视化[113]

核心图像区域。最后，通过以热力图的形式可视化类激活图得到最终的解释结果。研究结果表明，全局平均池化层的优势远不止于作为一个正则器来防止网络过拟合，事实上，通过稍加调整，全局平均池化还可以将CNN良好的定位能力保留到网络的最后一层[112]。

然而，CAM 方法需要修改网络结构并重训练模型，因而在实际应用中并不实用。因此，Selvaraju 等人[113]对 CAM 方法进行了改进，提出了一种将梯度信息与特征映射相结合的梯度加权类激活映射方法（Grad-CAM）。给定一个输入样本，Grad-CAM 首先计算目标类别相对于最后一个卷积层中每一个特征图的梯度并对梯度进行全局平均池化，以获得每个特征图的重要性权重；然后，基于重要性权重计算特征图的加权激活，以获得一个粗粒度的梯度加权类激活图，用于定位输入样本中具有类判别性的重要区域（如图 7(c)所示）。与 CAM 相比，Grad-CAM 无需修改网络架构或重训练模型，避免了模型的可解释性与准确性之间的权衡，因而可适用于多种任务以及任何基于 CNN 结构的模型，对于全卷积神经网络，Grad-CAM 退化为 CAM 方法。尽管 Grad-CAM 具有良好的类别判别能力并能很好地定位相关图像区域，但缺乏诸如 GuidedBP[100]和 DeconvNet[80]等像素级别梯度可视化解释方法显示细粒度特征重要性的能力[113]。为获得更细粒度的特征重要性，作者将 Grad-CAM 与 GuidedBP 方法相结合提出了导向梯度加权类激活映射方法（Guided Grad-CAM），该方法首先利用双线性插值将梯度加权类激活图上采样到输入图片分辨率大小，然后点乘 GuidedBP 方法的输出结果，得到细粒度的类判别性特征定位图（如图 7(d)所示）。研究结果表明，Guided Grad-CAM 方法解释效果优于 GuidedBP 和 Grad-CAM。

类激活映射解释方法实现简单、计算效率高，解释结果视觉效果好且易于理解，但这类方法只适用于解释 CNN 模型，很难扩展到全连接神经网络（FCN）以及 RNN 等模型。此外，CAM 方法需要修改网络结构并重训练模型，模型的准确性与可解释性之间始终存在一个权衡，且针对重训练模型做出的解释结果与原待解释模型的真实行为之间存在一定的不一致性，因而在真实应用场景中很难适用。Grad-CAM 虽然解决了 CAM 需要进行网络修改和模型重训练的问题，但仍然与 CAM 方法一样只能提供粗粒度的解释结果，无法满足安全敏感应用场景（如自动驾驶、医疗诊断等）中对精细化解释的需要。Guided Grad-CAM 方法作为 CAM 和 Grad-CAM 的加强版，既不需要修改网络结构或重训练模型，又能提供更细粒度的解释结果，但由于引入了导向反向传播方法，因而该方法同样存在由于负梯度归零导致无法定位与模型决策结果呈负相关的

样本特征的局限性[114]。

3.2.6 其他方法

除了上述几种典型的局部可解释方法外，其他研究者从不同的角度对模型可解释性进行了深入研究，并提出了一些新的局部解释方法，包括抽象解释[115]和准确一致解释[116]等。

针对 DNN 系统的可靠分析技术所面临的主要挑战是如何在解释神经网络某些特性的同时将其扩展到大规模的 DNN 分类器，因此，分析方法必须考虑到任何经过大量中间神经元处理的大规模输入集上所有可能的模型输出结果。由于模型的输入空间通常是巨大的，因而通过在所有可能的输入样本上运行模型来检查它们是否满足某一特性是不可行的。为解决这一挑战，避免状态空间爆炸，Gehr 等人[115]将程序分析中的经典抽象解释框架应用于 DNN 分析，首次提出了可扩展的、可用于验证和分析 DNN 安全性和鲁棒性的抽象解释系统（AI^2）。具体地，AI^2首先构造一个包含一系列逻辑约束和抽象元素的数值抽象域；由于 DNN 的每一层处理的是具体的数值，因而抽象元素无法在网络中传播。为解决此问题，AI^2通过定义一个被称之为抽象转换器（Abstract Transformer）的函数将 DNN 的每一层转换为对应的抽象层，并基于抽象元素过近似（Over-approximation）原神经网络每一层的处理函数以捕获其真实行为；最后，AI^2基于抽象转换器返回的抽象结果，分析并验证神经网络的鲁棒性和安全性。AI^2不要真正地运行 DNN 模型即可验证 DNN 的某些特定属性，因而计算效率高，可扩展到大规模、更复杂的 DNN 网络。但由于采用了过近似处理，尽管AI^2能提供可靠的解释但无法保证解释的准确性。

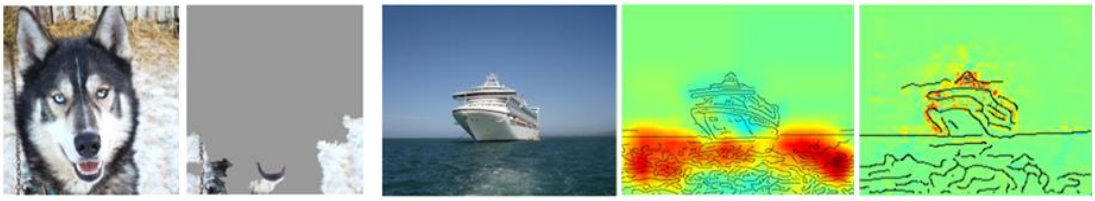
现有局部解释方法包括抽象解释都很难保证解释结果的准确性和一致性，为此，许多学者开始研究针对 DNN 模型的精确解释方法。Chu 等人[116]提出了一种准确一致的解释方法（OpenBox），可为分段线性神经网络（PLNN）家族模型提供精确一致的解释。作者研究证明，PLNN 在数学上等价于一系列的局部线性分类器，其中每一个线性分类器负责分类输入空间中的一组样本。因此，给定一个待解释 PLNN 模型，OpenBox 首先利用神经网络的前向传播机制和矩阵运算将给定的 PLNN 模型表示成数学上与之等价的、由一系列数据依赖的局部线性分类器组成的线性解释模型；然后，针对每一个待解释样本，OpenBox 基于该样本所对应的局部线性分类器提供对 PLNN 分类结果的解释。研究结果表明，由于线性解释模型数学上与待解释 PLNN 等价，因此基于线性解释模型给出的解释结果能精确地反映 PLNN 的真实决策行为，并且线性解释模型针对每一个输入的决策结果与待解释 PLNN 的决策结果完全一致，从而解决了模型的可解释性与准确性之间的权衡难题。此外，针对近似的样本，OpenBox 可以给出一致的解释，保证了解释结果的一致性。然而，OpenBox 作为针对 PLNN 家族的特定解释方法，只能解释线性神经网络模型，无法用于解释非线性神经网络模型。此外，如何将其扩展到 CNN、RNN 等更复杂的神经网络模型同样面临着巨大的挑战。

4 可解释性应用

机器学习模型可解释性相关技术潜在应用非常广泛，具体包括模型验证、模型诊断、辅助分析以及知识发现等。

4.1 模型验证

传统的模型验证方法通常是通过构造一个与训练集不相交的验证集，然后基于模型在验证集上的误差来评估模型的泛化性能，从而提供对模型好坏的一个粗粒度的验证。然而，由于数据集中可能存在偏差，并且验证集也可能与训练集同分布，我们很难简单地通过评估模型在验证集上的泛化能力来验证模型的可靠性，也很难验证模型是否从训练数据中学到了真正的决策知识。以冰原狼与哈士奇的分类为例，由于训练集中所有冰原狼样本图片的背景均为雪地，导致分类模型可能从训练集中学到数据偏差从而将雪作为冰原狼的分类特征，又由于验证集与训练集同分布，模型在验证集上的分类



(a) validation of husky classifier[13] (b)validation of boat classifier[117]

Fig.8 Examples of interpretation-based model validation.

图 8 基于可解释性的模型验证示例

性能与在训练集上的性能同样优异，因而导致传统的模型验证方法将该模型识别为一个好的分类模型[13]。很显然，这样的模型通常是不可靠的，一旦模型在推理阶段遇到背景为雪地的哈士奇样本图片，分类模型会做出错误的决策，而模型的这种行为将会给实际场景尤其是风险敏感场景中的真实应用带来潜在的威胁。

针对传统模型验证方法的不足，我们可以利用模型的可解释性及相关解释方法对模型可靠性进行更细粒度的评估和验证，从而消除模型在实际部署应用中的潜在风险。基于可解释性的模型验证方法一般思路如下：首先构造一个可信验证集，消除验证集中可能存在的数据偏差，保证验证数据的可靠性；然后，基于可信验证集，利用相关解释方法提供对模型整体决策行为（全局解释）或模型决策结果（局部解释）的解释；最后，基于解释方法给出的解释结果并结合人类认知，对模型决策行为和决策结果的可靠性进行验证，以检查模型是否在以符合人类认知的形式正常工作。

在冰原狼与哈士奇分类的例子中，Ribeiro 等人[13]利用局部解释方法 LIME 解释分类模型针对一个背景为雪的哈士奇图片的分类结果，发现分类模型将该图片错误地分类为冰原狼，而解释方法给出的解释结果表明模型做出决策的依据是图片背景中的雪（如图 8(a)所示）。很显然，该解释结果与人类的认知相违背，表明模型在学习的过程中错误地将雪作为冰原狼的决策特征，从而证明该模型是不可靠的。类似地，Lapuschkin 等人[117]利用 LRP 解释方法定性地分析一个从 ImageNet 中迁移训练得到的 CNN 模型和一个在 PASCAL VOC 2007 数据集上训练得到的 Fisher 向量（FV）分类器的决策结果，以检测训练数据中的潜在缺陷和偏差。研究结果表明，尽管两个模型具有相似的分类精度，但在对输入样本进行分类时却采用了完全不同的分类策略。从 LRP 解释方法给出的解释结果可以看出（如图 8(b)所示），在对轮船图片进行分类时，FV 分类器依据的是海水特征，而 CNN 模型则能正确地捕获到轮船的轮廓信息。与此同时，如果将位于水外的轮船作为测试样本，FV 分类器的分类性能将大幅下降，而 CNN 模型则几乎不受影响。这一验证结果表明，FV 分类器的决策行为存在偏差而 CNN 模型表现正常。因此，我们认为 CNN 模型比 FV 分类器更可靠，在进行模型选择时，我们将会选择 CNN 模型作为最终的分类模型。

而对于可解释方法所识别出的不可靠的模型，我们则可以采取相应的对策来进行改进。比如说，我们可以通过在训练模型时引入归纳偏置，提高模型在预测阶段的泛化能力，从而使其能对未知样本做出正确的决策。我们也可以通过修正训练集分布，消除数据中存在的偏差，并利用修正后的数据集重训练模型达到消除模型决策偏差的目的。

4.2 模型诊断

由于机器学习模型内部工作机制复杂、透明性低，模型开发人员往往缺乏可靠的推理或依据来辅助他们进行模型开发和调试，因而使得模型开发迭代过程变得更加耗时且容易出错。而模型可解释性相关技术作为一种细粒度分析和解释模型的有效手段，可用于分析和调试模型的错误决策行为，以“诊断”模型中存在的缺陷，并为修复模型中的缺陷提供有力的支撑。近年来，随着模

型可解释性研究不断取得新的突破，基于可解释性的机器学习模型诊断相关研究也吸引了越来越多的关注[118-121]。

研究表明，基于模型特征表示可视化以及中间层分析的解释方法（如激活最大化、特征反演等）可以有效地用于解释和诊断复杂模型。典型的解决方案包括可视化模型的中间激活状态或内部特征表示以及可视化模型中的数据流图[122-124]，以增强对复杂模型的解释和理解，同时分析和评估模型或算法的性能，为在模型开发的不同阶段（如前期特征工程、中期超参调整以及后期模型微调等）交互式改进模型提供有效的指导[125]。此外，一些其他的研究方法则通过识别与模型“漏洞”相关的重要特征或实例来进行模型诊断和调试。Krause 等人[126] 基于敏感性分析解释方法的思想，设计了一个名为 Prospector 的系统，通过修改特征值并检查预测结果的相应变化来确定敏感性特征。Cadamuro 等人[118]提出了一种概念分析和诊断循环的模型诊断方法，允许终端用户迭代地检测模型“漏洞”，以找到对模型“漏洞”贡献最大的训练实例，从而确定模型出错的根本原因。Krause 等人[127]提出了一个可视化模型诊断工作流，通过利用局部解释方法度量输入实例中的局部特征相关性，以帮助数据科学家和领域专家理解和诊断模型所做出的决策。具体地，该工作流首先利用聚合统计查看数据在正确决策和错误决策之间的分布；然后，基于解释方法理解用于做出这些决策的特征；最后基于原始数据，对影响模型决策的潜在根本原因进行深入分析。

针对已发现的模型“漏洞”，我们可以基于模型诊断方法给出的推理结果，采取相应的措施对模型进行“治疗”，如提高训练数据的质量、选择可靠特征以及调整模型超参等。Paiva 等人[128]提出了一种可视化数据分类方法，该方法通过点布局策略实现数据集的可视化，允许用户选择并指定用于模型学习过程的训练数据，从而提高训练集的整体质量。Brooks 等人[129]提出了一个用于改进特征工程的交互式可视化分析系统，该系统支持错误驱动的特征构思过程并为误分类样本提供交互式可视化摘要，允许在误分类样本和正确分类样本之间进行特征级别的比较，以选择能减小模型预测错误率的特征，从而提高模型性能并修复模型中的“漏洞”。

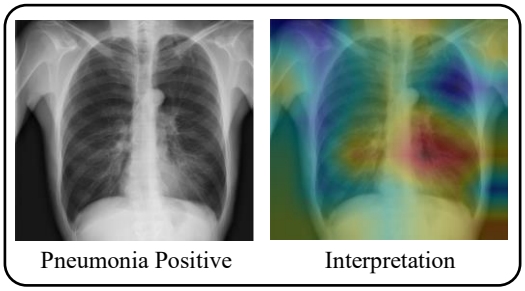


Fig.9 Application of interpretation in medical diagnosis

图 9 可解释方法在医疗诊断中的应用[6]

4.3 辅助分析

除了用于模型验证与模型诊断之外，可解释性相关技术还可用于辅助分析与决策，以提高人工分析和决策的效率。相关研究表明，基于可解释性的辅助分析技术在医疗数据分析、分子模拟以及基因分析等多个领域取得了巨大的成功，有效地解决了人工分析耗时费力的难题。

在智慧医疗领域，许多学者尝试将深度学习及可解释性技术应用于构建自动化智能诊断系统，以辅助医护人员分析病人的医疗诊断数据，从而提高人工诊断的效率[6; 130]。Rajpurkar 等人[6]基于大规模病人胸片数据开发了基于深度学习的肺炎检测系统（CheXNet），其检测性能甚至超过了放射科医师的诊断水平，该系统通过将可解释方法 CAM 应用于解释检测系统的决策依据并可视化对应的解释结果（如图 9 所示），可以为医师分析病人医疗影像数据以快速定位病人的病灶提供大量的辅助信息。Arvaniti 等人[130]研究结果表明，在给定一个良好标注的数据集的前提下，可以利用 CNN 模型成功地实现对前列腺癌组织微阵列的自动格里森分级。同时，利用解释方法给出自动分级系统的分级依据，可实现病理专家级的分级效果，从而为简化相对繁琐的分级任务提供了支撑。

在量子化学领域，分子动力学模拟是理解化学反应机理、速率和产率的关键，然而由于分子的完整波函数相对复杂，且难以计

算和近似，导致人们通常难以理解，因而如何创建人类可解释的分子表示成为 21 世纪物质模拟的一大挑战[131]。为解决这一难题，许多学者将机器学习及可解释性技术引入到分子模拟任务中，用于辅助分析分子结构与分子性质之间的关系[132-134]。其中，Schütt 等人[134]提出一种通过结合强大的结构和表示能力以实现较高预测性能和良好可解释性的深度张量神经网络（DTNN），用于预测分子结构与电子性质之间的关系。同时，作者利用基于测试电荷扰动的敏感性分析方法测量在给定的位置插入电荷对 DTNN 输出结果的影响，从而找到与解释分子结构与性质关系最相关的每个单独的分子空间结构。Häse 等人[133]提出一种利用机器学习来辅助分子动力学模拟的方法，该方法利用模拟产生的大量数据训练贝叶斯神经网络（BNN）来预测 1,2-二氧杂环丁烷从初始核位置的离解时间。为了构建一个可解释的 BNN 模型，作者将模型的权重和偏置分布参数化为拉普拉斯分布，以确定与准确预测离解时间以及实际的物理过程相关的输入特征。研究结果表明，该方法不仅可以准确地再现化合物的离解过程，而且能自动地从模拟数据中提取相关信息，而不需要预先了解相关化学反应。同时，通过解释 BNN 所捕获的特征与实际物理过程之间的相关关系，可以在不了解电子结构的情况下，确定核坐标与离解时间之间的物理相关性，从而为人们在化学领域取得概念性的突破提供灵感。

在基因组分析领域，由基因组学研究不断进步而产生的数据爆炸，给传统的基因组分析方法带来了巨大的挑战，同时也给数据驱动的深度学习技术在基因组分析研究中的发展和应用带来了机遇[135]。相关研究表明，深度学习在基因组分析中的应用已突显出了其强大的优势[136-139]。然而，人们期望深度学习模型不仅能成功地预测结果，还能识别有意义的基因序列，并对所研究的科学问题（如基因与疾病、药物之间的关系）提供进一步的见解，因而模型的可解释性在应用中显得至关重要。Lanchantin 等人[138]将三种 DNN 模型（即 CNN、RNN 以及 CNN-RNN）应用于预测给定的 DNA 序列中某一特定的转录因子是否有结合位点，并且提出了一套基于解释方法的可视化策略，用于解释对应的预测模型并从中提取隐含的序列模式。其中，作者基于反向传播解释方法，通过计算预测概率相对于输入 DNA 序列的梯度来构建显著图[73]，用于度量并显示核苷酸的重要性。同时，作者利用时间域输出分值来识别 DNN 序列中与特定转录因子结合位点相关的关键序列位置，并利用类激活最大化方法生成与特定预测结果相关的 Motif 模式。实验结果证明，这一系列的可视化策略可为研究人员分析 DNA 序列结构、组成成分与特定转录因子结合位点之间的关系提供大量的辅助信息。类似地，Alipanahi 等人[139]构建了一个名为 DeepBind 的系统，通过训练一个 CNN 模型将 DNA 和 RNA 序列映射到蛋白质结合位点上，以了解 DNA 和 RNA 结合蛋白的序列特异性。为了进一步探索遗传变异对蛋白质结合位点的影响，作者采用了基于扰动的敏感性分析方法，通过计算突变对 DeepBind 预测结果的影响生成“突变图”，以解释序列中每个可能的点突破对结合亲和力的影响。作者表明，DeepBind 可用于揭示 RNA 结合蛋白质在选择性剪接中的调节作用，并辅助研究人员分析、识别、分组及可视化可影响转录因子结合和基因表达的疾病相关遗传变异，从而有望实现精准医学。

4.4 知识发现

近年来，随着人工智能相关技术的发展，基于机器学习的自动决策系统被广泛地应用到各个领域，如恶意程序分析、自动化医疗诊断以及量化交易等。然而，由于实际任务的复杂性以及人类认知和领域知识的局限性，人们可能无法理解决策系统给出的结果，因而缺乏对相关领域问题更深入的理解，进而导致许多科学问题难以得到有效的解决。最新研究成果表明，通过将可解释性相关技术与基于机器学习的自动决策系统相结合，可有效地挖掘出自动决策系统从数据中学到的新知识，以提供对所研究科学问题的深入理解，从而弥补人类认知与领域知识的局限性。

在二进制分析领域，许多潜在的启发式方法都是针对某一个特定的函数的，而挖掘这些潜在的方法通常需要丰富的领域知识，因而很难通过人工的方式对所有的启发式方法进行汇总。Guo 等人[99]将可解释方法 LEMNA 应用于一个基于 LSTM 的二进制函数入口检测器，以提供对 LSTM 检测结果的解释。通过分析解释结果，作者发现检测模型确实从训练数据中学到了用于识别函数入口的潜在特征，这表明利用 LEMNA 解释方法可以挖掘出检测模型从数据中学到的新知识，从而对总结针对某个特殊函数的所有潜在的启发式方法提供帮助。

在医疗保健领域，由于病人病理错综复杂且因人而异，医护人员往往无法通过有限的医疗诊断知识挖掘潜在的致病因素及其之间的相互作用，而对潜在因素的忽视极其可能带来致命的威胁。Yang 等人[53]基于重症监护室（ICU）治疗记录数据构建了一个带注意力机制的 RNN 模型，用于分析医疗条件与 ICU 死亡率之间的关系，而这些关系在以往的医疗实践中往往没有得到很好的研究。作者研究结果表明，利用可解释性技术有助于发现与医疗保健中某些结果相关的潜在影响因素或相互作用，从而使得从自动化医疗诊断模型中学习新的诊断知识成为可能。

此外，作为知识发现的重要手段，模型可解释性及其相关解释方法还被广泛地应用到了数据挖掘领域，以从海量数据中自动地挖掘隐含的新知识[140-143]。这类研究核心思想是基于所研究的领域及科学目标构建海量数据集，然后对构建的数据集进行清洗并利用机器学习模型从清洗后的数据中提取数据映射模式，最后利用解释方法从挖掘到的数据模式识别代表新知识的模式并利用可视化技术将新知识呈现给用户。

5 可解释性与安全性分析

模型可解释性研究的初衷是通过构建可解释的模型或设计解释方法提高模型的透明性，同时验证和评估模型决策行为和决策结果的可靠性和安全性，消除模型在实际部署应用中的安全隐患。然而，模型可解释性相关技术同样可以被攻击者利用以探测机器学习模型中的“漏洞”，因而会给机器学习模型以及真实应用场景中尤其是风险敏感场景中的机器学习应用带来威胁。此外，由于解释方法与待解释模型之间可能存在不一致性，因而可解释系统或可解释方法本身就存在一定的安全风险。

5.1 安全隐患消除

如第 4 节中所述，模型可解释性及相关解释方法不仅可以用于评估和验证机器学习模型，以弥补传统模型验证方法的不足，保证模型决策行为和决策结果的可靠性和安全性，还可用于辅助模型开发人员和安全分析师诊断和调试模型以检测模型中的缺陷，并为安全分析师修复模型“漏洞”提供指导，从而消除模型在实际部署应用中的安全隐患。并且，通过同时向终端用户提供模型的预测结果及对应的解释结果，可提高模型决策的透明性，进而有助于建立终端用户与决策系统之间的信任关系。

除了用于消除上述内在安全隐患之外，模型可解释性相关技术还可以帮助抵御外在安全风险。人工智能安全领域相关研究表明即使决策“可靠”的机器学习模型也同样容易受到对抗样本攻击，只需要在输入样本中添加精心构造的、人眼不可察觉的扰动就可以轻松地让模型决策出错[8; 144; 145]。这种攻击危害性大，隐蔽性强，变种多且难以防御，严重地威胁着人工智能系统的安全。而现存防御方法大多数是针对某一个特定的对抗样本攻击设计的静态的经验性防御，因而防御能力极其有限。然而，不管是哪种攻击方法，其本质思想都是通过向输入中添加扰动以转移模型的决策注意力，最终使模型决策出错。由于这种攻击使得模型决策依据发生变化，因而解释方法针对对抗样本的解释结果必然与其针对对应的正常样本的解释结果不同。因此，我们可以通过对比并利用这种解释结果的反差来检测对抗样本，而这种方法并不特定于某一种对抗攻击，因而可以弥补传统经验性防御的不足。

除上述防御方法外，很多学者从不同的角度提出了一些新的基于可解释性技术的

对抗防御方法。其中，Tao 等人[146]认为对抗攻击与模型的可解释性密切相关，即对于正常样本的决策结果，我们可以基于人类可感知的特征或属性来进行推理，而对于对抗样本的决策结果我们则通常无法解释。基于这一认知，作者提出一种针对人脸识别模型的对抗样本检测方法，该方法首先利用敏感性分析解释方法识别与人类可感知属性相对应的神经元，称之为“属性见证”神经元；然后，通过加强见证神经元同时削弱其他神经元将原始模型转换为属性导向模型，对于正常样本，属性导向模型的预测结果与原始模型一致，对于对抗样本二者预测结果则不一致；最后，利用两个模型预测结果的不一致性来检测对抗样本，实现对对抗攻击的防御。Liu 等人[147]则基于对分类模型的解释，提出了一种新的对抗样本检测框架。给定一个恶意样本检测器，该框架首先选择一个以确定为恶意样本的样本子集作为种子样本，然后构建一个局部解释器解释种子样本被分类器视为恶意样本的原因，并通过朝着解释器确定的规避方向来扰动每一个种子样本的方式产生对抗样本。最后，通过利用原始数据和生成的对抗样本对检测器进行对抗训练，以提高检测器对对抗样本的鲁棒性，从而降低模型的外在安全风险。

5.2 安全威胁

尽管可解释性技术是为保证模型可靠性和安全性而设计的，但其同样可以被恶意用户滥用而给实际部署应用的机器学习系统带来安全威胁。比如说，攻击者可以利用解释方法探测能触发模型崩溃的模型漏洞，在对抗攻击中，攻击者还可以利用可解释方法探测模型的决策弱点或决策逻辑，从而为设计更强大的攻击提供详细的信息。在本文中，我们将以对抗攻击为例，阐述可解释性技术可能带来的安全风险。

在白盒对抗攻击中，攻击者可以获取目标模型的结构、参数信息，因而可以利用反向传播解释方法的思想来探测模型的弱点[154]。其中，Goodfellow 等人[144]提出了快速梯度符号攻击方法（FGSM），通过计算模型输出相对于输入样本的梯度信息来探测模型的敏感性，并通过朝着敏感方向添加一个固定规模的噪音来生成对抗样本。Papernot 等人[148]基于 Grad[73]解释方法提出了雅可比显著图攻击（JSMA），该攻击方法首先利用 Grad 解释方法生成显著图，然后基于选择图来选择最重要的特征进行攻击。利用 Grad 方法提供的特征重要性信息，JMSA 攻击只需要扰动少量的特征就能达到很高的攻击成功率，因而攻击的隐蔽性更强。对于黑盒对抗攻击，由于无法获取模型的结构信息，只能操纵模型的输入和输出[153]，因而攻击者可以利用模型无关解释方法的思想来设计攻击方法。其中，Papernot 等人[149]提出了一种针对黑盒机器学习模型的替代模型攻击方法。该方法首先利用模型蒸馏解释方法的思想训练一个替代模型来拟合目标黑盒模型的决策结果，以完成从黑盒模型到替代模型的知识迁移过程；然后，利用已有的攻击方法针对替代模型生成对抗样本；最后，利用生成的对抗样本对黑盒模型进行迁移攻击。Li 等人[9]提出了一种基于敏感性分析解释方法的文本对抗攻击方法（TextBugger），用于攻击真实场景中的情感分析模型和垃圾文本检测器。该方法首先通过观察去掉某个词前后模型决策结果的变化来定位文本中的重要单词，然后通过利用符合人类感知的噪音逐个扰动重要的单词直到达到攻击目标。该研究表明，利用 TextBugger 攻击方法可以轻松的攻破 Google Cloud、Microsoft Azure、Amazon AWS、IBM Watson、Facebook fastText 等平台提供的商业自然语言处理机器学习服务，并且攻击成功率高、隐蔽性强。

5.3 自身安全问题

由于采用了近似处理或是基于优化手段，大多数解释方法只能提供近似的解释，因而解释结果与模型的真实行为之间存在一定的不一致性。而最新研究表明，攻击者可以利用解释方法与待解释模型之间的这种不一致性设计针对可解释系统的新型对抗样本攻击，因而严重的威胁着可解释系统的自身安全。

根据攻击目的不同，现存针对可解释系统的新型对抗样本攻击可以分为两类，其中，

一类是在不改变模型的决策结果的前提下，使解释方法解释出错[150]；另一类是使模型决策出错而不改变解释方法的解释结果[151]。其中，Ghorbani 等人[150]首次将对抗攻击的概念引入到了神经网络的可解释性中并且提出了模型解释脆弱性的概念。具体地，作者将针对解释方法的对抗攻击定义为如下优化问题：

$$\arg\max_{\delta} D(I(x_t;N), I(x_t+\delta;N))$$
$$s.t.\|\delta\|_{\infty} \le \varepsilon,\ f(x_t+\delta) = f(x_t)$$

其中，$\boldsymbol{I}(\boldsymbol{x_t};\mathcal{N})$为解释系统对神经网络$\mathcal{N}$针对样本$\boldsymbol{x_t}$决策结果$f(\boldsymbol{x_t})$的解释，$\boldsymbol{\delta}$为样本中所需添加的扰动，$\mathcal{D}(\cdot)$用于度量扰动前后解释结果的变化。通过优化上述目标函数，可以在不改变模型决策结果的前提下，生成能让解释方法产生截然不同的解释结果的对抗样本。针对 Grad[73]、Integrated[101]以及 DeepLIFT[106]等反向传播解释方法的对抗攻击实验证明，上述解释方法均容易受到对抗样本攻击，因而只能提供脆弱的模型解释。与 Ghorbani 等人研究相反，Zhang 等人[151]提出了 Acid 攻击，旨在生成能让模型分类出错而不改变解释方法解释结果的对抗样本。通过对表示导向的（如激活最大化、特征反演等）、模型导向的（如基于掩码模型的显著性检测等[152]）以及扰动导向的（如敏感性分析等）三大类解释方法进行 Acid 攻击和经验性评估，作者发现生成欺骗分类器及其解释方法的对抗样本实际上并不比生成仅能欺骗分类器的对抗样本更困难。因此，这几类解释方法同样是脆弱的，在对抗的环境下，其提供的解释结果未必可靠。此外，这种攻击还会使基于对比攻击前后解释结果的防御方法失效，导致对抗攻击更难防御。

上述研究表明，现存解释方法大多数是脆弱的，因此只能提供有限的安全保证。但由于可解释性技术潜在应用广泛，因而其自身安全问题不容忽视。以医疗诊断中的可解释系统为例，在临床治疗中，医生会根据可解释系统提供的解释结果对病人进行相应的诊断和治疗，一旦解释系统被新型对抗攻击方法攻击，那么提供的解释结果必然会影响医生的诊断过程，甚至是误导医生的诊断而给病人带来致命的威胁。因此，仅有解释是不够的，为保证机器学习及可解释性技术在实际部署应用中的安全，解释方法本身必须是安全的，而设计更精确的解释方法以消除解释方法与决策系统之间的不一致性则是提高解释方法鲁棒性进而消除其外在安全隐患的重要途径。

6 当前挑战与未来方向

尽管模型可解释性研究已取得一系列瞩目的研究成果，但其研究还处于初级阶段，依然面临着许多的挑战且存在许多的关键问题尚待解决。其中，可解释性研究当前面临的一个挑战是如何设计更精确、更友好的解释方法，消除解释结果与模型真实行为之间的不一致。第二个挑战是如何设计更科学、更统一的可解释性评估指标，以评估可解释方法解释性能和安全性。

6.1 解释方法设计

精确地理解机器学习的工作原理，研究透明的、可解释且可证明机器学习技术，有助于推动机器学习研究的进一步发展，同时有助于促进人工智能相关技术的落地应用。这要求机器学习可解释性研究必须具备能精确地揭示模型内部工作逻辑同时向人类提供可以足够准确理解模型决策的信息的能力。因此，无论是 ante-hoc 可解释性还是 post-hoc 可解释性，我们所设计的解释方法都必须是精确的，我们的解释方法提供的解释结果都必须忠实于模型的真实决策行为。

由于模型的决策准确性与模型自身可解释性之间存在一个权衡，现有关于 ante-hoc 可解释性的研究多局限于诸如线性回归、决策树等算法透明、结构简单的模型，对于复杂的 DNN 模型则只能依赖于注意力机制提供一个粗粒度的解释。因此，如何设计可解释的机器学习模型以消除模型准确性与可解释性之间的制约是 ante-hoc 可解释性研究所面临的一大挑战，也是未来可解释性研究发展的一个重要趋势。其中，一种直观的方法是将机器学习与因果模型相结合，让机器学习系统具备从观察数据中发现事物间的因果结构和定量推断的能力。同时，我们

还可以将机器学习与常识推理和类比计算等技术相结合，形成可解释的、能自动推理的学习系统。未来我们还可以考虑利用仿生学知识并结合更先进的认知理论对人类认知建模，以设计具备人类自我解释能力的机器学习模型，实现具有一定思维能力并且能自我推理自我解释的强人工智能系统。

对于 post-hoc 可解释性而言，大多数的研究都在尝试采用近似的方法来模拟模型的决策行为，以从全局的角度解释模型的整体决策逻辑或者从局部的角度解释模型的单个决策结果。然而，由于近似过程往往不够精确，解释方法给出的解释结果无法正确地反映待解释模型的实际运行状态和真实决策行为，而解释方法与决策模型之间的这种不一致性甚至严重地威胁着可解释系统自身的安全。因此，当前 post-hoc 可解释性相关研究面临的巨大挑战是如何设计忠实于决策模型的安全可保障的精确解释方法，以消除解释结果与模型真实行为之间的不一致性，从而保证解释结果的可靠性和安全性。未来一个有前景的潜在研究方向是设计数学上与待解释模型等价的解释方法或解释模型。对于全连接神经网络，Chu 等人[116]已经给出了相应的研究方法并取得了一定的研究成果，我们则可以基于具体模型的内部机理和神经网络的前向传播机制，将 Chu 等人提出的研究方法扩展到 CNN、RNN 等更复杂神经网络模型，从而实现对复杂模型的精确解释。

6.2 解释方法评估

目前，可解释性研究领域缺乏一个用于评估解释方法的科学评估体系，尤其是在计算机视觉领域，许多解释方法的评估还依赖于人类的认知，因而只能定性评估，无法对解释方法的性能进行量化，也无法对同类型的研究工作进行精确地比较。并且，由于人类认知的局限性，人们只能理解解释结果中揭示的显性知识，而通常无法理解其隐性知识，因而无法保证基于认知的评估方法的可靠性。

对于 ante-hoc 可解释性而言，其评估挑战在于如何量化模型的内在解释能力。对于同一应用场景，我们可能会采用不同的模型，同一模型也可能会应用到不同的场景中，而对于如何衡量和比较这些模型的可解释性目前仍没有达成共识。由于模型自身可解释性受实际应用场景、模型算法本身以及人类理解能力的制约，未来我们可以从应用场景、算法功能、人类认知这三个角度来设计评估指标。这些指标虽各有利弊但相互补充，可以实现多层次、细粒度的可解释性评估，以弥补单一评估指标的不足。

对于 post-hoc 可解释性而言，其评估挑战在于如何量化解释结果的保真度和一致性。如前所述，由于人类认知的局限性，解释方法针对机器学习模型给出的解释结果并不总是“合理”的，而我们很难判断这种与人类认知相违背的解释结果到底是由于模型自身的错误行为还是解释方法的局限性，抑或是人类认知的局限性造成的。因此，我们需要设计可靠的评估指标对解释方法进行定量的评估。Guo 等人[99]提出利用解释方法给出的预测结果与待解释模型预测结果之间的均方根误差（RMSE）来评估解释方法的保真度，然而这种评估指标无法用于评估激活最大化、敏感性分析、反向传播以及特征反演等不提供预测结果的解释方法。Chu 等人[116]提出利用输入样本及其邻近样本的解释结果的余弦相似性来评估解释方法，然而这种方法无法用于评估解释结果的保真度。此外，目前还缺乏用于评估针对同一模型的不同解释方法的评估指标。因此，未来我们需要从解释结果的保真度、一致性以及不同解释方法的差异性等角度设计评价指标，对解释方法进行综合评估。

7 结束语

机器学习可解释性是一个非常有前景的研究领域，该领域已经成为了国内外学者的研究热点，并且取得了许多瞩目的研究成果。但到目前为止，机器学习可解释性研究还处于初级阶段，依然存在许多关键问题尚待解决。为了总结现有研究成果的优势与不足，探讨未来研究方向，本文从可解释性相关技术、潜在应用、安全性分析等方面对现有研究成果进行了归类、总结和分析，同时

讨论了当前研究面临的挑战和未来潜在的研究方向，旨在为推动模型可解释性研究的进一步发展和应用提供一定帮助。

参考文献

[1] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 815-823

[2] Sun Yi, Liang Ding, Wang Xiaogang, et al. Deepid3: Face recognition with very deep neural networks [J]. arXiv preprint arXiv:1502.00873, 2015

[3] Taigman Y, Yang Ming, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 1701-1708

[4] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 3354-3361

[5] Tobiyama S, Yamaguchi Y, Shimada H, et al. Malware detection with deep neural network using process behavior [C] //Proc of the 40th Annual Computer Software and Applications Conf. Piscataway, NJ: IEEE, 2016: 577-582

[6] Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning [J]. arXiv preprint arXiv:1711.05225, 2017

[7] Ibrahim M, Louie M, Modarres C, et al. Global Explanations of Neural Networks: Mapping the Landscape of Predictions [J]. arXiv preprint arXiv:1902.02384, 2019

[8] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [C] //Proc of the 2nd Int Conf on Learning Representations, 2014.

[9] Li Jinfeng, Ji Shouling, Du Tianyu, et al. TextBugger: Generating Adversarial Text Against Real-world Applications [C] //Proc of 26th Annual Network and Distributed Systems Security Symp. Reston, VA: ISOC, 2019

[10] Du Tianyu, Ji Shouling, Li Jinfeng, et al. SirenAttack: Generating Adversarial Audio for End-to-End Acoustic Systems [J]. arXiv preprint arXiv:1901.07846, 2019

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

[12] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models [J]. ACM Computing Surveys, 2018, 51(5): 93

[13] Ribeiro M T, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144

[14] Baehrens D, Schroeter T, Harmeling S, et al. How to explain individual classification decisions [J]. Journal of Machine Learning Research, 2010, 11(Jun): 1803-1831

[15] Melis D A, Jaakkola T. Towards robust interpretability with self-explaining neural networks [C] //Proc of the 32nd Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 2018: 7775-7784

[16] Poulin B, Eisner R, Szafron D, et al. Visual explanation of evidence with additive classifiers [C] //Proc of the 18th Conf on Innovative Applications of Artificial Intelligence. Palo Alto, CA: AAAI Press, 2006: 1822-1829

[17] Kononenko I. An efficient explanation of individual classifications using game theory [J]. Journal of Machine Learning Research, 2010, 11(Jan): 1-18

[18] Haufe S, Meinecke F, Görgen K, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging [J]. NeuroImage, 2014, 87: 96-110

[19] Huysmans J, Dejaeger K, Mues C, et al. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models [J]. Decision Support Systems, 2011, 51(1): 141-154

[20] Breslow L A, Aha D W. Simplifying decision trees: A survey [J]. The Knowledge Engineering Review, 1997, 12(1): 1-40

[21] Frank E, Witten I H. Generating accurate rule sets without global optimization [C] //Proc. of the 15th Int Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc, 1998: 144-151

[22] Quinlan J R. Generating production rules from decision trees [C] //Proc of the 10th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1987: 304-307

[23] Deng Houtao. Interpreting tree ensembles with intrees [J]. Int Journal of Data Science and Analytics, 2019, 7(4): 277-287

[24] Lou Yin, Caruana R, Gehrke J. Intelligible models for classification and regression [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data mining. New York: ACM, 2012: 150-158

[25] Henson R K. The Logic and Interpretation of Structure

Coefficients in Multivariate General Linear Model Analyses [R]. New Orleans, LA: ERIC Document Reproduction Service No. ED467381, 2002

[26] Hastie T J. Statistical models in S [M]. New York: Routledge, 2017: 249-307

[27] Wood S N. Generalized additive models: an introduction with R [M]. New York: Chapman and Hall/CRC, 2017

[28] Ravikumar P, Lafferty J, Liu Han, et al. Sparse additive models [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2009, 71(5): 1009-1030

[29] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C] //Proc of the 3rd Int Conf on Learning Representations, 2015

[30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Proc of the 31st Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 2017: 6000-6010

[31] Choi E, Bahadori M T, Sun Jimeng, et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism [C] //Proc of the 30th Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 2016: 3504-3512

[32] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention [C] //Proc of the 32nd Int Conf on Machine Learning. Tahoe City, CA: International Machine Learning Society, 2015: 2048-2057

[33] Chaudhari S, Polatkan G, Ramanath R, et al. An Attentive Survey of Attention Models [J]. arXiv:1904.02874, 2019

[34] Yang Zichao, Yang Diyi, Dyer C, et al. Hierarchical attention networks for document classification [C] //Proc of the 2016 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg PA: Association for Computational Linguistics, 2016: 1480-1489

[35] He Xiangnan, He Zhankui, Song Jingkuan, et al. NAIS: Neural attentive item similarity model for recommendation [J]. IEEE Trans on Knowledge and Data Engineering, 2018, 30(12): 2354-2366

[36] Ying Haochao, Zhuang Fuzheng, Zhang Fuzheng, et al. Sequential recommender system based on hierarchical attention networks [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2018

[37] Zhou Chang, Bai Jinze, Song Junshuai, et al. ATRank: An attention-based user behavior modeling framework for recommendation [C] //Proc of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018

[38] Yu Shuai, Wang Yongbo, Yang Min, et al. NAIRS: A Neural Attentive Interpretable Recommendation System [C] //Proc of the 12th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2019: 790-793

[39] Seo S, Huang Jing, Yang Hao, et al. Interpretable convolutional neural networks with dual local and global attention for review rating prediction [C] //Proc of the 11th ACM Conference on Recommender Systems. New York: ACM, 2017: 297-305

[40] Andrews R, Diederich J, Tickle A B. Survey and critique of techniques for extracting rules from trained artificial neural networks [J]. Knowledge-based systems, 1995, 8(6): 373-389

[41] Tickle A B, Andrews R, Golea M, et al. The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks [J]. IEEE Trans on Neural Networks, 1998, 9(6): 1057-1068

[42] Tickle A B, Orlowski M, Diederich J. DEDEC: A methodology for extracting rules from trained artificial neural networks [C] //Proc of the AISB'96 Workshop on Rule Extraction from Trained Neural Networks. The Netherlands: IOS Press Amsterdam,, 1996: 90-102

[43] Fu Limin. Rule generation from neural networks [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1994, 24(8): 1114-1124

[44] Bastani O, Kim C, Bastani H. Interpreting blackbox models via model extraction [J]. arXiv preprint arXiv:1705.08504, 2017

[45] Craven M W. Extracting comprehensible models from trained neural networks [D]. Madison, WI: University of Wisconsin-Madison Department of Computer Sciences, 1996

[46] Boz O. Extracting decision trees from trained neural networks[C] //Proc of the 8th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 456-461

[47] Mashayekhi M, Gras R. Rule extraction from decision trees ensembles: new algorithms based on heuristic search and sparse group lasso methods [J]. Int Journal of Information Technology & Decision Making, 2017, 16(06): 1707-1727

[48] Mashayekhi M, Gras R. Rule extraction from random forest: the RF+HC methods [G] // LNCS 9091: Proc of the 28th Canadian Conf on Artificial Intelligence. Berlin: Springer,

2015: 223-237

[49] Hara S, Hayashi K. Making tree ensembles interpretable: A Bayesian model selection approach [J]. arXiv preprint arXiv:1606.09066, 2016

[50] Hara S, Hayashi K. Making tree ensembles interpretable [J]. arXiv preprint arXiv:1606.05390, 2016

[51] Krishnan R. A Systematic Method for Decompositional Rule Extraction from from Neural Networks [C] //Proc of the 10th Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 1996

[52] Bondarenko A, Zmanovska T, Borisov A. Decompositional Rules Extraction Methods from Neural Networks [C] //Proc of the 16th Int Conf on Soft Computing. Berlin: Springer, 2010: 256-262

[53] Yang Chengliang, Rangarajan A, Ranka S. Global model interpretation via recursive partitioning [C] //Proc of the IEEE 4th Int Conf on Data Science and Systems. Piscataway, NJ: IEEE, 2018: 1563-1570

[54] Craven M W, Shavlik J W. Using sampling and queries to extract rules from trained neural networks [C] //Proc of the 8th Int Conf on Machine Learning. Tahoe City, CA: International Machine Learning Society, 1994: 37-45

[55] Zhou Zhihua, Jiang Yuan, Chen Shifu. Extracting symbolic rules from trained neural network ensembles [J]. AI Communications, 2003, 16(1): 3-15

[56] De Fortuny E J, Martens D. Active learning-based pedagogical rule extraction [J]. IEEE Trans on Neural Networks and Learning Systems, 2015, 26(11): 2664-2677

[57] Lakkaraju H, Kamar E, Caruana R, et al. Interpretable & explorable approximations of black box models [J]. arXiv preprint arXiv:1707.01154, 2017

[58] Puri N, Gupta P, Agarwal P, et al. MAGIX: Model Agnostic Globally Interpretable Explanations [J]. arXiv preprint arXiv:1706.07160, 2017

[59] Liu Xuan, Wang Xiaoguang, Matwin S. Improving the Interpretability of Deep Neural Networks with Knowledge Distillation[C] //Proc of the IEEE Int Conf on Data Mining Workshops, Piscataway, NJ: IEEE, 2018: 905-912

[60] Buciluǎ C, Caruana R, Niculescu-Mizil A. Model compression [C] //Proc of the 12th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, 2006: 535-541

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

[62] Craven M, Shavlik J W. Extracting tree-structured representations of trained networks [C] //Proc of the 10th Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 1996: 24-30

[63] Frosst N, Hinton G. Distilling a neural network into a soft decision tree [J]. arXiv preprint arXiv:1711.09784, 2017

[64] Tan S, Caruana R, Hooker G, et al. Learning Global Additive Explanations for Neural Nets Using Model Distillation [J]. arXiv preprint arXiv:1801.08640, 2018

[65] Che Z, Purushotham S, Khemani R, et al. Interpretable deep models for ICU outcome prediction [C] //Proc of the AMIA Annual Symp. Bethesda, MD: American Medical Informatics Association, 2016: 371

[66] Ding T, Hasan F, Bickel W K, et al. Interpreting Social Media-Based Substance Use Prediction Models with Knowledge Distillation [C] //Proc of the IEEE 30th Int Conf on Tools with Artificial Intelligence. Piscataway, NJ: IEEE, 2018: 623-630

[67] Xu K, Park D H, Yi C, et al. Interpreting Deep Classifier by Visual Distillation of Dark Knowledge [J]. arXiv preprint arXiv:1803.04042, 2018

[68] Tan S, Caruana R, Hooker G, et al. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation [C] //Proc of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, 2018: 303-310

[69] Tan S, Caruana R, Hooker G, et al. Detecting bias in black-box models using transparent model distillation [J]. arXiv preprint arXiv:1710.06169, 2017

[70] Wang Junpeng, Gou Liang, Zhang Wei, et al. DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation [J]. IEEE Trans on Visualization and Computer Graphics, 2019, 25(6): 2168-2180

[71] Zhang Quanshi, Wang Wenguan, Zhu Songchun. Examining cnn representations with respect to dataset bias [C] //Proc of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018

[72] Zhang Quanshi, Wu Yingnian, Zhu Songchun. Interpretable convolutional neural networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8827-8836

[73] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps [J]. arXiv preprint arXiv:1312.6034, 2013

[74] Berkes P, Wiskott L. On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields [J]. Neural

Computation, 2006, 18(8): 1868-1895

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

[76] Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks [J]. Digital Signal Processing, 2018, 73: 1-15

[77] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 5188-5196

[78] Nguyen A, Clune J, Bengio Y, et al. Plug & play generative networks: Conditional iterative generation of images in latent space [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4467-4477

[79] Nguyen A, Dosovitskiy A, Yosinski J, et al. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks [C] //Proc of the 30th Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 2016: 3387-3395

[80] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C] //Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 818-833

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

[82] Yuan Hao, Chen Yongjun, Hu Xia, et al. Interpreting Deep Models for Text Analysis via Optimization and Regularization Methods [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019

[83] Saltelli A, Tarantola S, Campolongo F, et al. Sensitivity analysis in practice: a guide to assessing scientific models [M]. Hoboken NJ: John Wiley & Sons, 2004

[84] Chatterjee S, Hadi A S. Sensitivity analysis in linear regression [M]. Hoboken NJ: John Wiley & Sons, 2009

[85] Homma T, Saltelli A. Importance measures in global sensitivity analysis of nonlinear models [J]. Reliability Engineering & System Safety, 1996, 52(1): 1-17

[86] Saltelli A, Tarantola S, Chan K-S. A quantitative model-independent method for global sensitivity analysis of model output [J]. Technometrics, 1999, 41(1): 39-56

[87] Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models [J]. Ecological Modelling, 2003, 160(3): 249-264

[88] Khan J, Wei J S, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks [J]. Nature Medicine, 2001, 7(6): 673

[89] Engelbrecht A P, Cloete I, Zurada J M. Determining the significance of input parameters using sensitivity analysis [C] //Proc of the Int Workshop on Artificial Neural Networks. Berlin: Springer, 1995: 382-388

[90] Harrington P D B, Wan C. Sensitivity Analysis Applied to Artificial Neural Networks: What has my neural network actually learned? [J]. Anal. Chem, 1998, 70: 2983-2990

[91] Sung A. Ranking importance of input parameters of neural networks [J]. Expert Systems with Applications, 1998, 15(3-4): 405-411

[92] Robnik-Šikonja M, Kononenko I. Explaining classifications for individual instances [J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(5): 589-600

[93] Fong R C, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 3429-3437

[94] Liu Lingqiao, Wang Lei. What has my classifier learned? visualizing the classification rules of bag-of-feature model by support region detection [C] //Proc of the 2012 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 3586-3593

[95] Li Jiwei, Monroe W, Jurafsky D. Understanding neural networks through representation erasure [J]. arXiv preprint arXiv:1612.08220, 2016

[96] Guidotti R, Monreale A, Ruggieri S, et al. Local rule-based explanations of black box decision systems [J]. arXiv preprint arXiv:1805.10820, 2018

[97] Ribeiro M T, Singh S, Guestrin C. Nothing else matters: model-agnostic explanations by identifying prediction invariance [J]. arXiv preprint arXiv:1611.05817, 2016

[98] Ribeiro M T, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018

[99] Guo Wenbo, Mu Dongliang, Xu Jun, et al. Lemna: Explaining deep learning based security applications [C] //Proc of the 2018 ACM SIGSAC Conf on Computer and Communications

Security. New York: ACM, 2018: 364-379

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

[101] Sundararajan M, Taly A, Yan Qiqi. Gradients of counterfactuals [J]. arXiv preprint arXiv:1611.02639, 2016

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

[103] Du Mengnan, Liu Ninghao, Song Qingquan, et al. Towards explanation of dnn-based prediction with guided feature inversion [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1358-1367

[104] Landecker W, Thomure M D, Bettencourt L M, et al. Interpreting individual classifications of hierarchical networks [C] //Proc of the IEEE Symp on Computational Intelligence and Data Mining. Piscataway, NJ: IEEE, 2013: 32-38

[105] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation [J]. PloS one, 2015, 10(7): e0130140.

[106] Shrikumar A, Greenside P, Shcherbina A, et al. Not just a black box: Learning important features through propagating activation differences [J]. arXiv preprint arXiv:1605.01713, 2016.

[107] Ding Yanzhuo, Liu Yang, Luan Huanbo, et al. Visualizing and understanding neural machine translation [C] //Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2017: 1150-1159.

[108] Arras L, Horn F, Montavon G, et al. " What is relevant in a text document?": An interpretable machine learning approach [J]. PloS one, 2017, 12(8): e0181142.

[109] Carter S. Exploring Neural Networks with Activation Atlases. https://ai.googleblog.com/2019/03/exploring-neural-networks.html

[110] Dosovitskiy A, Brox T. Inverting visual representations with convolutional networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4829-4837

[111] Zhou Bolei, Khosla A, Lapedriza A, et al. Object detectors emerge in deep scene cnns [C] //Proc of the 3rd Int Conf on Learning Representations, 2015

[112] Zhou Bolei, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2921-2929

[113] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization [C] //Proc of the IEEE Intel Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 618-626

[114] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences [C] //Proc of the 34th Int Conf on Machine Learning. Tahoe City, CA: International Machine Learning Society, 2017: 3145-3153

[115] Gehr T, Mirman M, Drachsler-Cohen D, et al. Ai2: Safety and robustness certification of neural networks with abstract interpretation [C] //Proc of the 2018 IEEE Symposium on Security and Privacy. Piscataway, NJ: IEEE, 2018: 3-18

[116] Chu Lingyang, Hu Xia, Hu Juhua, et al. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1244-1253

[117] Lapuschkin S, Binder A, Montavon G, et al. Analyzing classifiers: Fisher vectors and deep neural networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2912-2920

[118] Cadamuro G, Gilad-Bachrach R, Zhu X. Debugging machine learning models [C] //Proc of the 33rd ICML Workshop on Reliable Machine Learning in the Wild. Tahoe City, CA: International Machine Learning Society, 2016

[119] Kulesza T, Burnett M, Wong W-K, et al. Principles of explanatory debugging to personalize interactive machine learning [C] //Proc of the 20th ACM Int Conf on Intelligent User Interfaces. New York: ACM, 2015: 126-137

[120] Kulesza T, Stumpf S, Burnett M, et al. Explanatory debugging: Supporting end-user debugging of machine-learned programs [C] //Proc of the IEEE Symp on Visual Languages and Human-Centric Computing. Piscataway, NJ: IEEE, 2010: 41-48

[121] Bastani O, Kim C, Bastani H. Interpretability via model extraction[J]. arXiv preprint arXiv:1706.09773, 2017

[122] Kahng M, Andrews P Y, Kalro A, et al. Activis: Visual exploration of industry-scale deep neural network models [J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 88-97

[123] Strobelt H, Gehrmann S, Pfister H, et al. Lstmvis: A tool for

visual analysis of hidden state dynamics in recurrent neural networks [J]. IEEE Trans on Visualization and Computer Graphics, 2018, 24(1): 667-676

[124] Wongsuphasawat K, Smilkov D, Wexler J, et al. Visualizing dataflow graphs of deep learning models in tensorflow [J]. IEEE Trans on Visualization and Computer Graphics, 2018, 24(1): 1-12

[125] Zhang Jiawei, Wang Yang, Molino P, et al. Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models [J]. IEEE Trans on Visualization and Computer Graphics, 2019, 25(1): 364-373

[126] Krause J, Perer A, Ng K. Interacting with predictions: Visual inspection of black-box machine learning models [C] //Proc of the ACM CHI Conf on Human Factors in Computing Systems. New York: ACM, 2016: 5686-5697

[127] Krause J, Dasgupta A, Swartz J, et al. A workflow for visual diagnostics of binary classifiers using instance-level explanations [C] //Proc of the IEEE Conf on Visual Analytics Science and Technology. Piscataway, NJ: IEEE, 2017: 162-172

[128] Paiva J G S, Schwartz W R, Pedrini H, et al. An approach to supporting incremental visual data classification [J]. IEEE Trans on Visualization and Computer Graphics, 2015, 21(1): 4-17

[129] Brooks M, Amershi S, Lee B, et al. FeatureInsight: Visual support for error-driven feature ideation in text classification [C] //Proc of the IEEE Conference on Visual Analytics Science and Technology. Piscataway, NJ: IEEE, 2015: 105-112

[130] Arvaniti E, Fricker K S, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning [J]. Scientific Reports, 2018, 8(1): 12054-12054

[131] Aspuru-Guzik A, Lindh R, Reiher M. The matter simulation (r) evolution [J]. ACS Central Science, 2018, 4(2): 144-152

[132] Boukouvalas Z, Elton D C, Chung P W, et al. Independent Vector Analysis for Data Fusion Prior to Molecular Property Prediction with Machine Learning [J]. arXiv preprint arXiv:1811.00628, 2018

[133] Häse F, Galván I F, Aspuru-Guzik A, et al. How machine learning can assist the interpretation of ab initio molecular dynamics simulations and conceptual understanding of chemistry [J]. Chemical Science, 2019, 10(8): 2298-2307

[134] Schütt K T, Arbabzadah F, Chmiela S, et al. Quantum-chemical insights from deep tensor neural networks [J]. Nature Communications, 2017, 8: 13890

[135] Yue Tianwei, Wang Haohan. Deep learning for genomics: A concise overview [J]. arXiv preprint arXiv:1802.00810, 2018

[136] Angermueller C, Lee H J, Reik W, et al. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning [J]. Genome biology, 2017, 18(1): 67

[137] Quang D, Xie Xiaohui. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences [J]. Nucleic acids research, 2016, 44(11): 107

[138] Lanchantin J, Singh R, Wang B, et al. Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks [C] //Proc of the 22nd Pacific Symp on Biocomputing. Singapore: World Scientific Publishing Co., 2017: 254-265

[139] Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning [J]. Nature biotechnology, 2015, 33(8): 831

[140] Gulia N, Singh S, Sapra L. A study on different classification models for knowledge discovery [J]. Int. J. Comput. Sci. Mob. Comput, 2015, 4(6): 241-248

[141] Helma C. Data mining and knowledge discovery in predictive toxicology [J]. SAR and QSAR in Environmental Research, 2004, 15(5-6): 367-383

[142] Bertini E, Lalanne D. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery [J]. ACM SIGKDD Explorations Newsletter, 2010, 11(2): 9-18

[143] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases [J]. AI Magazine, 1996, 17(3): 37-37

[144] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. arXiv preprint arXiv:1412.6572, 2014

[145] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of the 38th IEEE Symposium on Security and Privacy. Piscataway, NJ: IEEE, 2017: 39-57

[146] Tao Guanhong, Ma Shiqing, Liu Yingqi, et al. Attacks meet interpretability: Attribute-steered detection of adversarial samples [C] //Proc of the 32st Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 2018: 7717-7728

[147] Liu Ninghao, Yang Hongxia, Hu Xia. Adversarial detection with model interpretation [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1803-1811

[148] Papernot N, Mcdaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C] //Proc of the 1st IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2016: 372-387

[149] Papernot N, Mcdaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning [C] //Proc of the 12th ACM Asia Conf on Computer and Communications Security. New York: ACM, 2017: 506-519

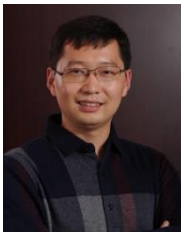
[150] Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile [J]. arXiv preprint arXiv:1710.10547, 2017

[151] Zhang Xinyang, Wang Ningfei, Ji Shouling, et al. Interpretable Deep Learning under Fire [C] //Proc of the 29th USENIX Security Symp. Berkele, CA: USENIX Association, 2020

[152] Dabkowski P, Gal Y. Real time image saliency for black box classifiers [C] //Proc of the 31st Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 2017: 6967-6976

[153] Li Xurong, Ji Shouling, Han Meng, et al. Adversarial Examples versus Cloud-based Detectors: A Black-box Empirical Study [J]. arXiv preprint arXiv:1901.01223, 2019

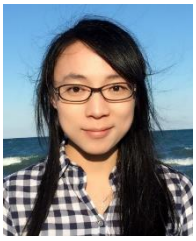
[154] Shi Chenghui, Xu Xiaogang, Ji Shouling, et al. Adversarial CAPTCHAs [J]. arXiv preprint arXiv:1901.01107, 2019



纪守领，1986 年生，获美国佐治亚理工学院电子与计算机工程博士学位、佐治亚州立大学计算机科学博士学位，现任浙江大学“百人计划”研究员、博士生导师、信息安全专业系主任，兼任佐治亚理工学院 Research Faculty，CCF、IEEE 和 ACM 会员。主要研究方向为人工智能安全、数据驱动安全、软件与系统安全、大数据分析等，发表论文 100 余篇。

李进锋，1994 年生，于 2017 年获得武汉理工大学软件工程学士学位，现为浙江大学计算机科学于技术学院硕士研究生。其主要研究方向为数据驱动安全、人工智能安全以及深度学习可解释性.

杜天宇，1996 年生，于 2017 年获得厦门大学通信工程学士学位，现为浙江大学计算机科学与技术学院博士研究生。其主要研究方向为数据驱动安全和人工智能安全。



李博，1989 年生，获美国范德堡大学计算机科学博士学位，现任伊利诺伊大学香槟分校计算机科学系助理教授、博士生导师，IEEE 和 ACM 会员，曾获赛门铁克学术奖金，主要研究方向为人工智能安全、数据安全、大数据分析、博弈论和区块链等，发表论文 100 余篇。