可解释机器学习技术*

作者：杜梦楠　刘宁昊　胡　侠
译者：胡欣宇　岳亚伟

关键词：可解释机器学习

机器学习模型决策机制揭秘。

随着诸如集成模型、深度神经网络 (DNNs) 之类的复杂模型技术的发展，机器学习已经取得了巨大的进步，但其自身仍存在局限性和缺陷，其中很重要的一点就是它们的行为背后缺乏透明性，用户很难理解机器学习模型是如何做出决策的。以先进的自动驾驶汽车为例，虽然它使用了各种机器学习算法，但面对一辆停止的消防车时却无法刹车或减速。这种非预期行为让用户非常失望，并疑惑为什么会出现这种情况。如果自动驾驶车辆正在高速行驶，错误的决策会导致更为严重的后果，可能会使其最终撞上消防车。复杂模型的黑盒特性，已经阻碍了其在自动驾驶汽车等需要关键决策领域的应用。

可解释机器学习是可以缓解该类问题的有效工具。可解释机器学习使得机器学习模型能够以易于理解的方式向用户解释或呈现其行为 [10]，我们称这种特性为可解释性 (interpretability) 或者解释性 (explainability)，在本文中，这两个术语可以互换使用。为了使机器学习更好地为人类服务以造福社会，可解释性是必不可少的。可解释性将会增强终端用户对机器学习系统的信任，并鼓励他们使用机器学习系统。对于机器学习的开发者与研究者，可解释性让其更好地理解问题、数据以及模型失败的原因，最终提升系统的安全性。

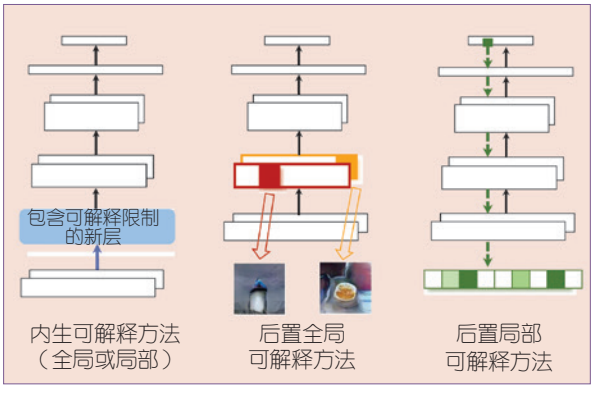

图 1　可解释机器学习的三条技术线（以 DNN 为例）

根据获取可解释能力的时间，可以将可解释机器学习技术分为两类：内生可解释与后置可解释 [23]。内生可解释通过构建自身可解释模型实现，将可解释性直接融入到模型结构。这一类模型包括决策树、规则模型、线性模型、注意力模型等。相比之下，后置可解释需要创建一个额外的辅助模型对原有模型进行解释。这两类模型之间的主要区别在于模型精度和可解释能力之间的权衡。内生可解释模型能够提供准确、无失真的解释，但可能会在一定程度上影响预测性能。后置可解释模型可以保证模型的准确性，但其可解释能力受限于其近似特性。

基于上面的分类，我们进一步将可解释性分为

*本文译自 *Communications of the ACM*, “Techniques for Interpretable Machine Learning”, 2020, 63(1): 68-77 一文，有删节。

全局可解释性和局部可解释性两类。全局可解释性意味着用户可以通过考察复杂模型的结构和参数来理解模型的整体工作方式。而局部可解释性则针对模型的单个预测结果进行局部解释，试图找出模型做出决策的原因。以图 1 中的 DNN 为例，可通过对中间层神经元捕获的表示进行理解得到全局可解释性，而局部可解释性通过识别特定输入中每个特征对 DNN 预测结果的影响程度获得。这两类方法带来的效果不同。全局可解释性可以揭示机器学习模型的内在工作机制，从而提高模型的透明度。局部可解释性有助于揭示特定输入与其相应模型预测之间的因果关系。这两种方法均有助于提升用户对模型和预测结果的信任。

内生可解释模型

内生可解释性可以通过设计自身可解释模型来实现，将可解释性直接融入到模型结构。构造的可解释模型既可以是全局可解释的，也可以为单个预测提供解释。

全局可解释模型

全局可解释模型可以通过两种方式构建：直接使用数据训练，并添加可解释性约束；从复杂且不透明的模型中提取。

添加可解释性约束 引入可解释性约束可以提高模型的可解释性。一些典型的例子包括在分类模型中强制使用稀疏项或强制使用语义单调性约束 [14]。其中，稀疏性意味着模型要尽量使用相对较少的特征进行预测，而单调性使得特征与预测具有单调关系。对于决策树，将子树替换成叶节点进行剪枝，以获得长而深的决策树，而不是宽而平衡的决策树 [29]。这些约束使得模型变得更简单，从而提高模型的可理解性。

此外，可以在模型中添加更多语义约束，以进一步提高可解释性。例如，可解释卷积神经网络 (CNN) 在较高卷积层中添加一个正则损失以学习得到非耦合特征表示，从而得到自然对象的语义检测滤波器 [39]。另外一个工作将称为胶囊的新型神经单元进行组合，构成胶囊网络 [32]。被激活的胶囊的响应向量表示各类语义概念，如特定物体的位置和姿态。这种特点使得胶囊网络更容易被人理解。

然而，当在模型中加入约束时，往往需要在预测精度和可解释性之间进行权衡。具有较好可解释性模型的预测精度可能要比具有较差可解释性模型的预测精度要低。

可解释模型提取 一种替代方案是进行可解释模型的提取，也被称为模拟学习 [36]，此方法不必过分牺牲模型的性能。模拟学习的目标是使用一个易于解释的模型（如决策树、规则模型或线性模型）来近似复杂模型。只要模拟模型和原始模型足够近似，可解释模型就可以反映复杂模型的统计特性。最终，我们得到了一个预测性能相当的模型，其行为也更易理解。例如，将树集成模型转化为单个决策树 [36]。另外，利用 DNN 对决策树进行训练，以模拟神经网络的输入输出函数，从而将 DNN 编码的知识迁移到决策树中 [5]。为了避免决策树过拟合，可采用主动学习方法进行训练。这些技术将原始模型转化为具有更好可解释性的决策树，同时保持了与原始模型相当的预测性能。

局部可解释模型

局部可解释模型通常通过设计更为合理的模型架构来实现，这种架构可以解释机器学习模型做出特定决策的原因。

局部可解释模型的典型框架采用注意力机制 [4, 38]，这种机制被广泛用于解释序列模型所做的预测，如递归神经网络 (RNNs)。注意力机制的优势在于，它通过对单个预测的注意力权重矩阵进行可视化，使用户能够理解模型更关注输入的哪些部分。注意力机制已经用于解决图像标签生成问题 [38]。该方法采用 CNN 将输入图像编码成矢量，使用基于注意力机制的 RNN 来生成图像描述。生成每个描述词时，模型都会改变其注意力，以展现图像中与之关联的部分。注意力机制也被引入到机器翻译任务中 [4]。在解码阶段，将神经注意力模块添加到神经机器翻

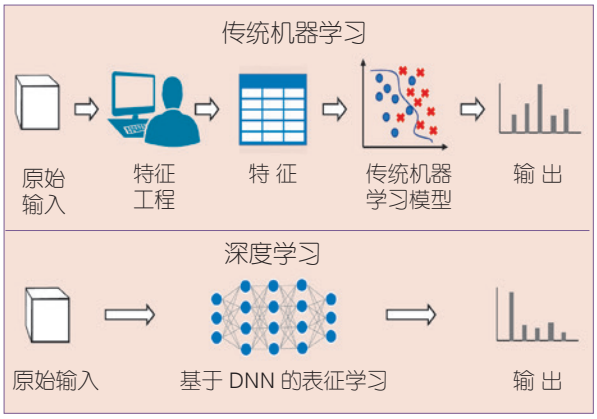

图 2　一个使用特征工程的传统机器学习流水线和一个基于 DNN 表征学习的深度学习流水线

译 (NMT) 模型中，为解码器的隐状态分配不同的权重，使解码器在生成输出的每一步都能有选择地聚焦在句子的不同部分。通过对注意力权重的可视化，用户可以理解一种语言中的单词如何依赖另一种语言中的单词进行正确的翻译。

后置全局可解释

机器学习模型从大量训练数据中自动地学习有用模式，并将学到的知识保存到模型结构和参数中。

后置全局可解释旨在为预先训练的模型所获得的知识提供全局解释，并以直观的方式对模型参数或学习得到的表示进行说明。我们将现有的模型分为传统机器学习和深度学习两类（见图 2），这样分类是因为我们能够从每个类别中提取一些类似的解释范例。

传统机器学习解释

传统的机器学习流水线通常依赖于特征工程，它将原始数据转换为更好地表示预测任务的特征，如图 2 所示。这些特征通常是可解释的，机器学习的作用是将特征表示映射到输出。我们考虑了一种简单有效的解释方法，称为特征重要性 (feature importance)，适用于大多数传统机器学习模型，它表明机器学习模型在进行决策时每个特征的贡献程度。

模型无关解释　与模型无关的特征重要性广泛适用于各种机器学习模型。它将模型视为黑盒，并且不检查内部模型参数。

典型的方法是“置换特征重要性”[1]，其核心思想是，通过对给定特征值置换后模型预测精度的变化进行计算，可以确定给定特征对模型整体性能的重要性。更具体地说，给定一个具有 n 个特征的预先训练模型和一个测试集，该模型在测试集上的平均预测得分为 p，即基线精度。我们将测试集上的特征值重新排序，并在修改后的数据集上计算模型的平均预测得分。

每个特征迭代执行上述过程，最终可以分别得到 n 个特征的 n 个预测得分。然后，我们根据 n 个特征相对于基线精度 p 的得分下降程度来对其重要性进行排序。这种方法有几个优点：首先，我们不需要对人工特征进行归一化；其次，该方法可以推广到几乎所有以人工特征为输入的机器学习模型；最后，该策略在应用中被证明是稳健和有效的。

模型相关解释　针对不同模型也存在特定的解释方法。模型相关解释方法通常通过检查内部模型的结构和参数来得到模型的解释。下面，我们将介绍针对两类机器学习模型计算特征的重要性。

广义线性模型 (GLM) 由一系列模型组成，这些模型是输入特征和模型参数的线性组合，然后输入到一些转换函数（通常是非线性的）构成的模型[21]，如线性回归和逻辑回归。GLM 的权重直接反映了特征的重要性，因此用户可以通过检查权重并将其可视化来了解模型的工作方式。然而，当不同的特征没有被适当地归一化且测量尺度变化时，通过权重进行模型解释不再可靠。此外，当特征维数过大时，该方法所给出解释的可理解性会下降，这可能超出了人类的理解能力。

基于树的集成模型，如梯度提升算法、随机森林和 XGBoost[7]，通常人类是难以理解的。有几种方法可以测量每个特征的贡献度。第一种方法是计算在树分支中采用某特征时的精度提升。如果不为某个特征的分支添加额外分支，则可能存在一些错误分类的元素。在增加额外分支之后，会

存在两个分支，使得每个分支都更准确。第二种方法测量特征覆盖率，即计算与一个特征相关的样本的相对数量。第三种方法是计算一个特征用于数据分割的次数。

DNN 表征解释

DNNs 不仅要研究从表征到输出的映射，而且还要从原始数据中进行表征学习 [15]，如图 2 所示。学习到的深层表征通常是人类无法解释的 [19]，因此对 DNNs 模型的解释主要集中在理解 DNNs 中间层神经元捕获的表征。在这里，我们介绍两类主要的 DNN 模型，即 CNN 和 RNN 的表征解释方法。

CNN 表征解释 针对 CNN 不同层上神秘表征的理解和解释吸引了越来越多的关注。在 CNN 表征解释的不同策略中，最为有效和广泛使用的策略是针对特定层上的神经元确定首选输入。该策略通常是通过激活最大化 (AM) 框架进行描述 [33]。从随机初始化图像开始，我们对图像进行优化，以最大化激活神经元。通过迭代优化，利用神经元响应值相对于图像的导数对图像做出调整。最后，对生成的图像进行可视化，就可以知道单个神经元在其感受野中探索的是什么。事实上，我们可以对任意神经元进行此类操作，从第一层神经元一直到最后一层输出神经元，以此来理解各层上的编码表示。

虽然框架很简单，但实际运用中却面临着一些挑战，其中最大的挑战是会产生奇怪的伪影。优化过程中可能会产生包含噪声和高频模式的不真实图像。由于图像搜索空间大，如果没有进行合适的正则化，即使有满足条件的图像激活神经元，图像也无法辨识。为了解决这一问题，需要利用自然图像先验知识对优化过程进行约束，以生成和自然图像类似的合成图像。一些研究者启发式地提出了一些人为先验，包括总变差范数、α 范数和高斯模糊。此外，可以通过生成模型（如 GAN 或 VAE）生成更为强大的自然图像先验，此类生成模型可将隐空间中的编码映射到图像空间 [25]。这些方法不是直接对图像进行优化，而是对隐空间编码进行优化，以找到能够激活指定神经元的图像。实验结果表明，由生成模型产生的先验信息显著改善了可视化效果。

模型可视化的结果揭示了 CNN 表征的一些有趣的性质。首先，神经网络在多个抽象层次上进行表征学习，从第一层到最后一层，逐渐从一般特征学到了任务相关特征。其次，神经元可以对存在语义关联的不同图像做出响应，展示了神经元具有多面性 [27]。注意，这种现象并不局限于高层神经元，所有层级的神经元都具有多面性，只是高层神经元比低层神经元更具多面性，即高层神经元对某类输入的变化具有更强的不变性。第三，CNN 可以学到对象的分布式编码 [40]，可以使用基于部件的表示来描述对象，这些部件可以跨类别共享。

RNN 表征解释 在对 CNN 解释进行了大量研究后，揭示 RNN 表征（包括 GRUs 和 LSTMs）所编码的抽象知识近年来也引起了人们浓厚的兴趣。语言模型常被用于对 RNN 表征学习的分析，语言模型的目标是根据前一个标识来预测下一个标识。研究表明 RNN 确实能学习到一些有用的表征 [17, 18, 28]。

首先，一些研究通过对能够最大化激活某个单元响应的实际输入标识进行分析，检测了 RNN 最后的隐藏层的表征，并对该层上不同单元的功能进行研究。研究表明，一些 RNN 表征单元能够捕获复杂的语言特性，如语法、语义和长期依赖关系。另外一项研究通过字符级语言模型对 RNN 的响应模式进行分析 [18]。该研究发现，虽然大多数神经单元很难找到特定的含义，但在 RNN 隐层表征中确实存在某些维度，能够关注于某些特定的语言结构，如引号、括号及文本中的行长度。在另外一项研究中，采用词级语言模型对 RNN 各个隐层单元所编码的语言特征进行分析 [17]。可视化结果表明，一些神经元主要对特定的语义类别产生响应，而另外一些神经单元则会对特定的语法类型或依赖关系产生响应。值得一提的是，一些隐层神经单元可以将响应值传递到后续的时间步，这就解释了为什么 RNN 可以学习长期依赖关系和复杂的语义特征。

其次，通过对不同隐藏层学习到的表征进行比对，发现 RNN 可以学习到对象的层次化表征 [28]。该研究表明，RNN 表征与 CNN 表征之间存在相似

之处。例如，采用多层LSTM构建的双向语言模型[28]，分析此模型不同层上的表征可知，模型的下层捕获了上下文无关的语义信息，较高层次的LSTM表征则可对语义上下文进行编码。深度上下文表征层可以通过上下文进行词义消歧，因此可以用于需要考虑上下文的词义理解任务中。

后置局部可解释

局部可解释的目标是标识每个输入特征对模型特定预测结果的贡献。由于局部可解释方法通常将模型的决策输出归因于其输入特征，因此该方法也被称为归因方法。

模型无关解释

模型无关的解释方法可以解释任意机器学习模型，而不关心模型的具体实现。该方法将模型视为黑盒，提供了一种无需模型内部参数即可对模型预测结果做出解释的途径。但同时该方法也存在一定风险，因为我们不能保证给出的解释能够如实反映模型的决策过程。

局部逼近的解释方法基于如下假设，即给定输入的周边邻域内的机器学习预测可由一个可解释的白盒模型近似。该方法产生的可解释模型无须保证在全局范围内都有效，但必须保证在原始输入附近的小邻域内可以较好地对黑盒模型进行近似。后面可以通过考察白盒模型的参数来得到每个特征对模型预测的贡献程度。

一些研究假设某个样本周边邻域的预测可以由输入特征的线性加权组合而成[30]。基于这个假设的归因方法首先对样本周边邻域的特征空间进行采样，构建辅助训练集。然后使用生成的样本和标签训练稀疏线性模型（如Lasso）。这种近似模型与黑盒模型的局部工作原理相同，但更易于考察。最后，通过考察稀疏模型的权重来解释原始模型的输出。

即便是模型的局部行为在某些时候也可能是极其非线性的，线性解释方法可能会导致模型的性能下降。此时需要借助能够表征非线性关系的模型进行局部近似。例如，可以基于if-then规则构建局部逼近的解释框架[31]。在一系列任务上进行的实验表明，该框架可以有效地捕获非线性行为。更值得一提的是，产生的规则不仅能够对当前样本进行解释，并且能够泛化到其他样本的解释。

基于扰动的解释方法遵循的基本逻辑为：特征对模型输出的贡献可以通过预测性能在特征改变时的变化进行衡量。该方法试图回答以下问题：输入的哪些部分的移除会对模型输出产生最大影响？因此，其结果又可称为反事实推理解释。依次对输入特征进行扰动以确定其对输出的贡献，扰动可以通过删除和遮罩两种方式实现。删除是指直接将某特征从输入中删除，但这样做可能不切实际，因为很少有模型允许将特征设置为未知值。遮罩是指将该特征替换为给定参考值，例如词嵌入模型置零或将图像像素值设置为指定灰度值。但遮罩可能会引入新的模型决策依据，并对模型产生副作用[8]。例如，我们使用绿色对图像的一部分进行遮罩，则可能会为草类提供不必要的决策依据。

模型相关解释

也有一些专门针对特定类型模型的解释方法。针对DNN的模型解释方法将神经网络视为白盒，明确利用模型内部结构得出解释。我们将其分为三大类：自顶向下的基于反向传播的方法、自下而上的基于扰动的方法、中间层深度表征研究方法。

基于反向传播的方法　使用反向传播来计算给定输出相对于输入的梯度或其变体，从而得出特征的贡献。在最简单的情况下，我们可以反向传播梯度[33]。其基本假设是，梯度越大表明特征与输出之间的相关性越强。其他一些方法将不同形式的信号反向传播到输入层，例如在反向传播过程中丢弃负梯度值[34]，或将与最终预测结果相关得分反向传播到输入层[3]。这些方法被整合到一个统一框架中，并被重构为修改后的梯度函数[2]。这种整合有助于对不同方法进行全面比较，并有助于在TensorFlow和PyTorch等现代深度学习库下进行有效实施。基于反向传播的方法在实现方面非常高效，因为它们通常仅需要一些

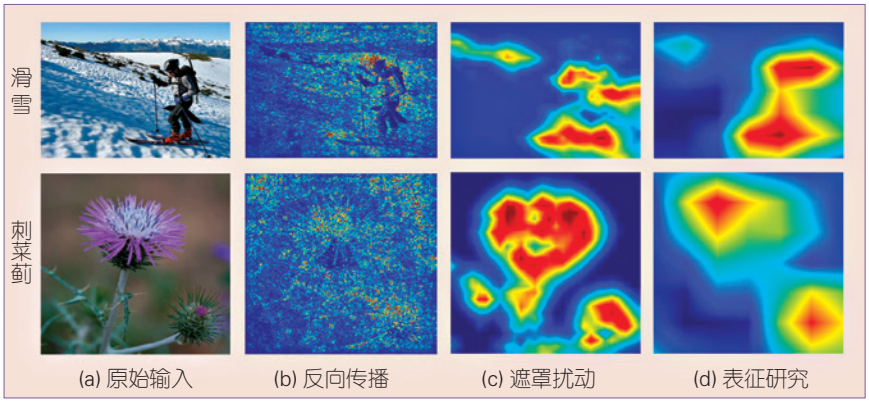

图 3　针对 DNN 局部解释的热图

前向和后向计算。然而，其启发式的性质使其受到限制，可能会产生不如人意的解释，比如解释中包含大量噪声或关注于无关特征，如图 3(b) 所示。

遮罩扰动　在处理高维样本时，由于需要按顺序对输入特征进行扰动，前面提到的模型无关扰动方法会带来较大的计算代价。相反，通过遮罩扰动和梯度下降，可以高效实现对 DNN 相关特征的扰动。一项典型的研究工作在优化框架中对扰动函数进行了定制，以学习遮罩扰动，该遮罩扰动可以显式地保留每个特征的贡献值 [13]。注意，该框架通常需要对遮罩扰动进行各种归一化，以产生有意义的解释，而不是奇怪的伪影 [13]。虽然基于优化的框架极大地提高了效率，但是生成解释依然需要数百个前向和后向操作。为了实现更高效的计算，可以通过训练额外的 DNN 模型预测归因遮罩 [8]。一旦获了得特征遮罩的神经网络模型，只需要进行一次正向传播就可以得出输入归因的得分。

深度表征研究　基于扰动和基于反向传播的解释方法都忽略了 DNN 的深层表征，而 DNN 深层表征中可能包含了丰富的解释信息。为了弥补这一缺陷，一些研究显式地利用网络的深层表征来进行归因。

CNN 深层表征可以捕获图像的高级语义内容及其空间排列，基于这一观察，有研究提出了一种导向性特征反演框架来提供局部解释 [11]。该框架将 CNN 高层表征进行反演转化为合成图像，同时将目标对象的位置信息也编码在遮罩中。分解是探索 DNN 深度表征的另外一种思路。例如，通过对 RNN 模型中隐层表征向量的信息流动过程进行建模，可将 RNN 输出结果分解为模型输入文本中每个单词贡献的叠加 [12]。分解结果可以量化每个单词对 RNN 输出结果的贡献。这两种解释框架在各类 DNN 架构上均取得了不错的结果，表明中间层信息确实对归因起到重要作用。此外，深层表征可以做完强大的正则化算子，从而增加在一般情况下能如实描述 DNN 行为的可能性。同时，它减少了产生奇怪伪影的风险，更够做出更有意义的解释。

应用

模型验证

模型解释可以帮助我们检验机器学习模型是否真正拟合了真实的现象，而非训练数据中存在的偏差。例如，一种后置归因方法分析了三个问答模型 [24]。归因热图显示，这些模型通常会忽视问题的重要部分，而仅依靠不相关的词语做出决策。研究者进一步指出，模型的缺陷是由训练数据不足引起的，解决此问题的方法包括在模型训练过程中对数据进行修改或者引入归纳偏置项。更严重的是，机器学习模型可能依赖性别或者种族偏见做出决策 [9]。模型解释可以帮助我们确定模型是否利用了这些偏见做出决策，以保证模型不违反道德和法律要求。

模型调试

当模型给出错误或意外的决策时，可以通过模型解释对模型的不当行为进行调试和分析。一个典型工作是对抗学习 [26]。最近研究表明，在处理意外或故意编造的输入样本时，可引导机器学习模型（如 DNN）做出高度可信的错误预测 [20, 26]。然而，这些样本非常容易被人类辨识。在这种情况下，模型解

释有助于帮助人们发现模型可能存在的缺陷，并协助分析模型失败的原因。更重要的是，我们可以进一步利用人类认知来找出可能的解决方案，以提升模型的性能和可理解性。

知识发现

模型解释还可以帮助人们理解机器学习模型的决策过程，并从中获得新的知识。基于解释，领域专家和最终用户可以提供较为真实的反馈。最终，可以得出原本隐藏在数据中的新科学和新知识。

研究挑战

解释方法设计

模型解释的第一个挑战与解释方法设计有关，尤其是后置解释方法的设计。我们认为模型解释方法应该局限于对正常条件下模型行为的解释。这句话有两层含义。第一，模型解释应该真实反映机器学习模型的底层机制[12]。后置模型解释方法是对模型行为的近似。有时，后置解释方法近似的精确度不够，其做出的解释可能无法准确反映原始模型的实际运行状态。例如，一种模型解释方法可能会给出人类可理解的解释，而真实机器学习模型则以完全不同的方式工作。第二，即使模型解释能够真实反映模型底层机制，它们也可能无法表征正常条件下的模型行为。模型解释和奇怪的伪影如同一枚硬币的正反面。解释过程可能会生成训练样本分布外的实例，包括无意义的样本和对抗样本[16]，这无疑超出了当前机器学习模型的能力。如果没有进行精心设计，全局和局部解释模型都可能产生机器学习模型的伪影，而非有意义的解释。

解释方法评估

内生可解释方法在评估过程中面临的挑战在于如何对可解释性进行量化。不同的假设会涉及许多可解释模型，并进行不同形式的实现。以推荐系统为例，可解释主题模型和注意力机制都可以提供一定程度的可解释性。但是，我们如何对全局可解释模型和局部可解释模型之间的可解释性进行对比？关于可解释性的含义以及如何对可解释性进行度量，目前仍未达成共识。菲纳利(Finale)和贝恩(Been)提出了三种度量标准：基于应用的度量、基于人的度量和基于功能的度量[10]。这些度量标准相互补充，在有效性和评估成本方面各有利弊。具体采用何种度量标准很大程度上取决于具体任务，以便做出更为有效的评估。

针对后置模型解释方法，评估其可解释性与评估其对原始模型解释的真实度同等重要，而现有研究经常忽略这些。如前所述，针对机器学习模型生成的解释对人类并不总是合理的。很难判断意外解释是由模型的不当行为造成的，还是模型解释方法自身的局限造成的。因此，需要设计更好的指标来度量解释的真实性，以补充现有评估指标。真实性可以衡量我们对一个解释的信任程度。尽管如此，设计合适的真实性度量指标依然是一个悬而未决的问题，值得进一步研究。

讨论

当前模型解释方法的局限

现有可解释机器学习的主要局限是，模型解释是依据研究人员的直觉设计的，而非最终用户的需求。当前的局部模型解释通常以特征重要性向量的形式给出，这些解释可以给出完整的因果归因及底层解释[23]。如果模型解释面向的对象是开发人员或研究人员，则这种形式的解释是令人满意的，他们可以利用特征重要性分布的统计分析来调试模型。如果模型解释面向的对象并不熟悉机器学习，则这种形式的模型解释就不太友好了。它描述了完整的模型决策逻辑，其中包含了大量的冗余信息，使得用户无所适从。模型解释的形式可以进一步改进，以更好提升用户体验。

通往用户友好的解释

基于社会科学和人类行为学的相关研究发现[22]，我们提供了一些以用户为导向的模型解释，使得模型解释可以更有效地被人理解。

比对解释 有时也可以称为差异化解释[22]。该方法不会解释特定决策是怎样得出的，而是解释为什么得出该决策而非其他决策，以便回答诸如“为什么是Q而非R?”之类的问题。其中Q是需要做出解释的事实，而R是比对项，可以是真实的也可是虚拟的。例如，某个用户按揭贷款被拒。用户可能会与其他情形进行对比，并质疑：“为什么我不能按揭，而我的邻居可以？”另一方面，用户可能会问：“为什么我的按揭被拒绝了？”这是一个隐含的比对解释案例,实际上用户正在寻求虚拟场景“如何让我通过按揭贷款？”的解释，因为它是与未发生事件进行比较，所以这里期待的解释也被称为反事实解释[37]。

为了对上面提到的场景提供模型输出的比对解释，可以使用类似的策略对其进行两两比对。首先针对两种情况分别生成特征重要性归因向量：一个针对用户按揭被拒的情况，另一个针对邻居按揭通过的情况（或者是用户设想的按揭可能通过的情况），然后对两个重要性归因向量进行对比。这里我们可以借助对抗性扰动来寻找可能通过的情形，即找出更多的对比案例，使得解释更有意义。最终，我们生成如下形式的解释：“您的按揭被拒绝，因为您的收入低于您的邻居，您的信用记录不如您邻居那么好……”或者“如果您的收入从x增加到y，您的按揭将被通过。”

选择性解释 通常用户不要求模型解释能够涵盖模型做出决策的全部原因。反之，用户希望得到的模型解释能够给出影响模型决策的最重要的因素[22]。一个稀疏解释包含一个最小特征子集，虽然不完整，但有助于帮助理解模型决策是可取的。

可信的解释 好的模型解释应与一般用户的先验知识相一致[23]。假设模型解释给出的按揭被拒的首要条件包括婚姻状况为单身、教育程度为高中毕业，那么这些解释的可信度将比信用记录不良、收入债务比低等解释的可信度要低，因为后两者是导致按揭被拒更合理的原因。可信度低既可能是因为给出的模型解释没有完整反映原始模型，也可能是因为机器学习模型的自身决策并未拟合真实情况。

对话解释 模型解释可以以对话的形式在解释者和解释受众之间传递[22]。这意味着我们必须考虑社会语境，即向谁提供解释[35]，以便明确解释的内容和格式。例如，如果我们向非专业用户解释，模型解释的首选格式是口头解释。

结论

本文旨在帮助各界人士更好地理解不同解释方法的功能和缺点。尽管可解释机器学习技术发展迅速，但仍存在一些尚未解决的关键问题，需要推出新的方案推动该领域的进一步发展。 ■

作 者：

杜梦楠（Mengnan Du）
美国德克萨斯农工大学计算机科学与工程系研究生。
dumengnan@tamu.edu

刘宁昊（Ninghao Liu）
美国德克萨斯农工大学计算机科学与工程系研究生。
nhliu43@tamu.edu

胡 侠（Xia Hu）
美国德克萨斯农工大学计算机科学与工程系助理教授。
xiahu@tamu.edu

译 者：

胡欣宇
CCF高级会员，*CCCF*特邀译者。山西云时代技术有限公司高级工程师。主要研究方向为大数据、物联网和人工智能等。
huxinyu109@126.com

岳亚伟
CCF专业会员。山西农业大学软件学院讲师。主要研究方向为图像处理、机器视觉。yue123161@sxau.edu.cn

（本期译文责任编委：**苗启广**）

参考文献

[1] Altmann, A., Tolos¸i, L., Sander, O. and Lengauer, T. Permutation importance: A corrected feature importance measure. Bioinformatics 26, 10 (2010), 1340–1347.

[2] Ancona, M., Ceolini, E., Oztireli, C. and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In Proceedings of the Intern. Conf. Learning Representations, 2018.

[3] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One 10, 7 (2015), e0130140.

[4] Bahdanau, D., Cho, K. and Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the Intern. Conf. Learning Representations, 2015.

[5] Bastani, O., Kim, C., and Bastani, H. Interpretability via model extraction. In Proceedings of the Fairness, Accountability, and Transparency in Machine Learning Workshop, 2017.

[6] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining. ACM, 2015.

[7] Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining. ACM, 2016.

[8] Dabkowski, P. and Gal, Y. Real time image saliency for black box classifiers. Advances in Neural Information Processing Systems (2017), 6970–6979.

[9] Dix, A. Human issues in the use of pattern recognition techniques. Neural Networks and Pattern Recognition in Human Computer Interaction (1992), 429–451.

[10]Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. 2017.

[11]Du, M., Liu, N., Song, Q. and Hu, X. Towards explanation of DNN-based prediction with guided feature inversion. In Proceedings of the ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining, 2018.

[12]Du, M., Liu, N., Yang, F. and Hu, X. On attribution of recurrent neural network predictions via additive decomposition. In Proceedings of the WWW Conf., 2019.

[13]Fong, R. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the Intern. Conf. Computer Vision, 2017.

[14]Freitas, A.A. Comprehensible classification models: A position paper. ACM SIGKDD Explorations Newsletter, 2014.

[15]Goodfellow, I., Bengio, Y and Courville, A. Deep Learning, Vol.1. MIT Press, Cambridge, MA, 2016.

[16]Goodfellow, I.J., Shlens, J. and Szegedy, C. Explaining and harnessing adversarial examples. In Proceedings of the Intern. Conf. Learning Representations, 2015.

[17]Kádár, A., Chrupa-la, G., and Alishahi, A. Representation of linguistic form and function in recurrent neural networks. Computational Linguistics 43, 4 (2017), 761–780.

[18]Karpathy, A., Johnson, J., and Fei-Fei, L. Visualizing and understanding recurrent networks. In Proceedings of the ICLR Workshop, 2016.

[19]Liu, N., Du, M., and Hu, X. Representation interpretation with spatial encoding and multimodal analytics. In Proceedings of the ACM Intern. Conf. Web Search and Data Mining, 2019.

[20]Liu, N., Yang, H., and Hu, X. Adversarial detection with model interpretation. In Proceedings of the ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining, 2018.

[21]McCullagh, P. and Nelder, J.A. Generalized Linear M, Vol. 37. CRC Press, 1989.

[22]Miller, T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018).

[23]Molnar, C. Interpretable Machine Learning (2018); https://christophm.github.io/interpretable-ml-book/.

[24]Mudrakarta, P.K., Taly, A., Sundararajan, M. and Dhamdhere, K. Did the model understand the question? In Proceedings of the 56th Annual Meeting of the Assoc. Computational Linguistics, 2018.

[25]Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T. and Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Advances in Neural Information Processing Systems, 2016.

[26]Nguyen, A., Yosinski, J. and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition, 2015.

[27]Nguyen, A., Yosinski, J. and Clune, J. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. In Proceedings of the ICLR Workshop, 2016.

[28]Peters, M.E. et al. Deep contextualized word representations. In Proceedings of the NAACL-HLT, 2018.

[29]Quinlan, J.R. Simplifying decision trees. Intern. J. Man-Machine Studies 27, 3 (1987), 221–234.

[30]Ribeiro, M.T., Singh, S. and Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining, 2016.

[31]Ribeiro, M.T., Singh, S. and Guestrin, C. Anchors: Highprecision model-agnostic explanations. In Proceedings of the AAAI Conf. Artificial Intelligence, 2018.

[32]Sabour, S., Frosst, N. and Hinton, G.E. Dynamic routing between capsules. Advances in Neural Information Processing Systems, 2017.

[33]Simonyan, K., Vedaldi, A. and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Proceedings of the ICLR Workshop, 2014.

[34]Springenberg, J.T., Dosovitskiy, A., Brox, T. and Riedmiller, M. Striving for simplicity: The all convolutional net. In Proceedings of the ICLR workshop, 2015.

[35]Tomsett, R., Braines, D., Harborne, D., Preece, A. and Chakraborty, S. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, 2018.

[36]Vandewiele, G., Janssens, G., Ongenae, O., and Van Hoecke, F.S. Genesim: Genetic extraction of a single, interpretable model. In Proceedings of the NIPS Workshop, 2016.

[37]Wachter, S., Mittelstadt, B. and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. 2017.

[38]Xu, K. et al. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the Intern. Conf. Machine Learning, 2015.

[39]Zhang, Q., Wu, Y.N. and Zhu, S.-C. Interpretable convolutional neural networks. In Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition, 2018.

[40]Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. Object detectors emerge in deep scene CNNs. In Proceedings of the Intern. Conf. Learning Representations, 2015.